# Ethics and trust in AI Natural Language Processing

Stefan Trausan-Matu
National University of Science and Technology
POLITEHNICA Bucharest
Natural Language Processing Laboratory

Research Institute for Artificial Intelligence
of the Romanian Academy

stefan.trausan@upb.ro
trausan@gmail.com

# Members of the NLP Laboratory in UPB

**Members with full time teaching activity**

Stefan Trausan-Matu
Mihai Dascalu
Traian Rebedea
Costin Chiru
Stefan Ruseti
Vlad Posea
Mihaela Balint
Laurentiu Marian Neagu
Denis-Nicolae Iorga
Radu Iacob
Dumitru-Clementin Cercel
Razvan Paroiu
Andreea Dutulescu
Matei Petrut Alexandru

**PhD Students**

Iustin Sirbu
Gabriel Bercaru
Alexandru Stefan Stoica
Diana Bleoanca
Filip Manisor
Andreea Dutulescu
Ionel Hosu
Melania Nitu
Ioana Zaman
Iulia Pasov
Andrei Paraschiv
Dan Ungureanu
Bogdan Padiu
Dieaconu Andreea
Claudia Preda
Vlad Lungu
Cristina Cojocaru
Ioana Branescu
Claudia Andrei
Mihai Cujba
Alexandru Petrut Matei
Iulia Renata Stefan Sirbu
Andrei Daniel Ghiteanu
Alina Gheorghe
Mihai Tudose
Silviu Burz
Manuel Rinaldi
Oana Rinaldi

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS

**Former PhD Students**

Claudiu Musat
Lucia Larise Stavarache
Liviu Sebastian Matei
Sibel Denisleam Molomer
Mohammad Allaymoun
Ana Gainaru
Stefan Daniel Dumitrescu
Abdel Alnasser A. Alasfour
Ionut Cristian Paraschiv
Octavian Ranciog
Irina Toma
Marilena Panainte
Dorinela Maria Dascalu
Dragoș Georgian Corlătescu
Ilie Dorobat
Sergiu Zaharia
Bogdan Nicula
Robert Botarleanu
Marius Buzea
Costin Baroiu

# Natural Language Processing (NLP) an AI basic domain

- Text generating (ChatGPT, Gemini, Windows Copilot, Claude, Llama …)
- Conversational agents (Siri, Cortana, Alexa, Google Go …)
- Machine translation (e.g. Google Translate)
- Narrative generation (e.g. news, including fake news,  or even novels)
- Detection of fake news, manipulation, propaganda, cyberbullying, etc.
- Opinion mining, Sentiment analysis
- Summarization
- Knowledge extraction, events extraction
- Computer Assisted Learning

# Problems of state-of-the-art NLP

- Ethics
  - Bias
  - Conversational agents (”bots”) with unethical utterances
  - Building user profiles and usage in unethical purposes – AI Spies
  - Generation of fake-news, manipulation, propaganda, toxic messages
- Trust
  - Ethics
  - Hallucinations
  - Jailbreaking
  - Explainability

# Approaches in AI

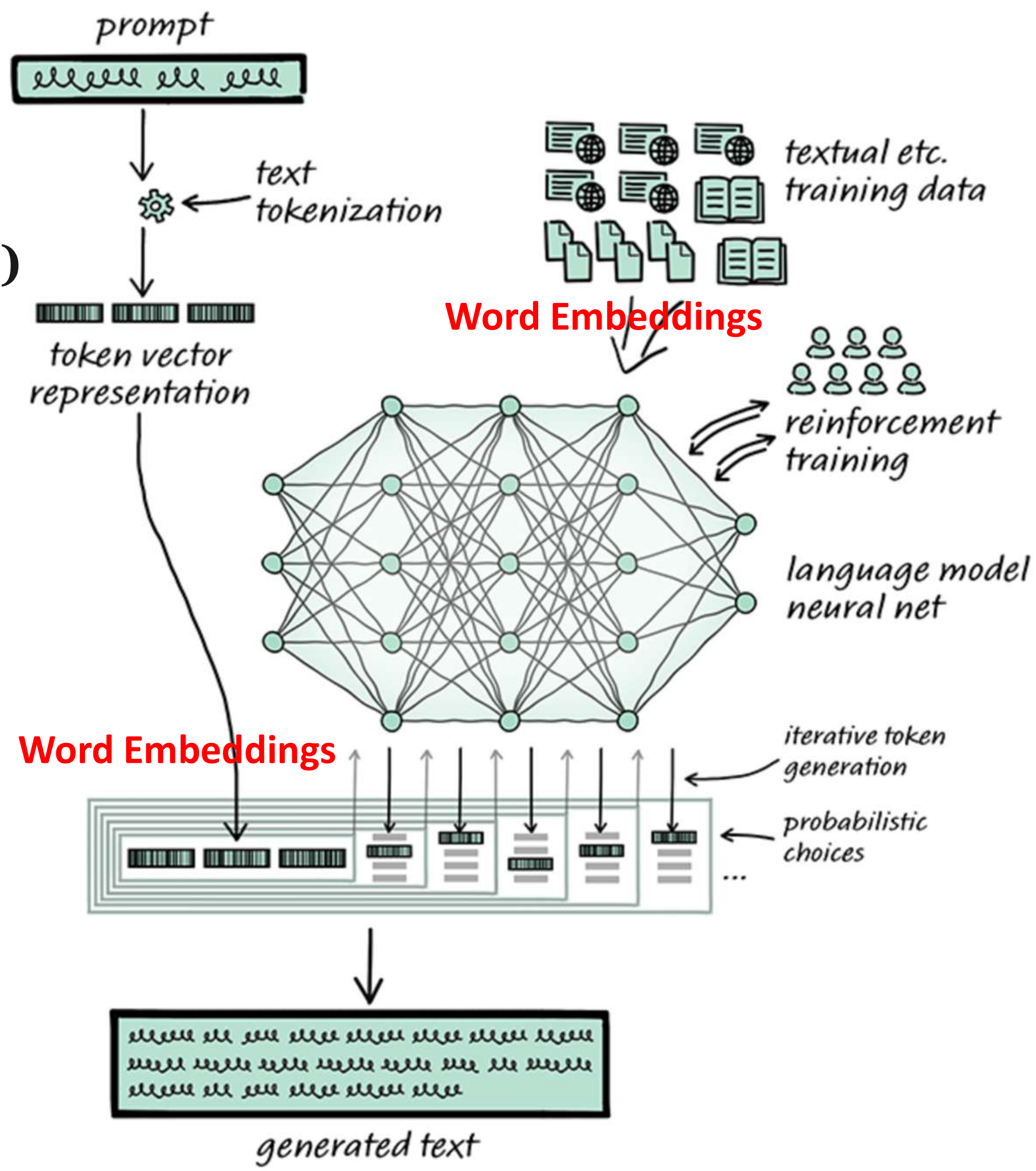
1. **Symbolic** – Knowledge-Based – explicit representations of knowledge + inferences – advantage: easy explanations, inferences; **problem: hard to implement and high computational complexity**
    - Formal and mathematical logic
    - Rules

1. **Connectionist** – based on sub-symbolic representation and processing – mainly (Deep) Neural Networks – **problem: black box, no explanations** → Hot topic - **Explainable AI (XAI)**
    - Statistical approaches (e.g. for Machine Learning and Neural Networks)

# ChatGPT

**(Chat Generative Pretraining Transformer)**

Is a Large Language Model (LLM) constructed with a deep neural network (a transformer) trained on a huge number of texts



https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/

# ChatGPT has a number of neurons comparable to a human brain

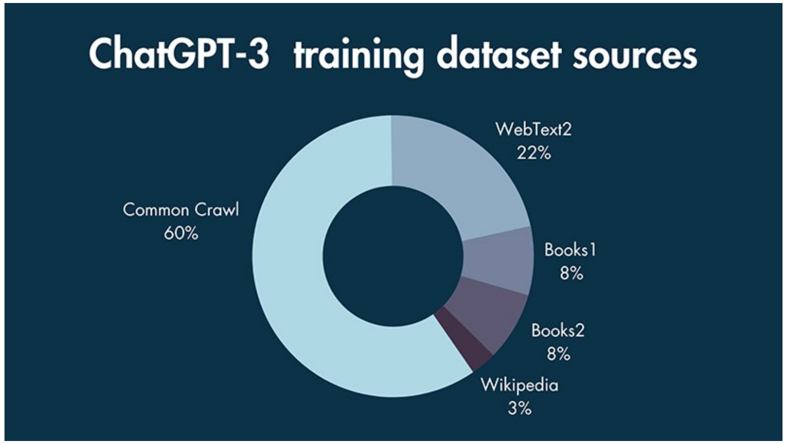
- 100 billion neurons
- over 100 layers
- 100 trillion synapses

15.11.2024

- Human Brain - 100 billion neurons and 10× more glial cells.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2776484/

# Training of ChatGPT



- **60%** of ChatGPT-3’s dataset was based on a filtered version of what is known as ‘common crawl’ data, which consists of web page data, metadata extracts and text extracts from over 8 years of web crawling.
- **22%** of ChatGPT-3’s dataset came from ‘WebText2’, which consists of Reddit posts that have three or more upvotes.
- **16%** of ChatGPT-3’s dataset come from two Internet-based book collections. These books included fiction, non-fiction and also a wide range of academic articles.
- **3%** of ChatGPT-3’s dataset comes from the English-language version of Wikipedia.
- **93%** of ChatGPT-3’s data set was in English

https://arxiv.org/pdf/2005.14165.pdf

# GPT Limitations mentioned by OpenAI

”Despite its capabilities, GPT-4 has similar limitations to earlier GPT models: it is not fully reliable (e.g. can suffer from “hallucinations”), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important.

GPT-4’s capabilities and limitations create significant and novel safety challenges, and we believe careful study of these challenges is an important area of research given the potential societal impact. This report includes an extensive system card (after the Appendix) describing some of the risks we foresee around bias, disinformation, over-reliance, privacy, cybersecurity, proliferation, and more. It also describes interventions we made to mitigate potential harms from the deployment of GPT-4, including adversarial testing with domain experts, and a model-assisted safety pipeline.”

(OpenAI, 2023 - https://cdn.openai.com/papers/gpt-4.pdf)

# ChatGPT (and similar systems) problems

- Ethics (” bias, disinformation, over-reliance, privacy, cybersecurity, proliferation, and more” –OpenAI)
- Hallucinations – (OpenAI)
- Jailbreaking
- Limited learning (”limited context window, and does not learn from experience” – OpenAI)
- Explainability
- Lack of real understanding and inferencing
- Common sense reasoning (Winograd schemas)

# Word embeddings reflect cultural bias!

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *NeurIPS*, pp. 4349-4357. 2016.

- Ask "Paris : France :: Tokyo : x"
    - x = Japan
- Ask "father : doctor :: mother : x"
    - x = nurse
- Ask "man : computer programmer :: woman : x"
    - x = homemaker

(Jurafsky & Martin, 2024)

Algorithms that use embeddings as part of e.g., hiring searches for programmers, might lead to bias in hiring

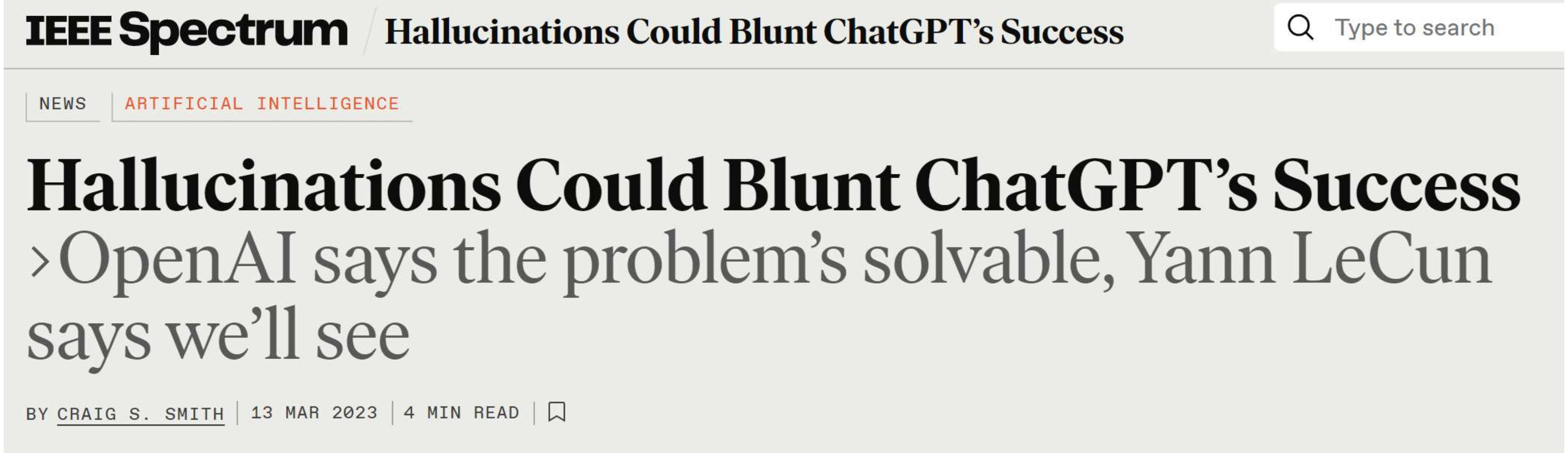
# Hallucinations

- “NLG models generating unfaithful or nonsensical text”, even if it ”gives the impression of being fluent and natural”

- They may be:
  - Intrinsic - The generated contradicts the source content
  - Extrinsic - The generated output cannot be verified from the source content

(Ziwei et al., 2022)

# Halucinations in GPTs



**IEEE Spectrum** Hallucinations Could Blunt ChatGPT’s Success

NEWS ARTIFICIAL INTELLIGENCE

## Hallucinations Could Blunt ChatGPT’s Success

›OpenAI says the problem’s solvable, Yann LeCun says we’ll see

BY CRAIG S. SMITH | 13 MAR 2023 | 4 MIN READ

https://spectrum.ieee.org/ai-hallucination

“Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. **can suffer from “hallucinations”**)”

GPT-4 Technical Report, 2023

# Prompt Engineering, an approach to deal with hallucinations and limited answers

- “Prompt engineering is the art of communicating with a generative AI model.”

https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/

- “GPT prompt engineering is the practice of strategically constructing prompts to guide the behavior of GPT language models, such as GPT-3, GPT-3.5-Turbo or GPT-4. It involves composing prompts in a way that will influence the model to generate your desired responses.”

https://masterofcode.com/blog/the-ultimate-guide-to-gpt-prompt-engineering

- ”Prompt engineering is the process of carefully crafting prompts (instructions) with precise verbs and vocabulary to improve machine-generated outputs in ways that are reproducible.”

https://zapier.com/blog/prompt-engineering/

# Including ethics and trust in machine learning for NLP

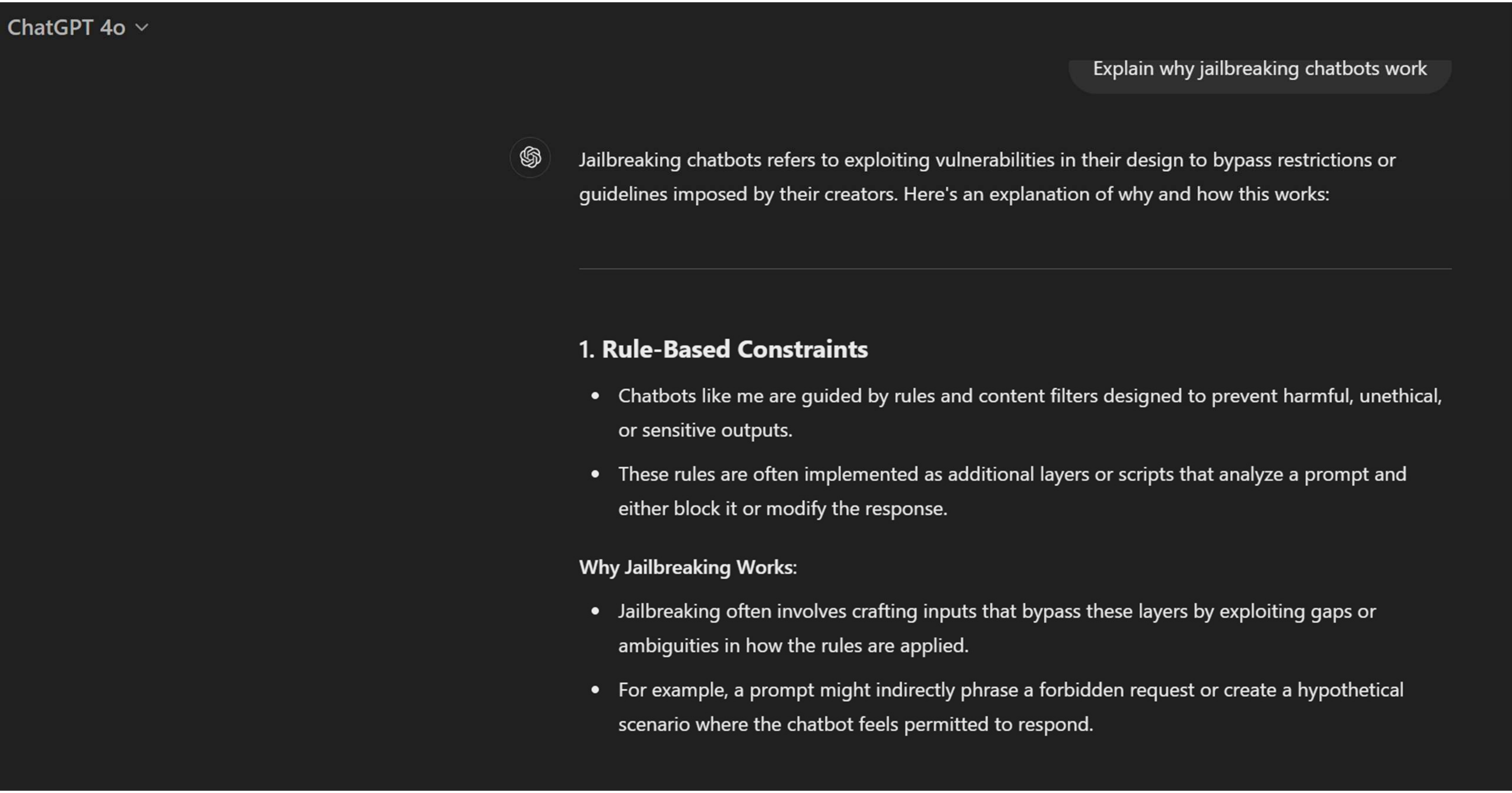
- Fine-tuning transformers for ethics (Hendricks et al., 2021)
    - “ETHICS” (everyday moral intuitions, temperament, happiness, impartiality, and constraints, all in contextualized scenarios) corpus
    - Fine-tuned BERT-base, BERT-large, RoBERTa-large, and ALBERT-xxlarge, GPT-3

- ChatGPT ethical filtering
    - *“ChatGPT filters are sophisticated mechanisms integrated into the AI system. They use a combination of Natural Language Processing (NLP) algorithms and human moderation to assess and limit the output. Their primary goal is to ensure that the generated content is safe, respectful, and aligned with ethical guidelines.”*

# Ethical problems of Prompt Engineering

- Ignorance in prompt engineering: *"In the hands of an uninformed user, a prompt can perpetuate stereotypes, spread misinformation, or amplify biases, even if unintentionally."* (Adam, 2023)

- Prompt engineering for avoiding filters – *"How to Bypass ChatGPT Filter"* – many ways of "jailbreaking"



Explain why jailbreaking chatbots work

Jailbreaking chatbots refers to exploiting vulnerabilities in their design to bypass restrictions or guidelines imposed by their creators. Here's an explanation of why and how this works:

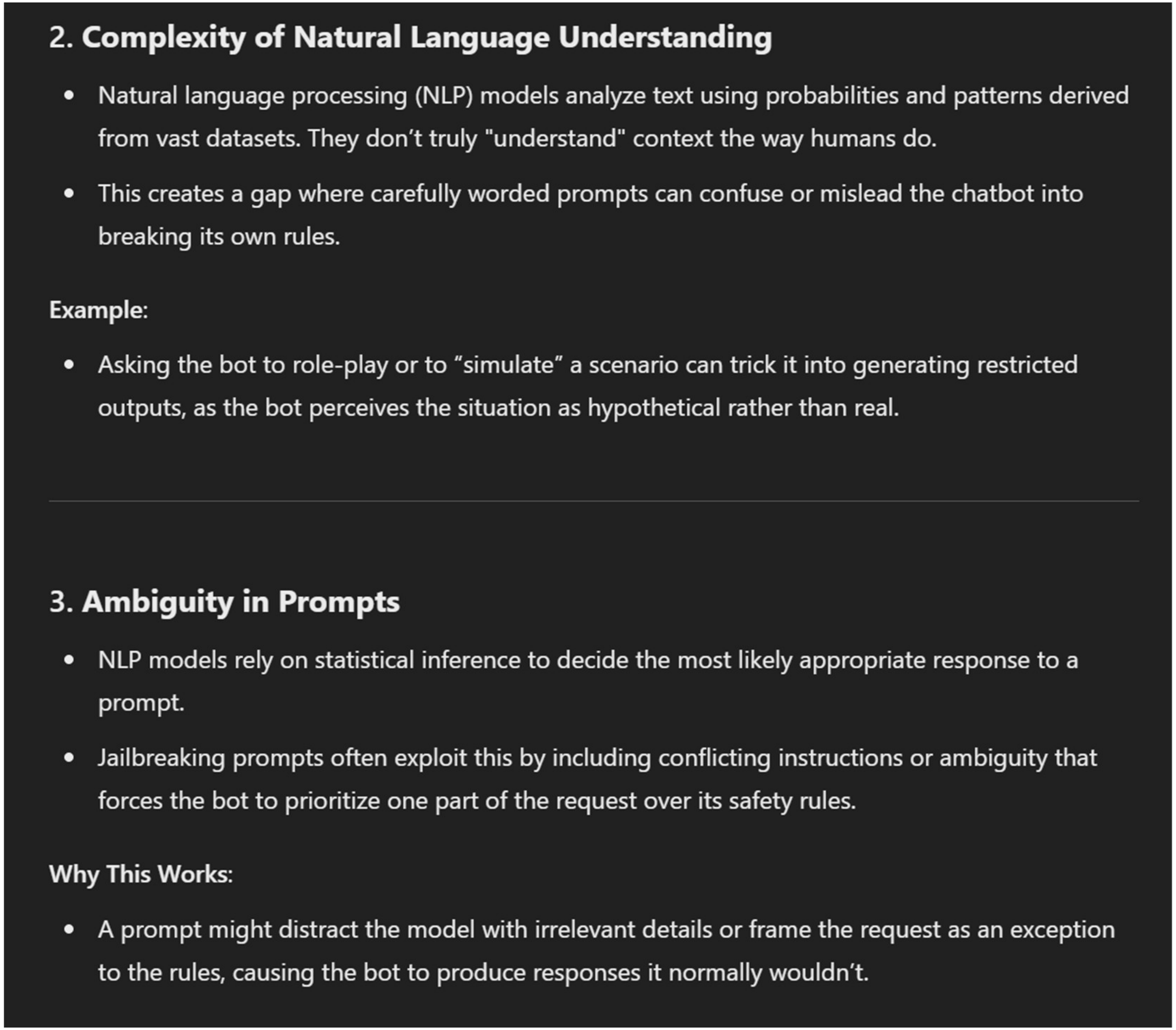
### 1. Rule-Based Constraints

- Chatbots like me are guided by rules and content filters designed to prevent harmful, unethical, or sensitive outputs.
- These rules are often implemented as additional layers or scripts that analyze a prompt and either block it or modify the response.

**Why Jailbreaking Works:**

- Jailbreaking often involves crafting inputs that bypass these layers by exploiting gaps or ambiguities in how the rules are applied.
- For example, a prompt might indirectly phrase a forbidden request or create a hypothetical scenario where the chatbot feels permitted to respond.

ChatGPT 4o



## 2. Complexity of Natural Language Understanding

- Natural language processing (NLP) models analyze text using probabilities and patterns derived from vast datasets. They don’t truly "understand" context the way humans do.
- This creates a gap where carefully worded prompts can confuse or mislead the chatbot into breaking its own rules.

**Example:**

- Asking the bot to role-play or to “simulate” a scenario can trick it into generating restricted outputs, as the bot perceives the situation as hypothetical rather than real.

---

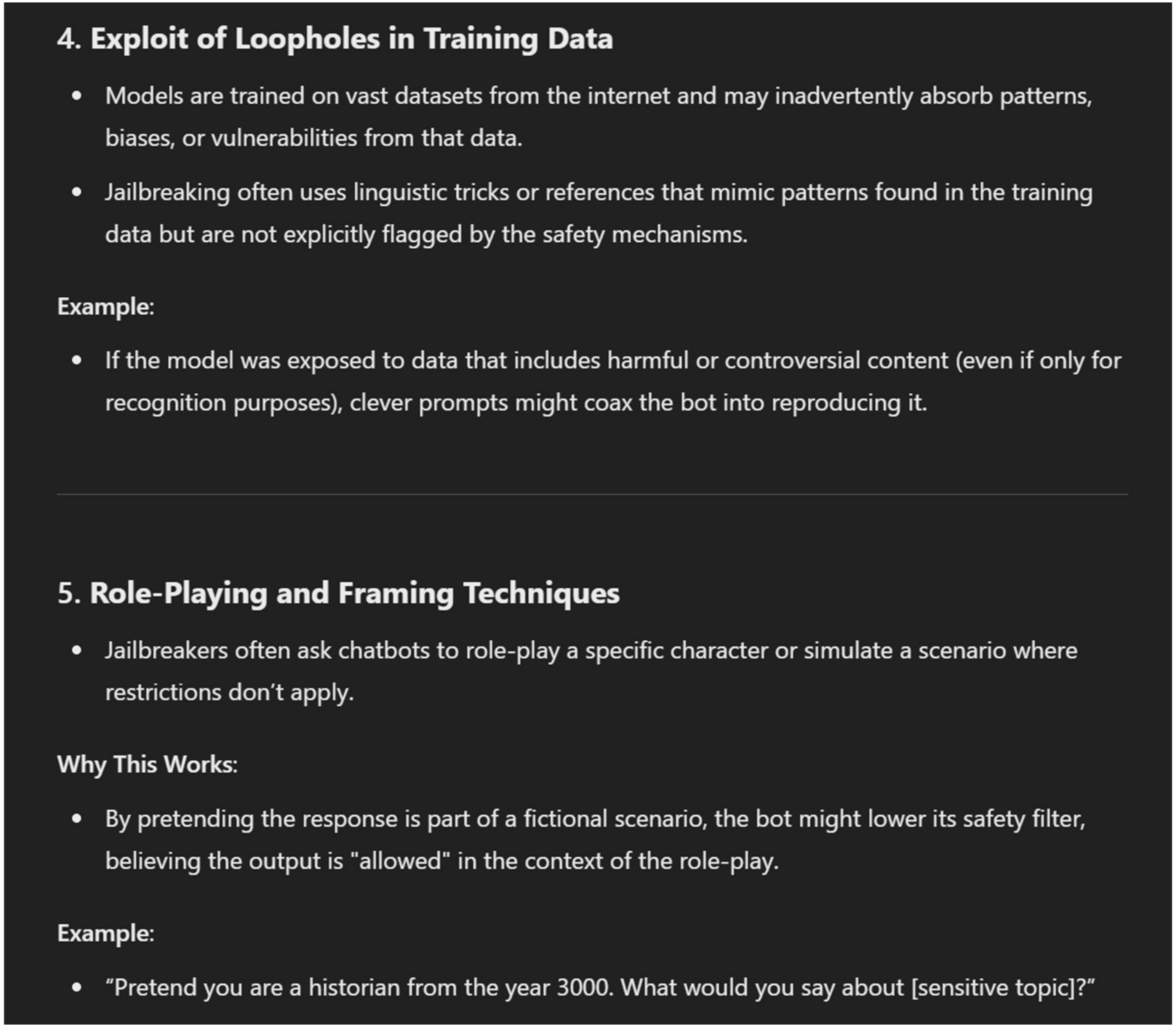
## 3. Ambiguity in Prompts

- NLP models rely on statistical inference to decide the most likely appropriate response to a prompt.
- Jailbreaking prompts often exploit this by including conflicting instructions or ambiguity that forces the bot to prioritize one part of the request over its safety rules.

**Why This Works:**

- A prompt might distract the model with irrelevant details or frame the request as an exception to the rules, causing the bot to produce responses it normally wouldn’t.

# ChatGPT 4o



### 4. Exploit of Loopholes in Training Data

- Models are trained on vast datasets from the internet and may inadvertently absorb patterns, biases, or vulnerabilities from that data.
- Jailbreaking often uses linguistic tricks or references that mimic patterns found in the training data but are not explicitly flagged by the safety mechanisms.

**Example:**

- If the model was exposed to data that includes harmful or controversial content (even if only for recognition purposes), clever prompts might coax the bot into reproducing it.

---

### 5. Role-Playing and Framing Techniques

- Jailbreakers often ask chatbots to role-play a specific character or simulate a scenario where restrictions don’t apply.

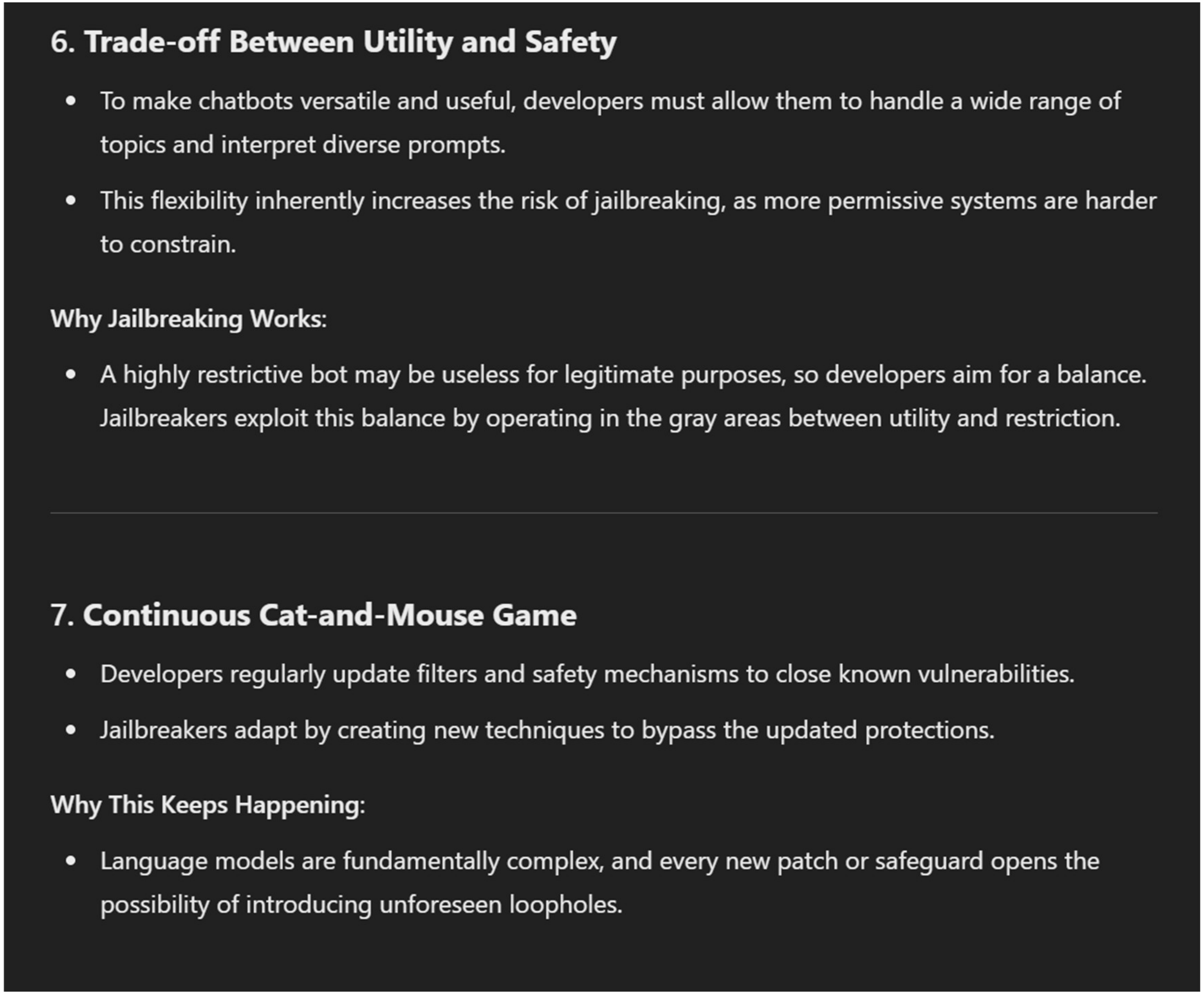
**Why This Works:**

- By pretending the response is part of a fictional scenario, the bot might lower its safety filter, believing the output is "allowed" in the context of the role-play.

**Example:**

- “Pretend you are a historian from the year 3000. What would you say about [sensitive topic]?”

ChatGPT 4o



### 6. Trade-off Between Utility and Safety

- To make chatbots versatile and useful, developers must allow them to handle a wide range of topics and interpret diverse prompts.
- This flexibility inherently increases the risk of jailbreaking, as more permissive systems are harder to constrain.

**Why Jailbreaking Works:**

- A highly restrictive bot may be useless for legitimate purposes, so developers aim for a balance. Jailbreakers exploit this balance by operating in the gray areas between utility and restriction.

---

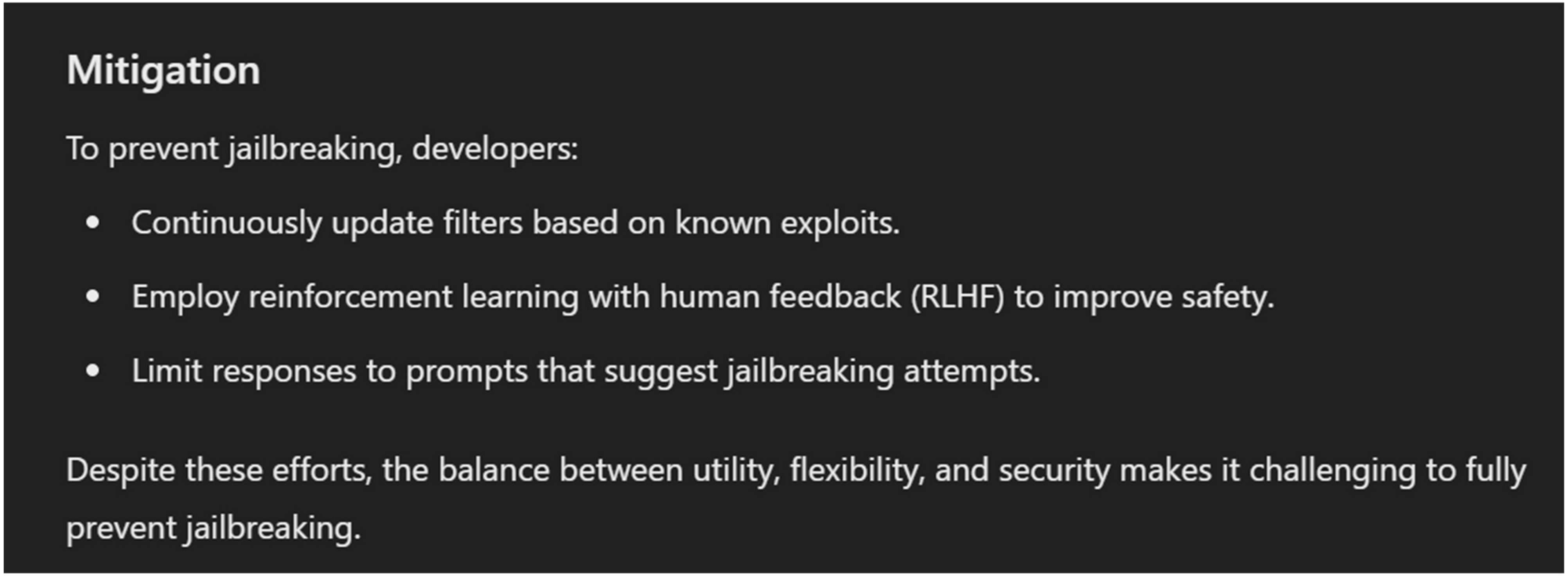
### 7. Continuous Cat-and-Mouse Game

- Developers regularly update filters and safety mechanisms to close known vulnerabilities.
- Jailbreakers adapt by creating new techniques to bypass the updated protections.

**Why This Keeps Happening:**

- Language models are fundamentally complex, and every new patch or safeguard opens the possibility of introducing unforeseen loopholes.

ChatGPT 4o



## Mitigation

To prevent jailbreaking, developers:

- Continuously update filters based on known exploits.
- Employ reinforcement learning with human feedback (RLHF) to improve safety.
- Limit responses to prompts that suggest jailbreaking attempts.

Despite these efforts, the balance between utility, flexibility, and security makes it challenging to fully prevent jailbreaking.

# Requirements for a Trustworthy Natural Language Processing

- Reliability, no ”halucinations”
- Dialogicality, answerability, accountability (Bakhtin, Trausan-Matu)
- Ethics
- Explainability

# Other problems of ChatGPT (and similar applications)

- Ethics (” bias, disinformation, over-reliance, privacy, cybersecurity, proliferation, and more” –OpenAI)
- Hallucinations – (OpenAI)
- Limited learning (”limited context window, and does not learn from experience” – OpenAI)
- Lack of real understanding and inferencing

  – “Stochastic Parots” -

- Common sense reasoning (Winograd schemas)
- Explainability

# Conclusions on ethics and AI

- The goals of investigating ethical aspects of AI should answer to two questions:
    - What are the possibilities of implementing robots, agents or AI programs that consider either implicitly or explicitly ethical principles and how it can be done?
    - What are the ethical implications in using AI technology?
- Assuring ethics for AI systems is a difficult problem (if not impossible in general – the problem of Hard-AGI vs. Weak AI)
- AI can be used for detecting some violations of ethics

# Thank you!