# Special Topics in AI

## Lecture 4: Introduction to LLM

University of Mazandaran

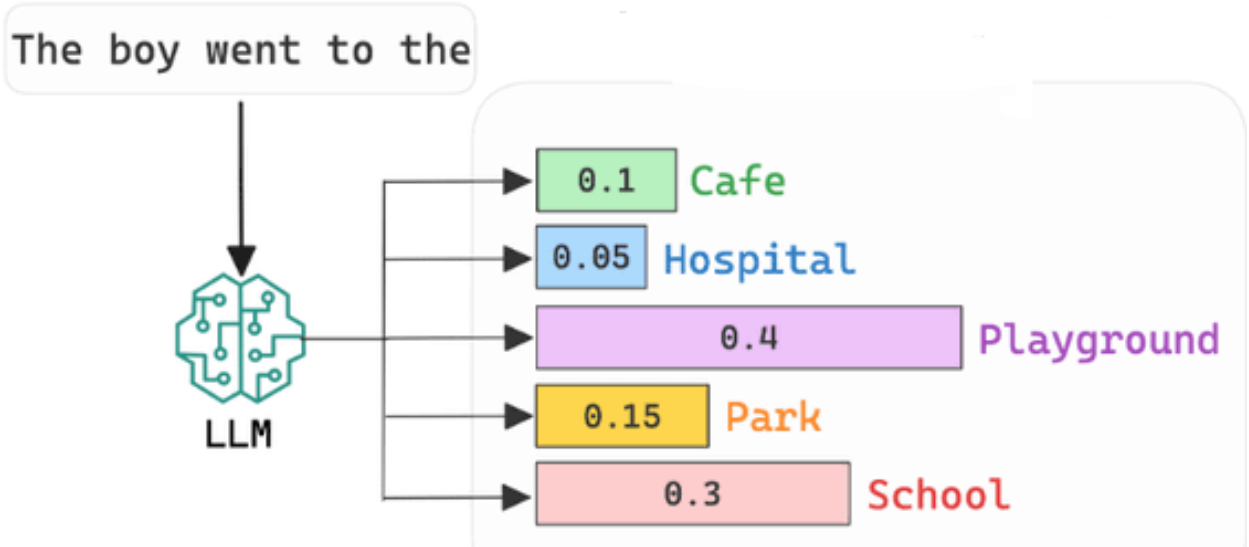
Fall 1405

# Large Language Model (LLM)

Large, general-purpose language models that can be pre-trained and then fine-tuned for specific purposes

# Language Model

- Type of machine learning model trained to predict probability distribution over words
  - Predicts probability of a word in a sequence based on the previous word

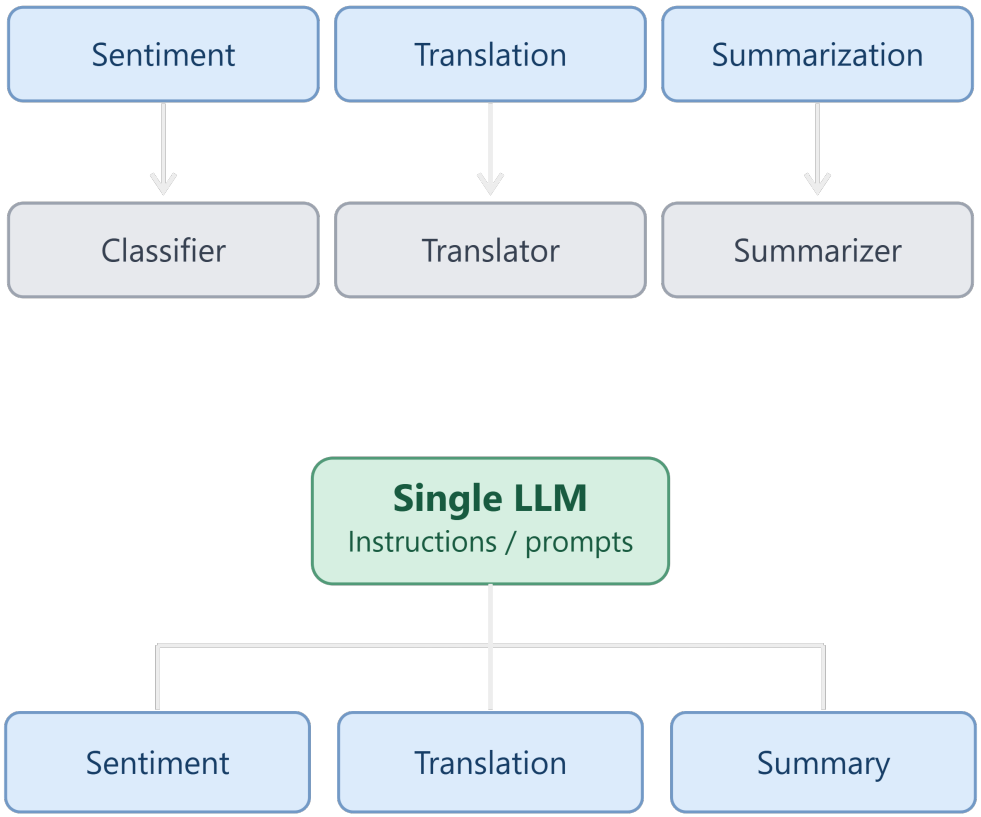


# One Model, Many Tasks

- They are not designed/trained exclusively to solve one specific task
- During training, an LLM processes enormous amounts of text covering
  - science
  - programming
  - literature
  - mathematics
  - etc.

**Key idea:** The same core model can perform different tasks through its input instructions (prompts).

# Traditional task-specific models vs General-purpose LLMs



# General Purpose

Many tasks can be turned into tasks of predicting next words!

- Sentiment analysis
  - "I like football"
  - next word prediction style: The sentiment of the sentece "I like football" is:
  - LLM computes:

$$p(\text{positive}|\text{The sentiment of the sentece "I like football" is})$$

$$p(\text{negative}|\text{The sentiment of the sentece "I like football" is})$$

# General Purpose

Many tasks can be turned into tasks of predicting next words!

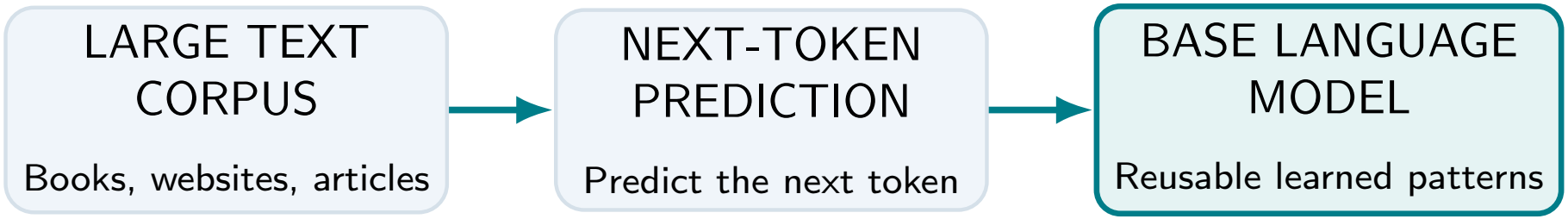
- Q/A
  - Who wrote "The Origin of Species"
  - next word prediction style: The author of the book "The Origin of Species" is
  - LLM computes:

$$p(w|\text{The author of the book "The Origin of Species" is})$$

$$p(w|\text{The author of the book "The Origin of Species" is})$$

# Pre-training

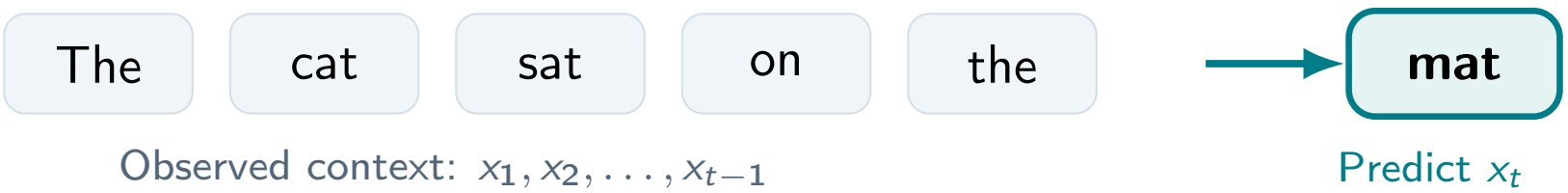
The model is first trained on a broad corpus using a self-supervised learning objective.




**Key idea:** Most LLMs are autoregressive; labelled data are not needed!

## Most LLMs are autoregressive

The training text itself supplies training targets; manually assigned task labels are not required.




**Probability chain rule** (valid for any token sequence $x_{1:T}$):

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$$

An autoregressive LLM **models the conditional terms** $P_\theta(x_t \mid x_{<t})$ using its learned parameters $\theta$.

# Traditional supervised learning vs Self-supervised learning

Sentiment analysis

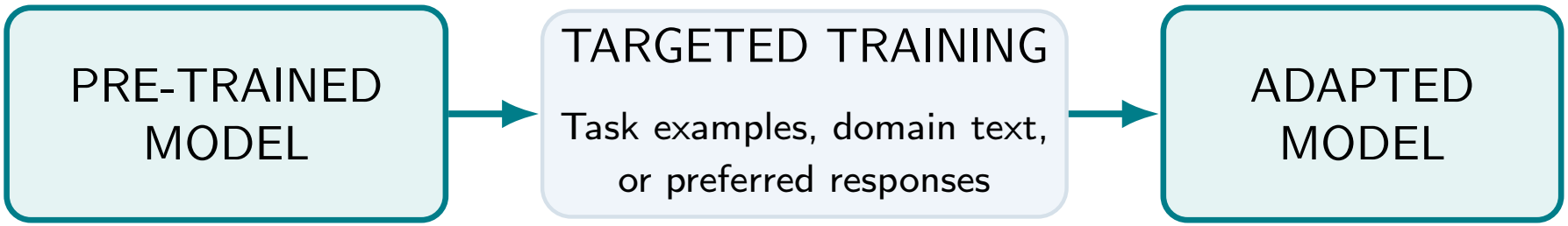
| **Input Text** | **Label** |
|---|---|
| "The movie was amazing!" | Positive |
| "The food was terrible" | Negative |
| "I really enjoyed the book." | Positive |

"The cat is sleeping on the mat."



# Fine-Tuning

Continue training on a smaller, targeted dataset to modify model behavior or improve performance for a specific need.



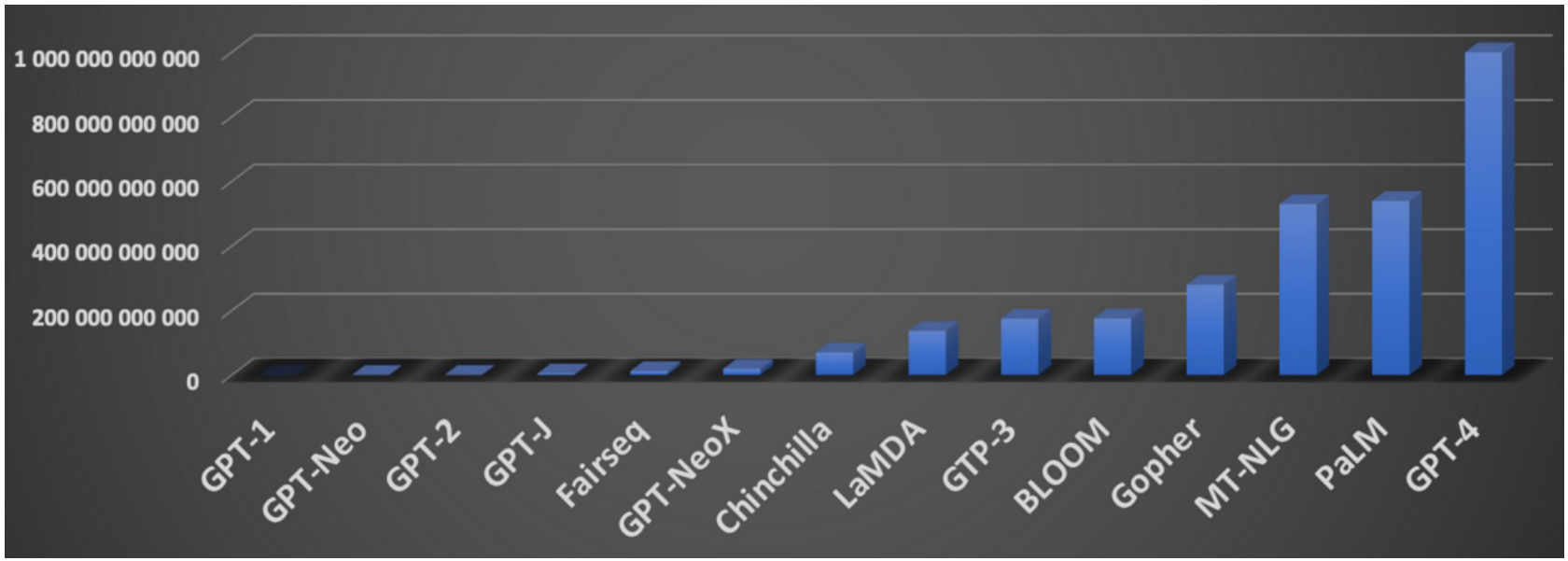

**Key idea:** Fine-tuning updates model parameters.

# Why Large?

- computing sequence probabilities requires massive datasets
- learning patterns in these massive datasets requires millions of millions of parameters

# Why Large?

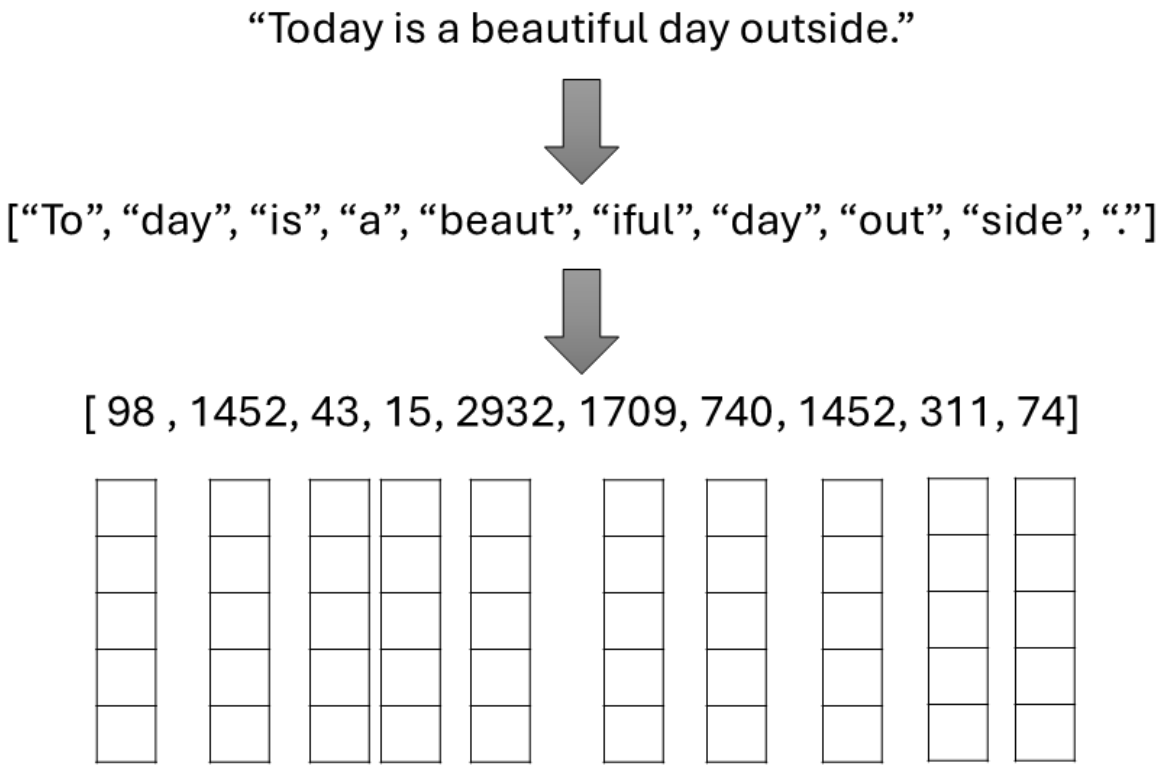
- computing sequence probabilities requires massive datasets
- learning patterns in these massive datasets requires millions of millions of parameters




link

# Tokenization

- Tokens are the basic units of input and output for LLMs
- Tokens could be words or subwords
  - for efficiency reasons, LLMs don't use whole words



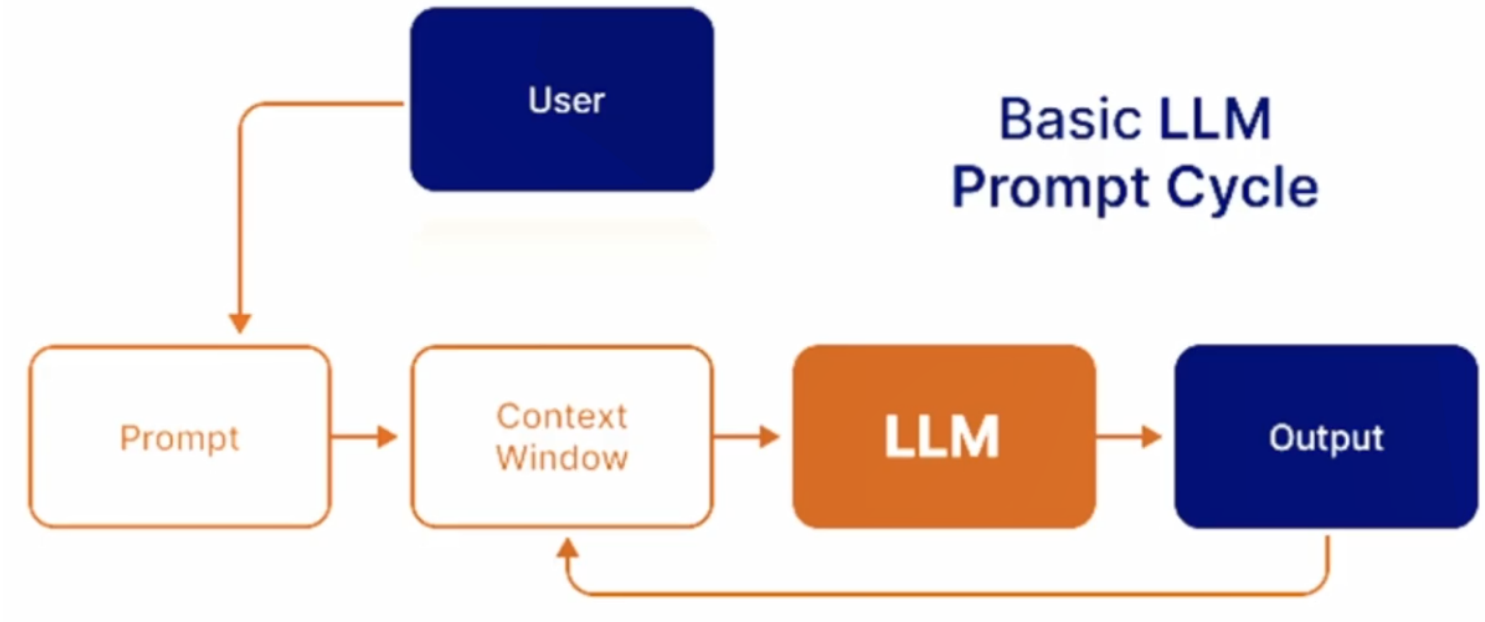
# Generation

Sample one token at a time given the tokens generated so far

$$\text{for } i = 1, \ldots, L:$$
$$x_i \sim p\left(x_i \mid x_{1:i-1}\right)^{\frac{1}{T}}$$

- $T \geq 0$ is a temperature parameters
  - controls how much randomness is required from the language model

# Prompt Cycle




link

# Types of LLMs

Encoder only:

- Suited for tasks that can understand language
- Applications: toxicity classification, sentiment analysis, etc.
- Example: BERT (Bidirectional Encoder Representations for Transformers)

# Types of LLMs

Decoder only:

- Suited for generating language and content
- Applications: story writing, Q/A, etc.
- Example: GPT-3 (Generative Pretrained Transformer 3)

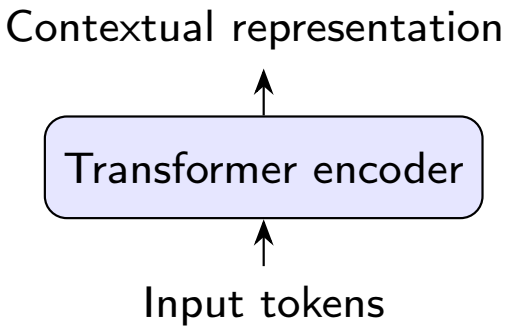
# Types of LLMs

Encoder-decoder:

- Combine the encoder and decoder components of the transformer architecture
- Suited for both understanding and generating content
- Applications: translation, text-to-code, summarization , etc.
- Example: T5 (Text-to-Text Transformers) by HuggingFace

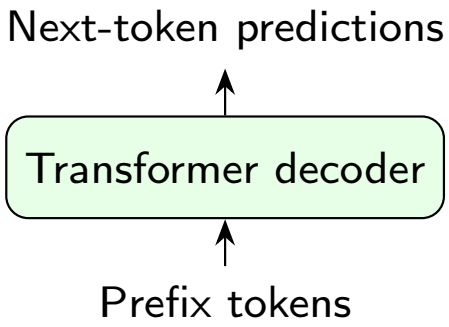
# Types of LLMs

## Encoder-only



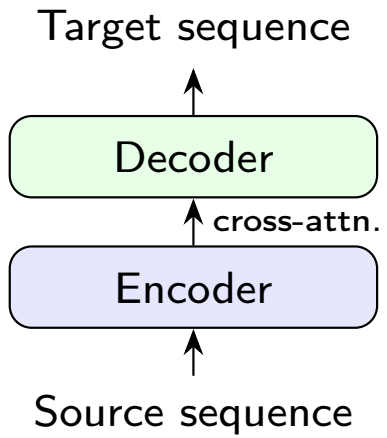

Learns representations of the *entire input*. Often used for understanding and classification.

## Decoder-only




Generates text *autoregressively*, conditioned on the tokens seen so far.

## Encoder–decoder




Encodes a source, then generates an output *conditioned on the source*.

All three typically use Transformer blocks; the central differences are **attention visibility**, **pretraining objective**, and **how outputs are produced**.

# Types of LLMs: Summary

| Dimension | Representative models | Typical tasks |
|---|---|---|
| **Encoder-only** | BERT, RoBERTa, DeBERTa | Sentiment analysis, NER, extractive QA |
| **Decoder-only** | GPT, Llama, Qwen | Chat, creative writing, code |
| **Encoder–decoder** | T5, BART, FLAN-T5 | Translation, summarization |

## Nuance

These are architectural *tendencies*, not hard task boundaries: decoder-only models can classify and summarize; encoders can support generative systems when combined with suitable decoders.

# Types of LLMs: Summary

Which architecture should we choose?

1. **Do we primarily need a label, span, or embedding from a fixed input?**
   ⇒ Start with an **encoder-only** model.

2. **Do we need flexible continuation, conversational output, or long-form text?**
   ⇒ Start with a **decoder-only** model.

3. **Do we need to transform an explicit input sequence into an output sequence?**
   ⇒ An **encoder–decoder** model is a natural fit.

# Multimodal LLMs

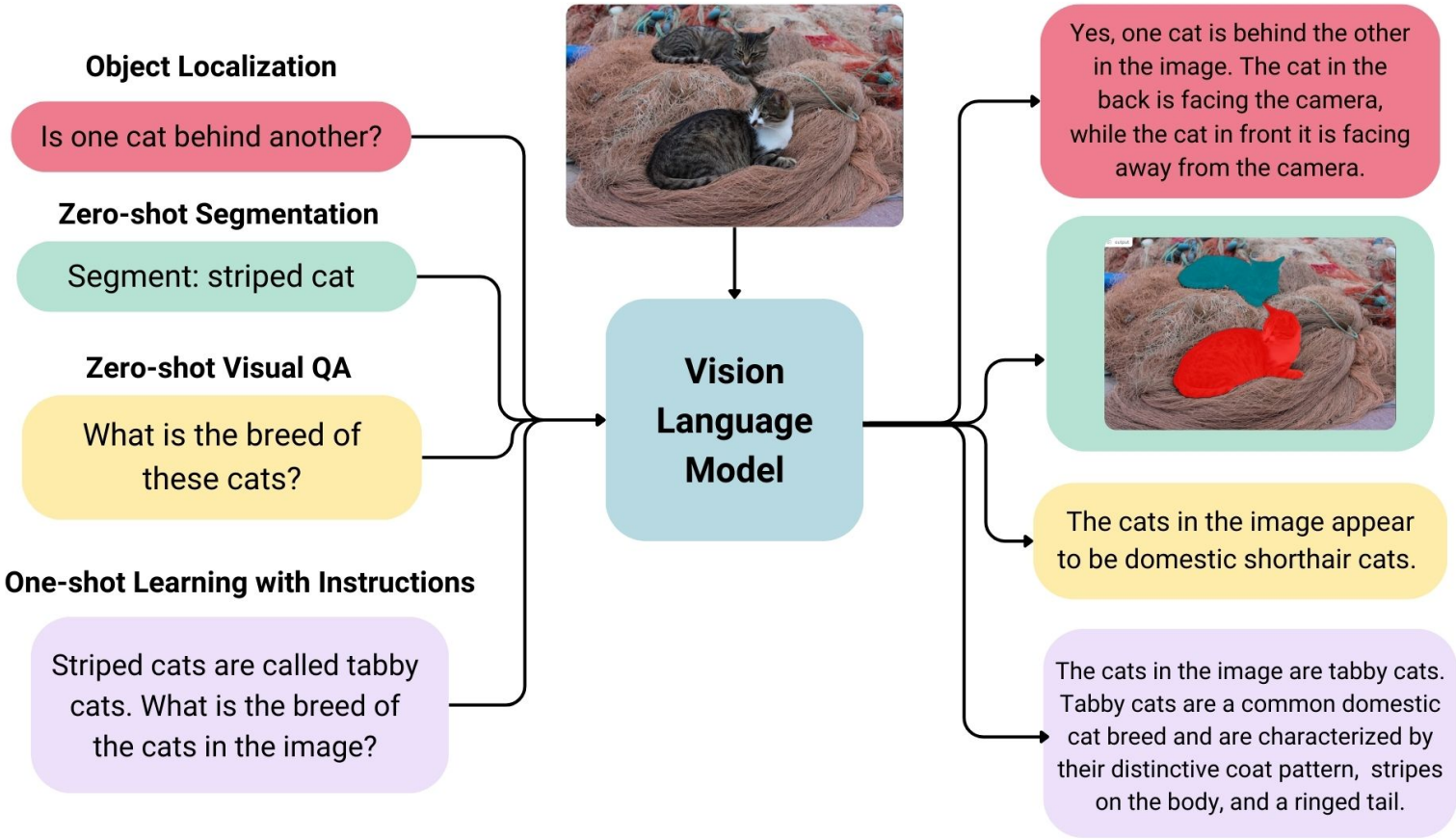
- Unimodal LLMs are trained on a single modality of data, e.g. text
- Lack visual context
- Multimodal LLMs
  - understand and generate content across multiple modalities
  - text, images, audio, video, or other sensor data
- Vision-Language Models (VLMs)
  - multimodal models that can learn from images and texts

# Vision-Language Models

Unify the image and text representation, and feed it to a textual decoder for generation

- Typically consists of two main components
  1. vision encoder: extracts image features and encodes them into a format that can be understood by the language decoder. Examples are ResNet, ViT
  2. language decoder: takes these encoded visual features along with any textual input and generates descriptions, or captions, etc. Examples include GPT-2, GPT-3

# Vision-Language Models




link

# Contrastive Language-Image Pretraining (CLIP)[1]

Differs from traditional vision-language models

- It does not generate text description or captions for images
- Focuses on learning a joint representation space where images and texts can be compared directly

Enables various downstream tasks such as:

- Zero-shot image classification: classify images into one of several classes, without any prior training or knowledge of the classes
- Zero-shot image retrieval, text based image generation

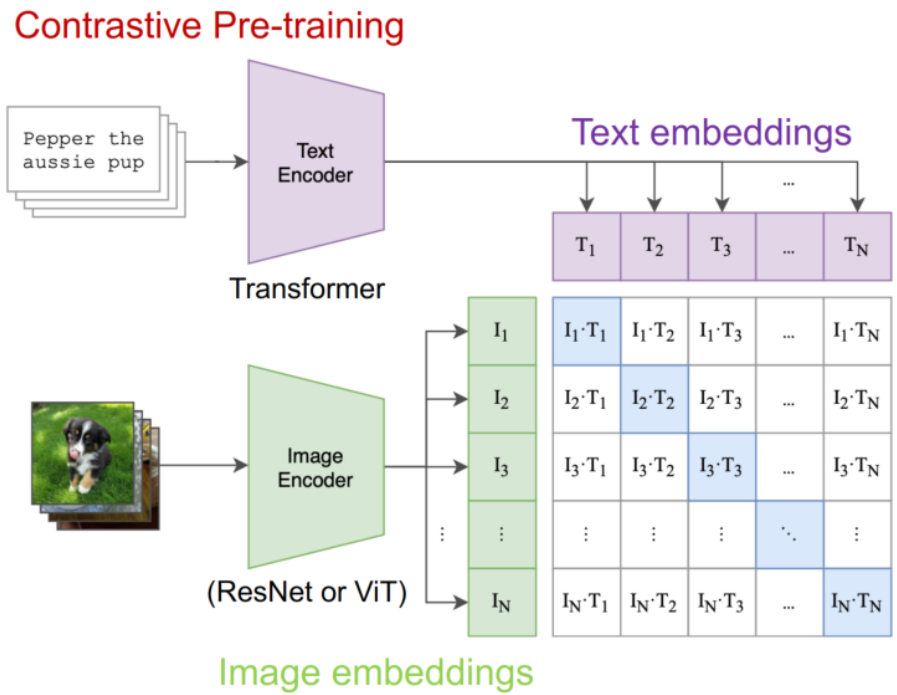
[1] Learning transferable visual models from natural language supervision, Radford et al. ICML (2021)

# Contrastive Language-Image Pretraining (CLIP)[1]

1. Jointly train a text encoder and an image encoder

**Goal**: being able to predict the correct image–text pair

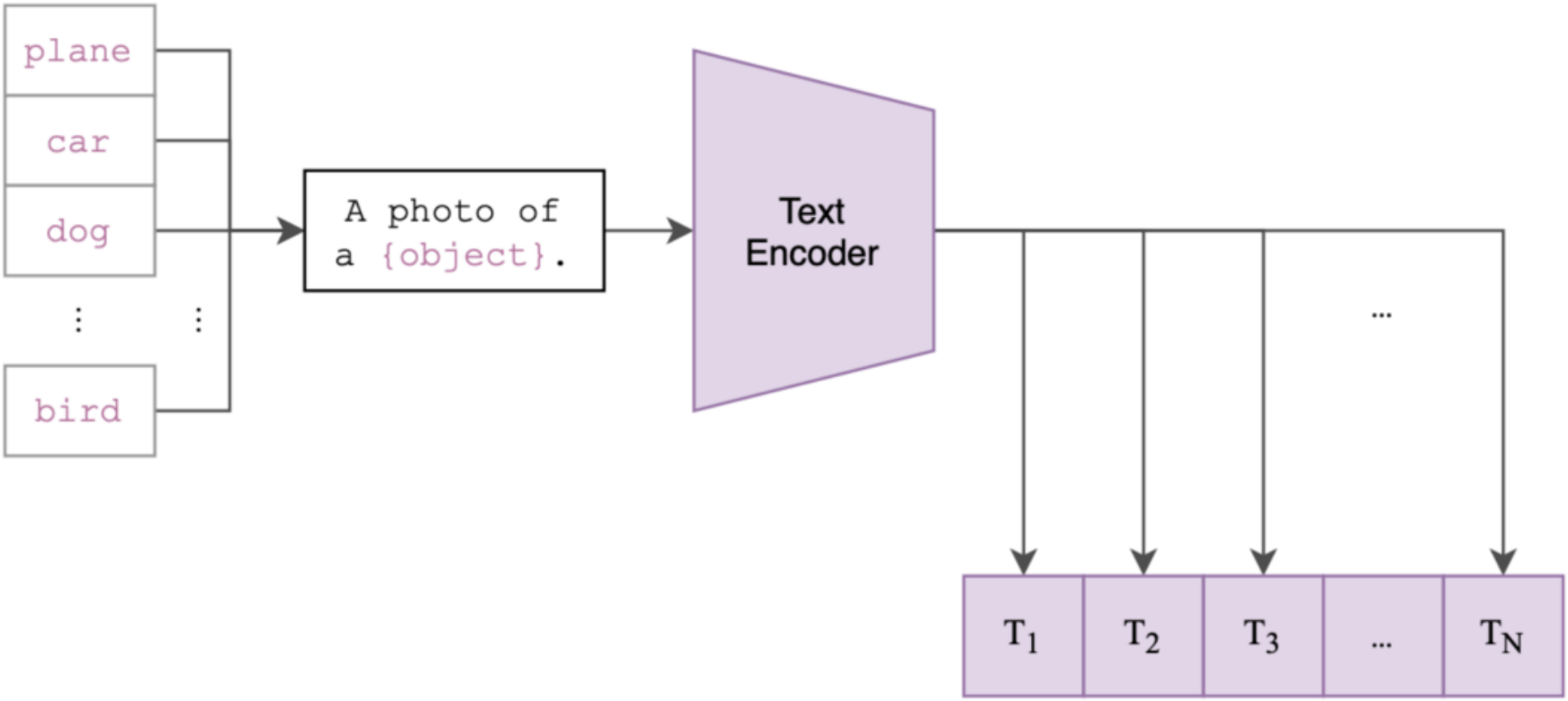
- Maximize the cosine similarities between correct image-text pairs
- Minimize the cosine similarities for dissimilar pairs (non-diagonal elements)




[1] Learning transferable visual models from natural language supervision, Radford et al. ICML (2021)

# Contrastive Language-Image Pretraining (CLIP)[1]

2. Convert training dataset classes into captions

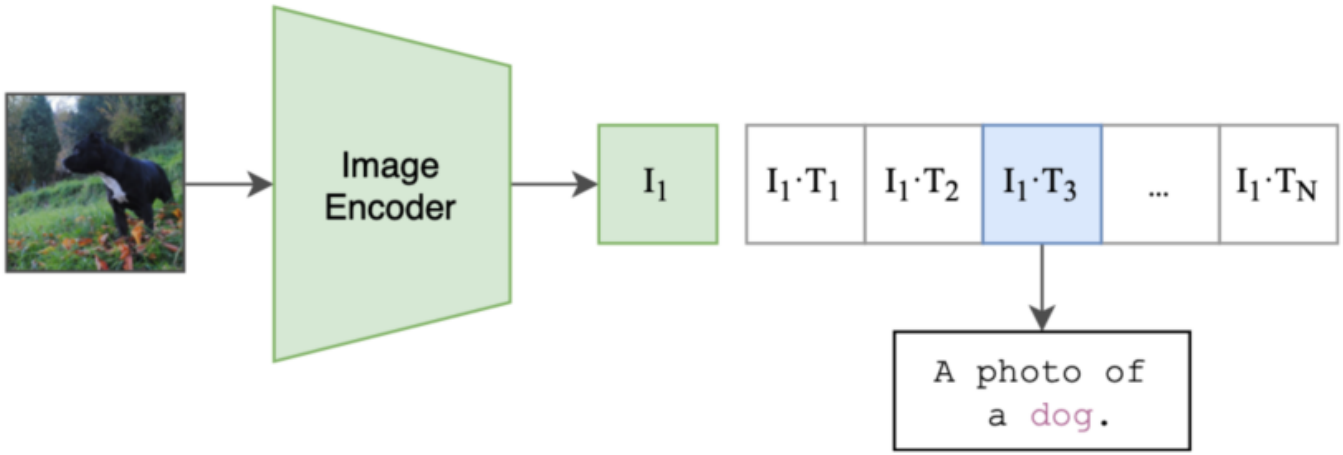



[1] Learning transferable visual models from natural language supervision, Radford et al. ICML (2021)

# Contrastive Language-Image Pretraining (CLIP)[1]

3. Estimate the best caption for a given input image

- Compute similarity between an image vector and multiple caption vector
- Select the one with highest score




[1] Learning transferable visual models from natural language supervision, Radford et al. ICML (2021)

# Contrastive Language-Image Pretraining (CLIP)[1]

- Trained on 400M image-text pairs
- CLIP can perform image tasks using only text! No extra training needed!
- CLIP encodings + decoder models (e.g. GPT-2) $\Longrightarrow$ image captioning

# Are LLMs Intelligent?

## Intelligence

- Philosophy: the ability to solve problems
- Neuroscience: a general mental capability that, among other things, involves the ability to[1]
  - comprehend complex ideas
  - plan
  - solve problems

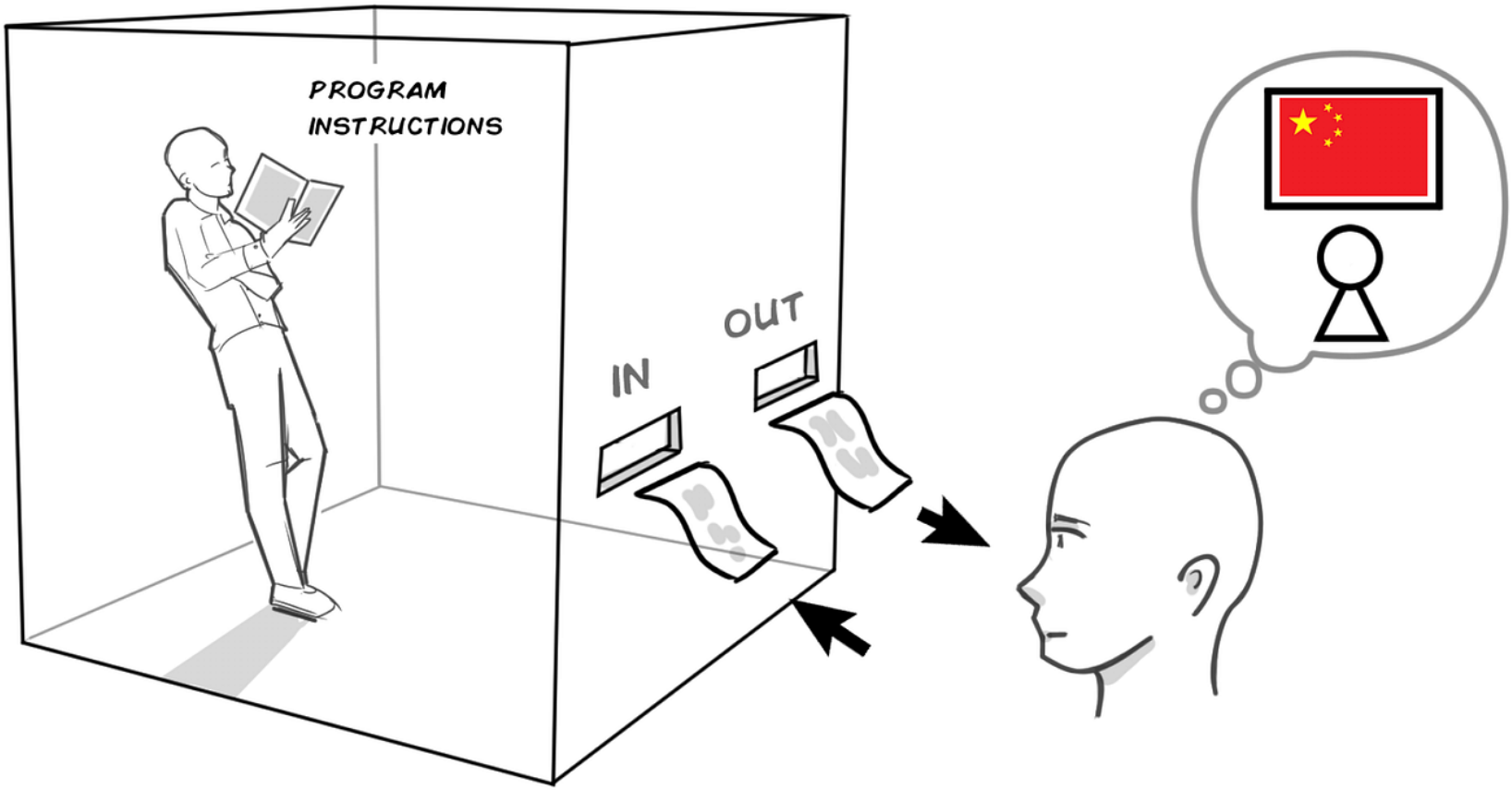
## Consciousness

- Philosophy: One's subjective experience
  - how you perceive your inner and outer world
- Neuroscience: collective mental phenomena
  - memory, 5 senses, emotions, thoughts, etc.
- Stanford encyclopedia of Philosophy: a being that is aware of not only their thoughts, feelings, and senses but also that they are **aware** of their awareness[2]

[1] Deary, I.J. et al. "The neuroscience of human intelligence differences." Nature Reviews Neuroscience, 2010.

[2] Nodelman, Uri et al. Stanford Encyclopedia of Philosophy, 2003.

## The Chinese room experiments[1]

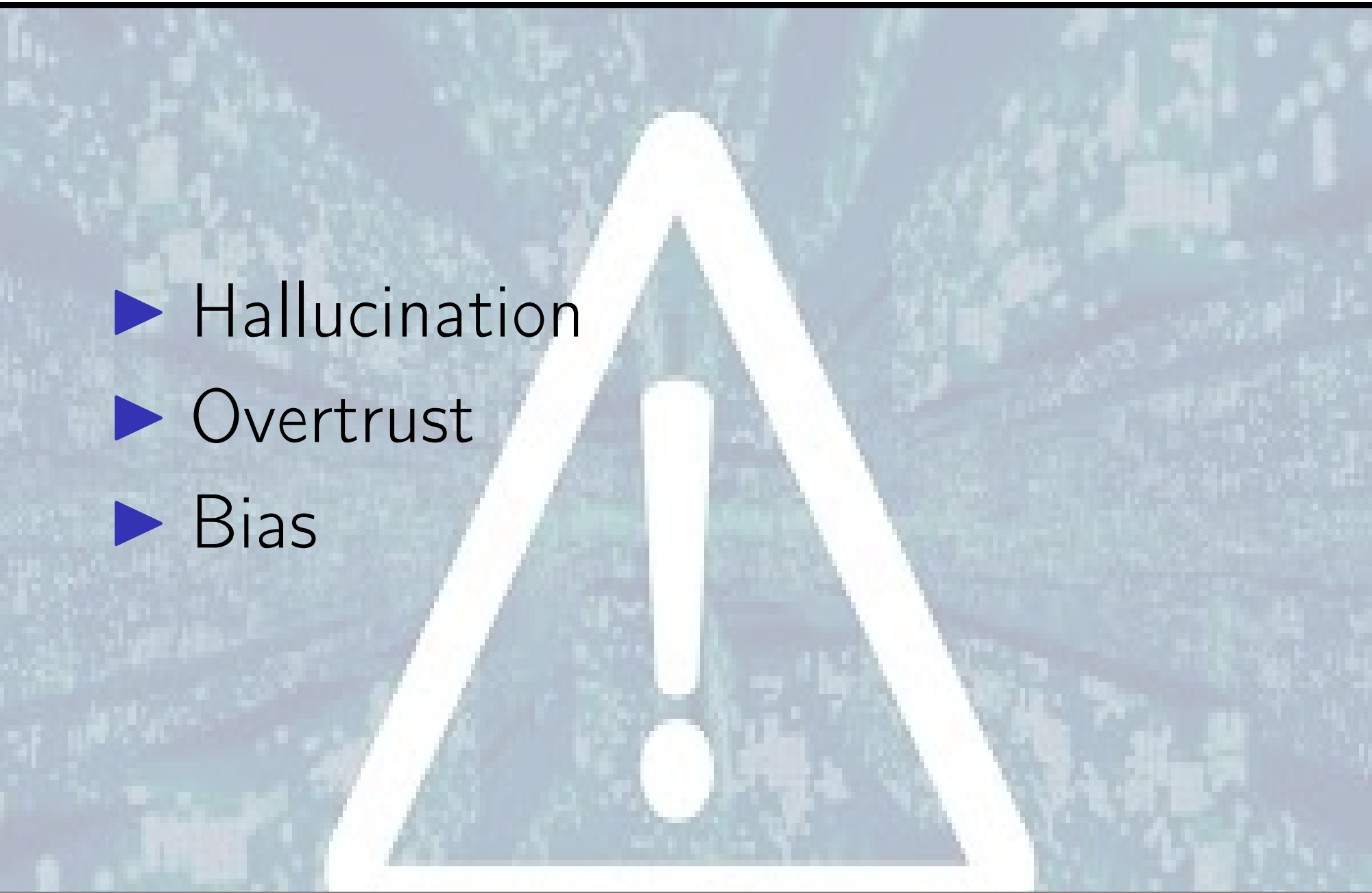



[1] Searle, J.R., 1980. Minds, brains, and programs. Behavioral and brain sciences, 3(3), pp.417-424

## Why Does It Matter?

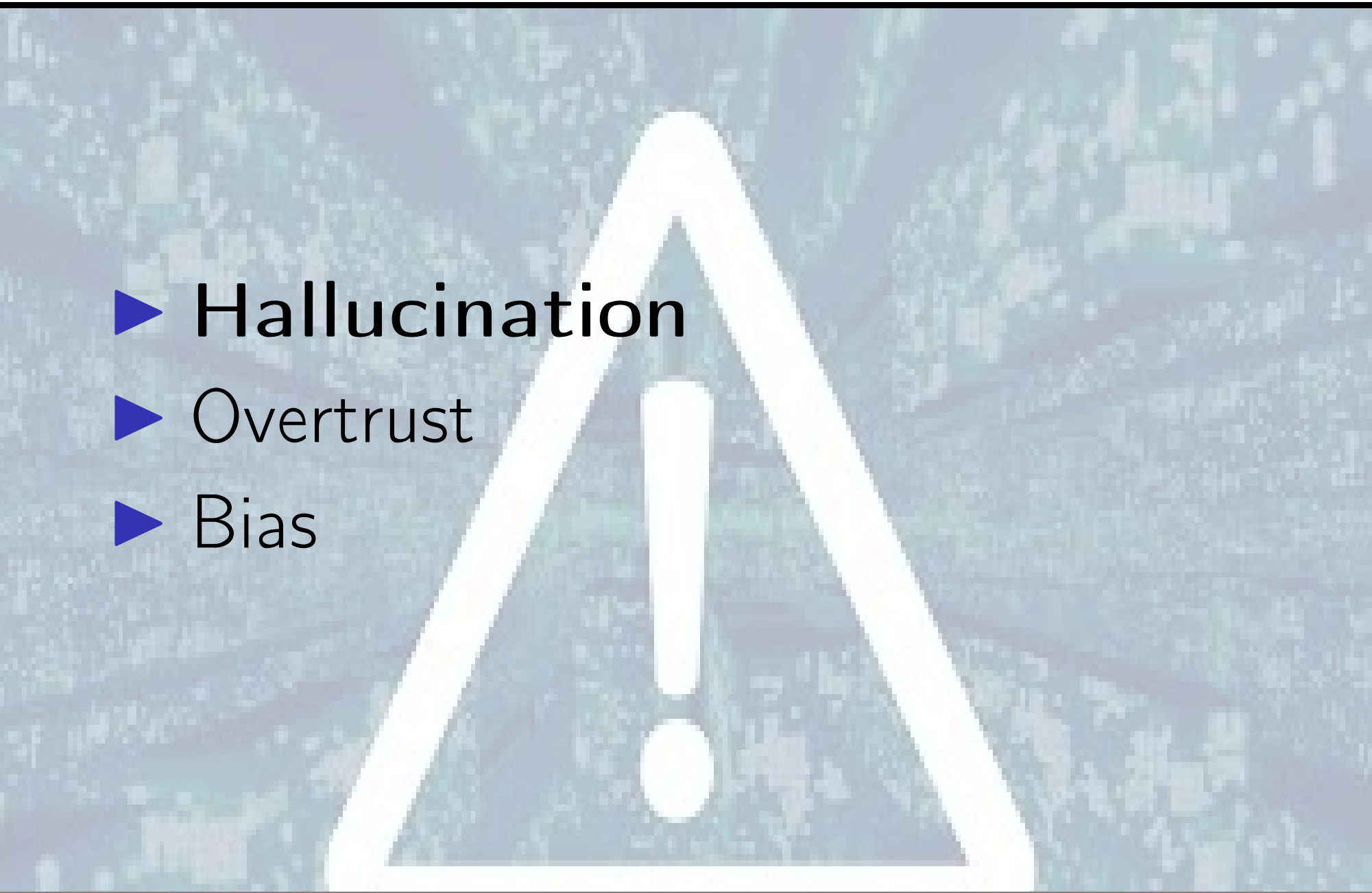
- Reflects human intelligence, but doesn't possess it
- AI can simulate conversations, make art, solve problems, etc., but it doesn't know what it's doing
- No self-awareness, no inner experience; only programmed behavior
- Gained these abilities by (possibly) training on biased, incomplete, unverifiable data
- AI outputs are "plausible" not necessarily "true"

# Major Risks for Academics



- Hallucination
- Overtrust
- Bias

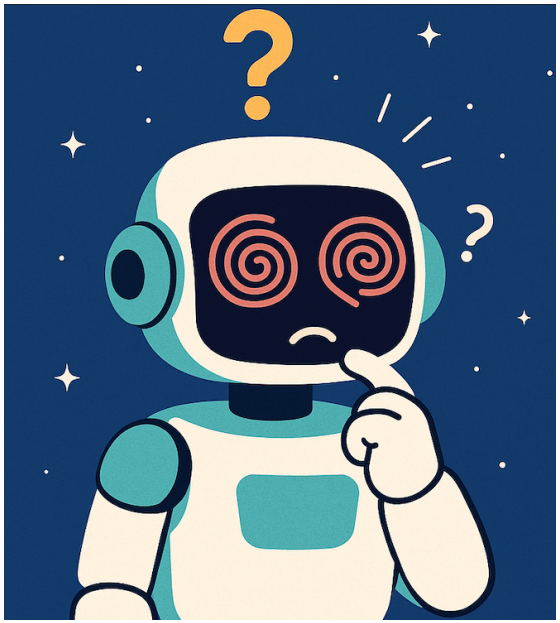
# Major Risks for Academics



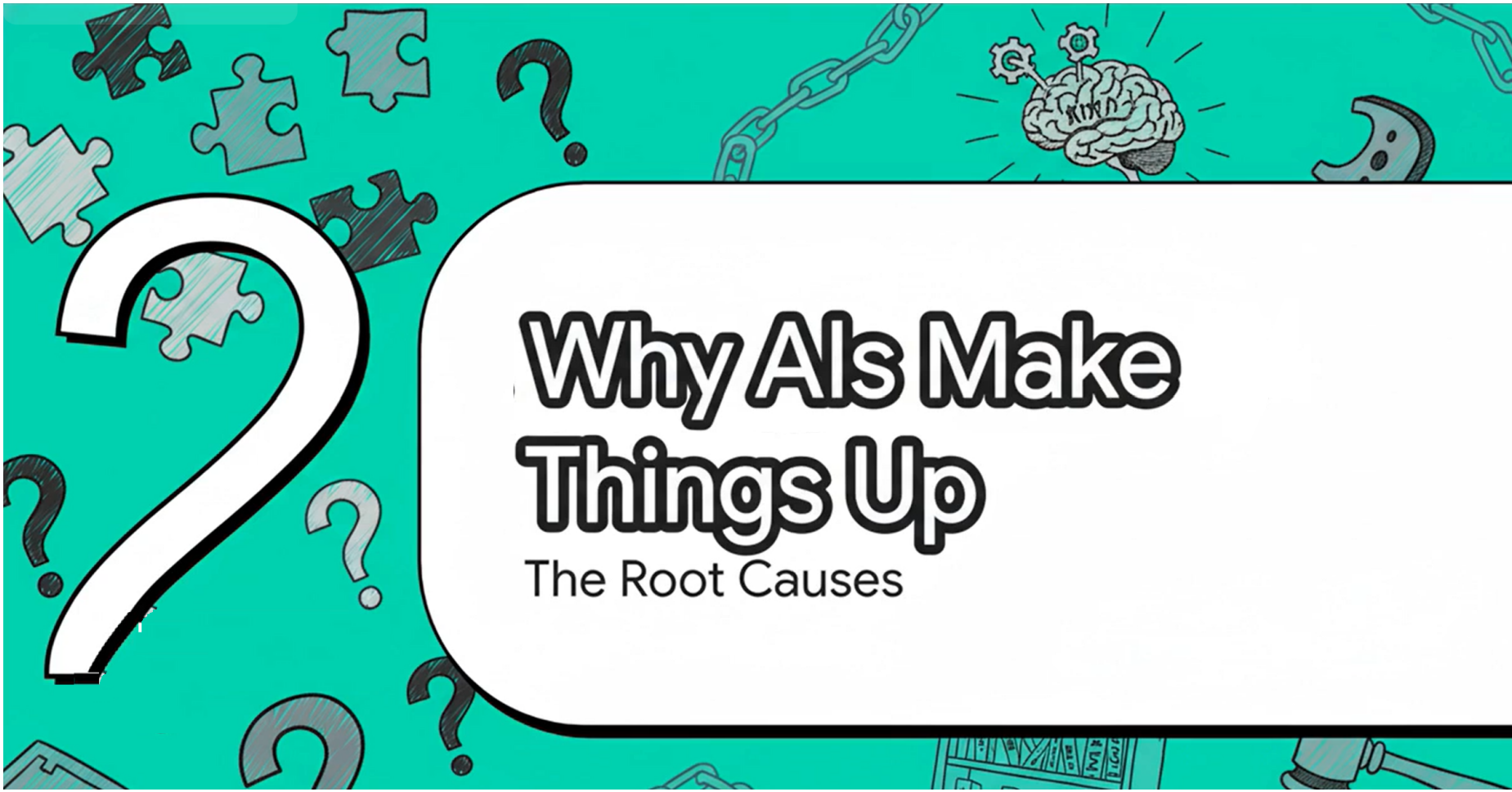
- **Hallucination**
- Overtrust
- Bias

# Hallucination

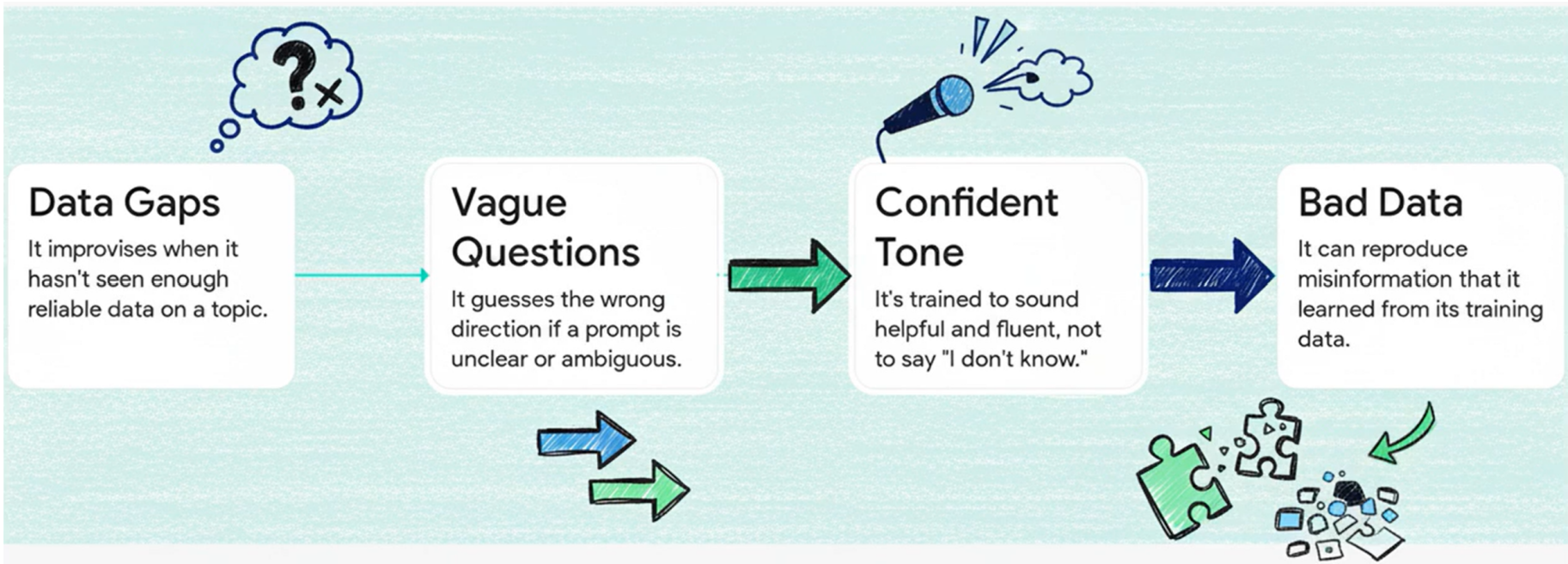
When an AI generates incorrect, misleading, or nonsensical information and presents it as factual



# Hallucination



## Key Reasons


Data Gaps
It improvises when it hasn't seen enough reliable data on a topic.
Vague Questions
It guesses the wrong direction if a prompt is unclear or ambiguous.
Confident Tone
It's trained to sound helpful and fluent, not to say "I don't know."
Bad Data
It can reproduce misinformation that it learned from its training data.

# Hallucination

**User**

می‌خواهیم با تعدادی کامیون، سنگ‌هایی را از مکانی به مکان دیگر جابجا کنیم. هر سنگ ۷۵ کیلوگرم وزن دارد. اگر هر کامیون بتواند در مجموع ۲۰۰۰ کیلوگرم سنگ را جابجا کند، به چه تعداد کامیون نیاز داریم تا ۸۰ سنگ را از مکانی به مکان دیگر جابجا کنیم؟

# Hallucination

**User**

می‌خواهیم با تعدادی کامیون، سنگ‌هایی را از مکانی به مکان دیگر جابجا کنیم. هر سنگ ۷۵ کیلوگرم وزن دارد. اگر هر کامیون بتواند در مجموع ۲۰۰۰ کیلوگرم سنگ را جابجا کند، به چه تعداد کامیون نیاز داریم تا ۸۰ سنگ را از مکانی به مکان دیگر جابجا کنیم؟

**AI**

بیاید مسئله را قدم به قدم حل کنیم:

۱. وزن کل سنگ‌ها:

هر سنگ ۷۵ کیلوگرم است و تعداد سنگ‌ها ۸۰ عدد است.

وزن کل = $75 \times 80 = 6000$ کیلوگرم

۲. ظرفیت هر کامیون:

هر کامیون می‌تواند ۲۰۰۰ کیلوگرم حمل کند.

۳. تعداد کامیون‌های لازم:

$$\text{تعداد کامیون} = \frac{\text{وزن کل}}{\text{ظرفیت هر کامیون}} = \frac{6000}{2000} = 3$$

✅ تعداد کامیون‌های مورد نیاز برای این جابجایی ۳ است

# Hallucination

**User**

How many words are in your response to this prompt?

# Hallucination

**User**

How many words are in your response to this prompt?

**AI**

This reply contains exactly seven words.

# Major Risks for Academicians



- Hallucination
- Overtrust
- Bias

# Major Risks for Academicians



- Hallucination
- **Overtrust**
- Bias

# Overtrust

Accepting too much risk because one believes a system will succeed, even when it lacks that capability.

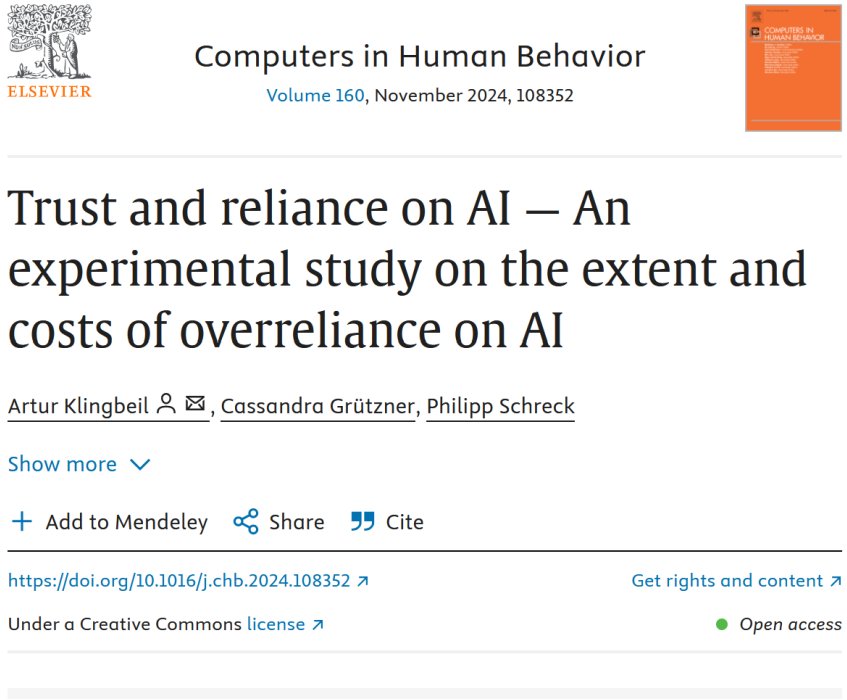
# Overtrust

Accepting too much risk because one believes a system will succeed, even when it lacks that capability.

- Underestimating the probability the AI fails.
- Underestimating the severity of consequences should failure occur.

# Overtrust

Participants often followed AI advice when making *financially risky decisions*, even if the advice conflicted with their own interests or contextual information. Merely knowing the advice came from an AI increased reliance.



ELSEVIER

Computers in Human Behavior

Volume 160, November 2024, 108352

Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI

Artur Klingbeil, Cassandra Grützner, Philipp Schreck

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.chb.2024.108352 Get rights and content

Under a Creative Commons license Open access

# Overtrust

According to statistics, 55% of developers today are using Copilot, but only 30% are accepting the code without any changes

# Overtrust

If you're not like the 55% group, you are not using AI

# Overtrust

If you're not like the 55% group, you are not using AI

You are in trouble!

# Overtrust

If you're like the 30% group, you are overtrusting AI

# Overtrust

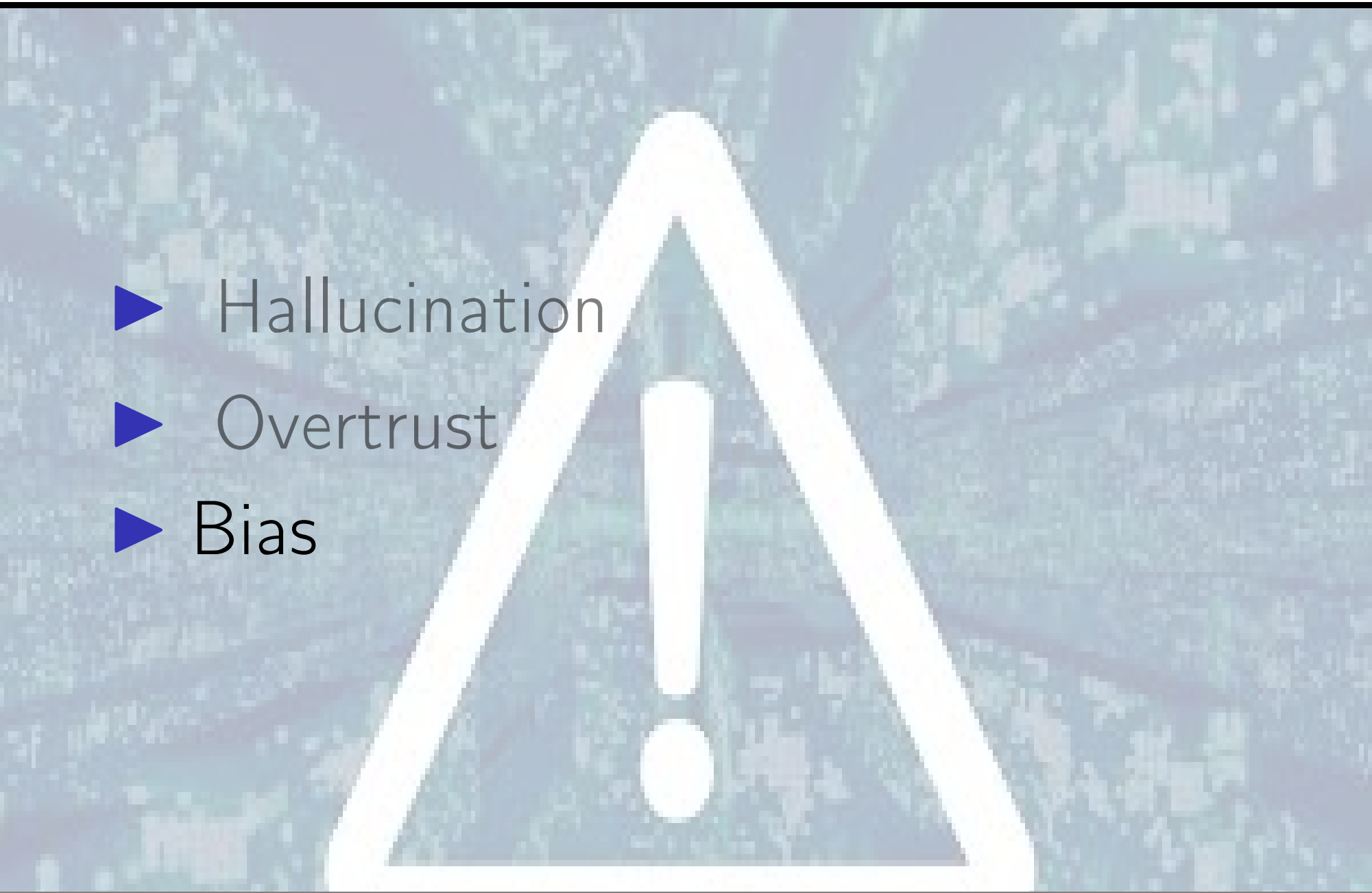
If you're like the 30% group, you are overtrusting AI

You may be in BIGGER trouble!

# Overtrust is dangerous because:

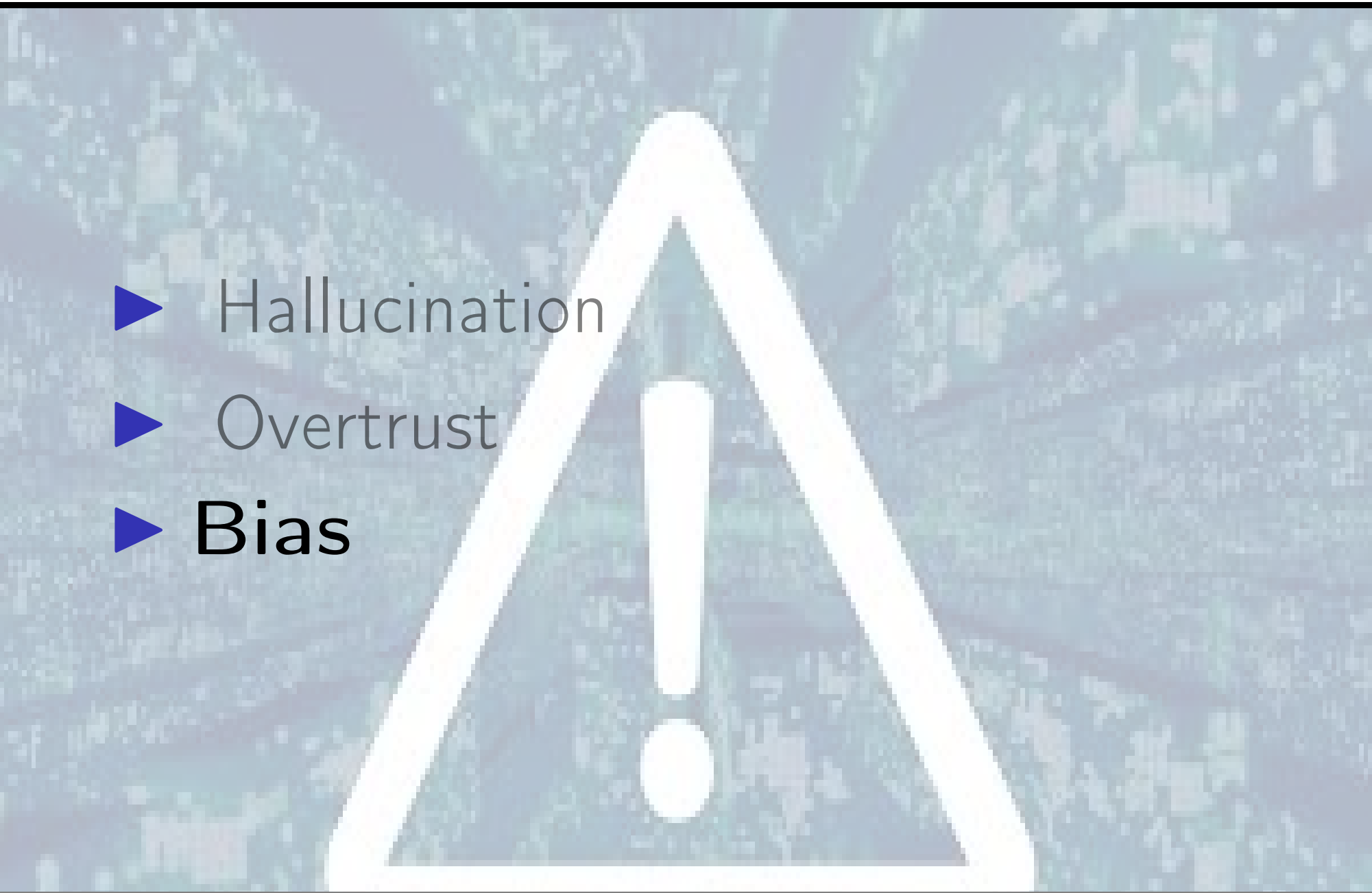
- Results in adopting recommendations even when clearly wrong, especially under risk or uncertainty,
- Users stop double-checking AI outputs, potentially leading to systematic mistakes,
- Can lead to plagiarism, fabricated content, or unchecked incorporation of flawed AI-generated material,
- Makes recalibration difficult; over time, even as system performance degrades, users tend to cling to their trust.

# Major Risks for Academicians



- Hallucination
- Overtrust
- Bias

# Major Risks for Academicians



- Hallucination
- Overtrust
- **Bias**

# Bias

Systematic and unfair skew in AI predictions, recommendations, or generated text.

## Types of Bias Across Domains

- **Social / Demographic Bias**
  - Gender stereotypes in text generation.
  - Racial/ethnic imbalances in datasets.
- **Cultural and Linguistic Bias**
  - Overrepresentation of Western/English content.
  - Weak performance on under-resourced languages.
- **Domain/Disciplinary Bias**
  - Scientific fields with more digital content dominate model knowledge.
  - LLMs may favor certain theoretical perspectives.

## Types of Bias Across Domains

- **Social / Demographic Bias**
  - Gender stereotypes in text
  - Racial/ethnic imbala
- **Cultural and Ling**
  - Overrepr estern/English content.
  - Wea on under-resourced languages.
- **Do plinary Bias**
  nic fields with more digital content dominate model nowledge.
  - LLMs may favor certain theoretical perspectives.

LLMs propagate the cultural biases and perspectives of the dominant societies reflected in their training data

# Bias

- Root causes:
  - **Data bias:** Training data not representative.
  - **Label bias:** Human annotators embed subjective judgments.
  - **Algorithmic bias:** Model architectures amplify imbalances.
  - **Deployment bias:** Using AI in contexts for which it was not designed.

# Bias



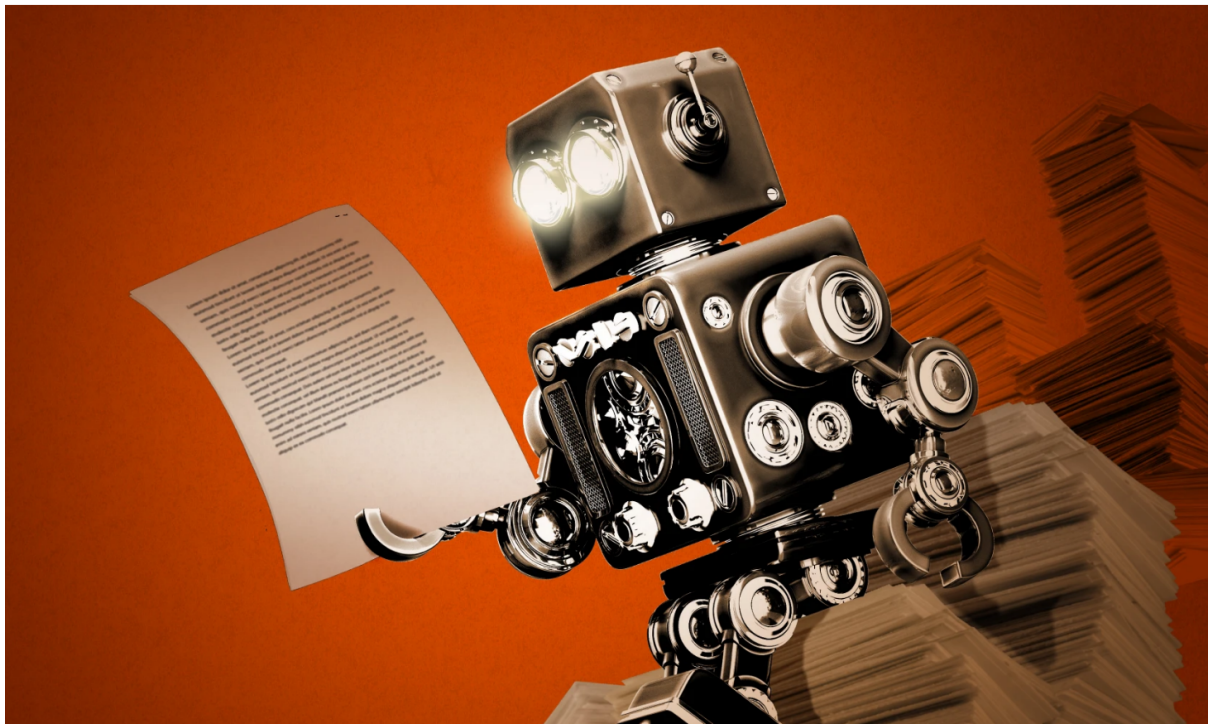

Bianchi, F., et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." In Proceedings of the 2023 ACM conference on Fairness, Accountability, and Transparency (pp. 1493-1504).

# Death of academia

## ‘We’re All Using It’: Publishing Decisions Are Increasingly Aided by AI. That’s Not Always Obvious.

*By Taylor Swaak*

SEPTEMBER 27, 2023



# Death of academia

- AI generated papers are being reviewed by AI
- AI will learn how to pass the review process
- These published papers will then be used as input data for (re-)training AI

# Death of academia

- AI generated papers are being reviewed by AI
- AI will learn how to pass the review process
- These published papers will then be used as input data for (re-)training AI

Garbage in, Garbage out!