👋 Welcome!

Today…

Lecture 8 – ⚙️ Introduction to Generative AI

not assessed

Lecture 9 – 🧪 Code quality 1: Testing and documentation

minor additions to assessed material

! Class Test 1 is this Thursday

Check your timetable to see whether you are in the 3pm or 4:30pm slot.
Both sessions are in AT 6:06.

⚠️ Treat this as an exam. **Do not be late!**

The test will consist of several mini-quizzes.
You need to pass each one to pass the entire test.
That way, you display **mastery** over the material in the course.

📝 **Assessed material:** Lectures 1 – 7 + 9

! Marking

The **Class Test** is automatically marked.
You will see your score immediately.

The **Assignment** is marked by your tutors.
We are aiming to have the marks ready by next Monday, 26th.

If you have passed both, you can move on to Part 2 of the course!
🥳 ...and you will have secured a minimum of 40% in this course.

If not, you will retake the parts you failed in the next assessment slot.

INF1-IP

Lecture 8 – Introduction to Generative AI

Rayo Verweij

19/10/2026

THE UNIVERSITY of EDINBURGH
informatics

> In this lecture

Part 1 – What is AI?

Part 2 – How do large language models work?

Part 3 – What does ‘responsible use’ of LLMs mean?

🎯 **Aim:** Know the technical workings of LLMs enough to understand their strengths and weaknesses, and make **informed decisions** about your own use.

> Defining AI

"I invented the term artificial intelligence... and I invented it because we had to do something when we were trying to get money for a summer study in 1956."



—
John McCarthy *during the Lighthill Debate on Artificial Intelligence, 1973*

https://www.youtube.com/watch?v=03p2CADwGF8 @ 45:58

It worked out for him... but this was not the best start for a rigorous definition!

> Defining AI, with difficulty

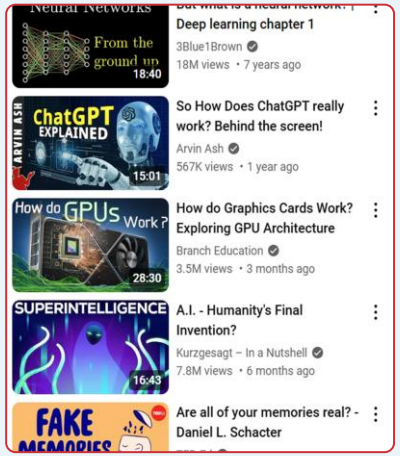
"**AI** comprises any technique that **enables computers to mimic human behaviour** and reproduce or excel over human decision-making to solve complex tasks independently or with minimal human intervention."

—Stuart J. Russell and Peter Norvig. 2021. Artificial Intelligence: A Modern Approach (4th. ed.). Pearson Education. (Emphasis added.)

Okay, but...

- What is human behaviour and what isn't?
- What is a complex task?
- What is *minimal* human intervention?

> Applications of AI



Recommender systems

Voice recognition



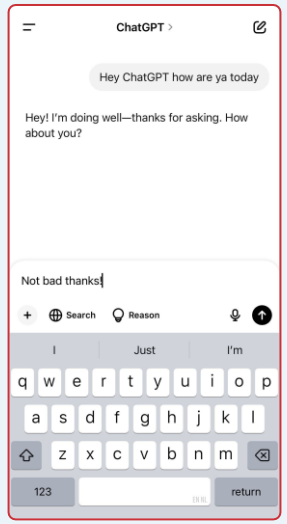
Speech generation

Robotics





Facial recognition



Chatbots

> Subfields of AI

🪧 **Symbolic AI**	Like EDIZA! Uses logic and rules
➗ **Classifying AI**	Drawing boundaries between categories of data
⚡ **Reinforcement learning**	Optimise for a reward signal
✒️ **Generative AI**	Create new content based on training data

Statistical approaches classifying patterns from data are called **machine learning**.
Neural networks are a type of ML system.

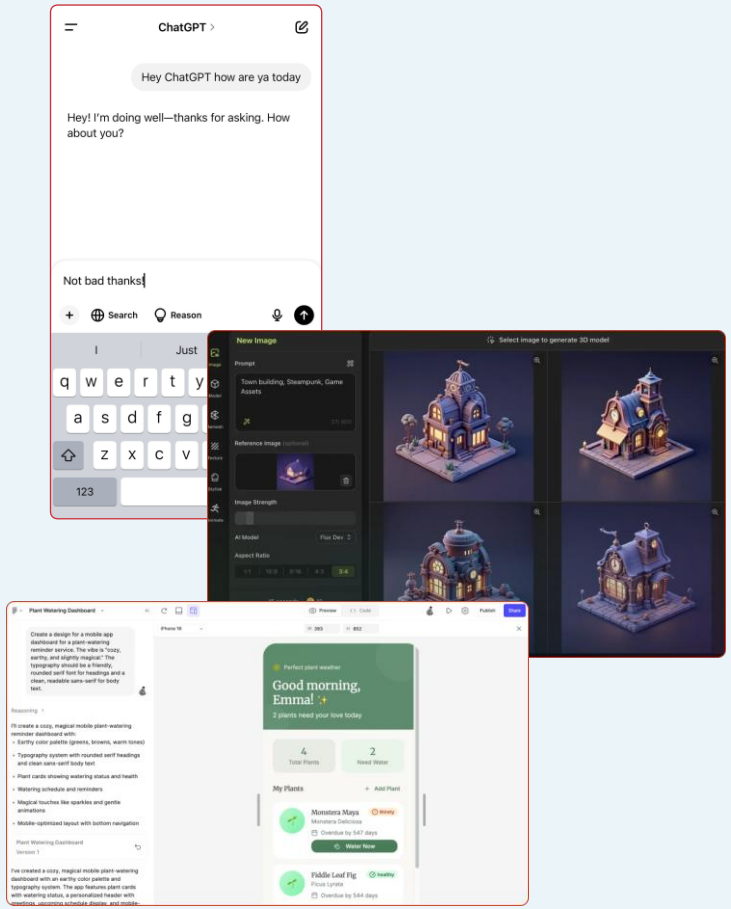
> Summary #1

The field of AI comprises **many techniques and research directions**, loosely united in their ability to mimic human behaviour.

🤔 When evaluating claims about AI, ask...

- What kind of AI do you mean?
- How is it applied?
- Why this type?

> Generative AI



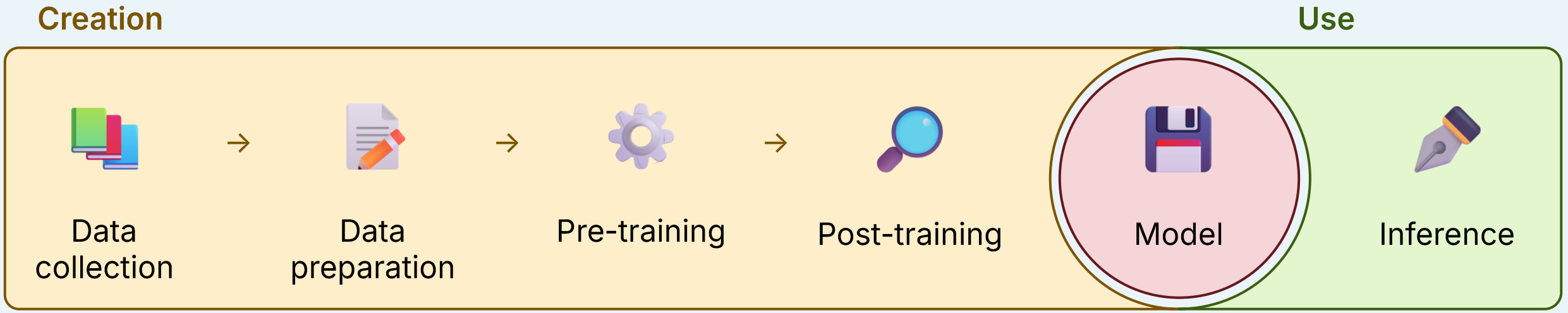
Train **models** that you then use to create new content... such as

- Language
- Vision
- Multimodal → combines modalities

Generally speaking, models are

- *Small* when running on general devices
- *Large* when requiring custom hardware

> Building and using models



> 📚📝 Data collection & preparation

Predicting sensible answers to almost any question needs a lot of examples...

→ *the equivalent of **10s-100s of millions of novels**' worth!*

This data is then:

- 🧹 **Cleaned:** de-duplication, filtering out personal details, etc.

- 🎀 **Tokenised:** broken into short chunks and labelled with numbers
  - Commonly used short words will usually be a token
  - Larger words will be broken up into chunks

...you did some tokenisation with EDIZA!

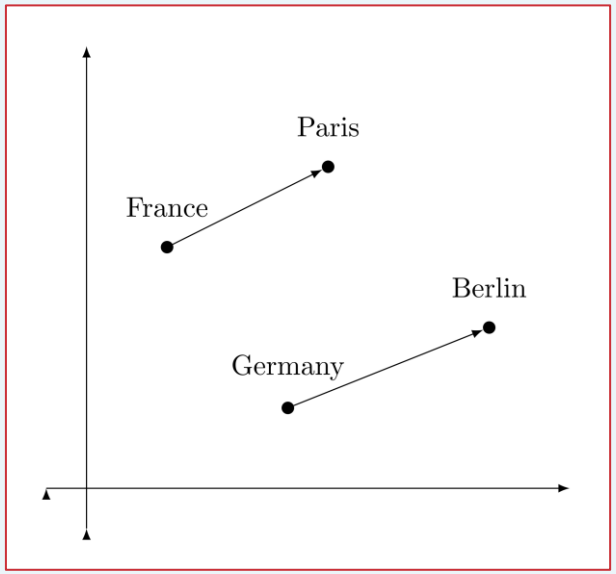
> ⚙️ Pre-training: embeddings

Next, data is 🔢 **vectorised:** each token is turned into an **embedding -**
a long **list of numbers** that encodes **the token and its position**

e.g. `[-0.004, 0.008, -0.016, 0.015, -0.052, -0.016, …]`

↳ often ~1000-2000 numbers long!

Think of an embedding like a point in a graph:



*…except a graph that has **1500 dimensions** instead of 2!*

*It's all **linear algebra…***

Image: Word embedding illustration - Word embedding - Wikipedia

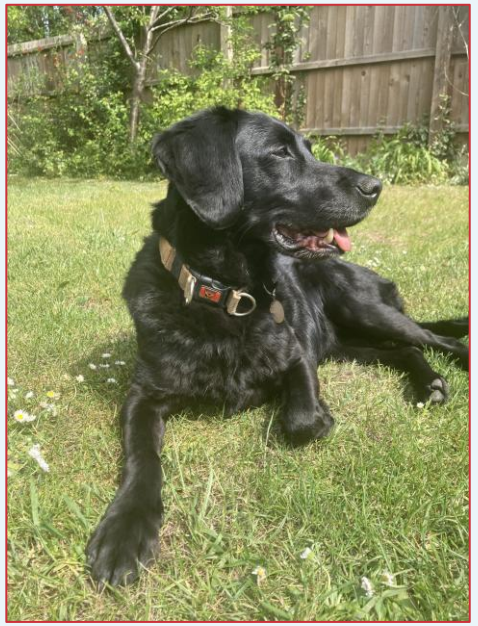
> ⚙️ Pre-training: attention

The embeddings are iterated on again, this time to include information from the **other tokens preceding it.**



This process is called 🔁 **attention**.

Neural networks based on attention are called **transformers**.
This architecture was introduced by Google researchers in 2017.



Thanks James!

> 🔎 Post-training: weights

After this process, we end up with a **massive file** that contains

- 🔢 Our final embedding vectors
- 🔁 The values used in the transformer process

⚖️ Each number – the weights and biases – in this file is a **parameter**.

These are a bit like the parameters of Python functions:

- Each parameter is a float
- There's only several billion of them, nbd

  1B parameters ~ 2GB of memory

> 🔎 Post-training: fine-tuning

Fine-tuning updates our model to **align** them with our desired behaviour.

Fine-tuning can be...

- *automatic*: e.g. 🟰 **supervised fine-tuning**
  → Give the model input-output pairs you want it to follow

- *done by humans*: e.g. 🏅 **reinforcement learning from human feedback**
  → Model generates several responses and humans rank them

> 💾 Finally... we have a model!

Common model families...

 GPT by OpenAI

 Claude by Anthropic

 Grok by xAI / SpaceX / SpaceXSI / whatever it's called next week

 LLaMA by Meta open-weight

 Qwen by Alibaba open-weight

> 🖋️ Inference: probability scores

Inference is the process of predicting the next token given certain input.

The program goes over **every token in the model** and calculates the probability that it will come next.

```
Sarang  was  eating  a
```

treat	42%
bisc	17%
warm	11%
…	…%
spoon	0.001%



Thanks Ash!

> 🖋️ Inference: temperature

Always choosing the token with the highest likelihood would mean you always get the same answer for a question.

To vary our responses, we introduce **randomness** in the selection.

Randomness is controlled by a setting called 🌡️ **temperature**.

- Temperature = 0 ⇒ always choose the most likely token
- Temperature > 0 ⇒ sometimes choose a less likely one

> 🖋️ Inference: sampling

To keep our responses somewhat likely, we **sample** the most-likely next tokens and choose from them.

Top-k sampling: select the *k* most likely tokens and choose from them.

Top-p sampling: select the tokens which together add up to *p*% likelihood and choose from them.

> ✒️ Inference: system prompt & harness

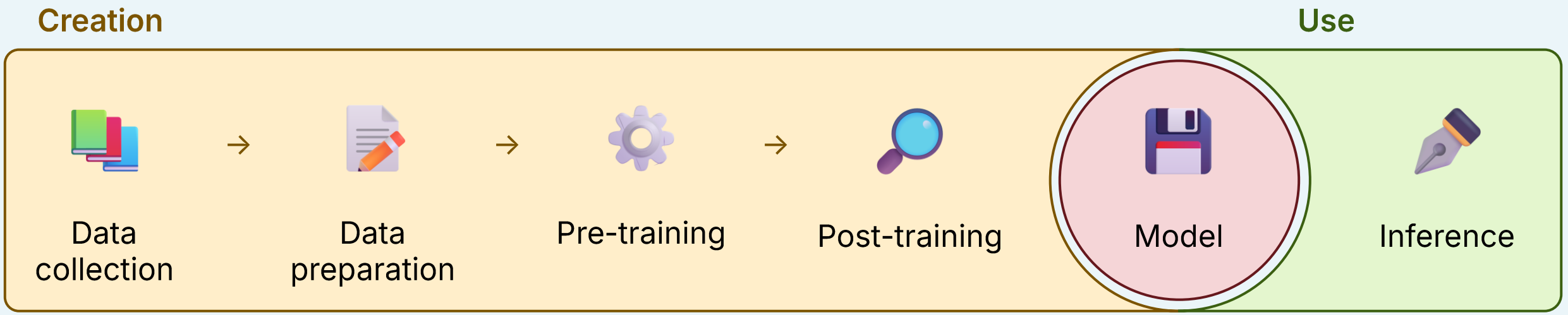
In addition to your prompt, the model will always have a **system prompt** predefined that gives meta-instructions.

On top of the model, we can build a 👛 **harness**:
'regular' programs that extend the capability of the software

- Web search, file access, calculator, etc.

Agents are tools built around LLMs that perform specific tasks.

> Understanding impact



> Copyright infringement

If models need the equiv. of 10s-100s of millions of novels to be trained...

where does the data come from?



AI firm Anthropic agrees to pay authors $1.5bn to settle piracy lawsuit

Lily Jamali

North America Technology correspondent in San Francisco

5 September 2025

Artificial intelligence (AI) firm Anthropic has agreed to pay $1.5bn (£1.11bn) to settle a class action lawsuit filed by authors who said the company stole their work to train its AI models.

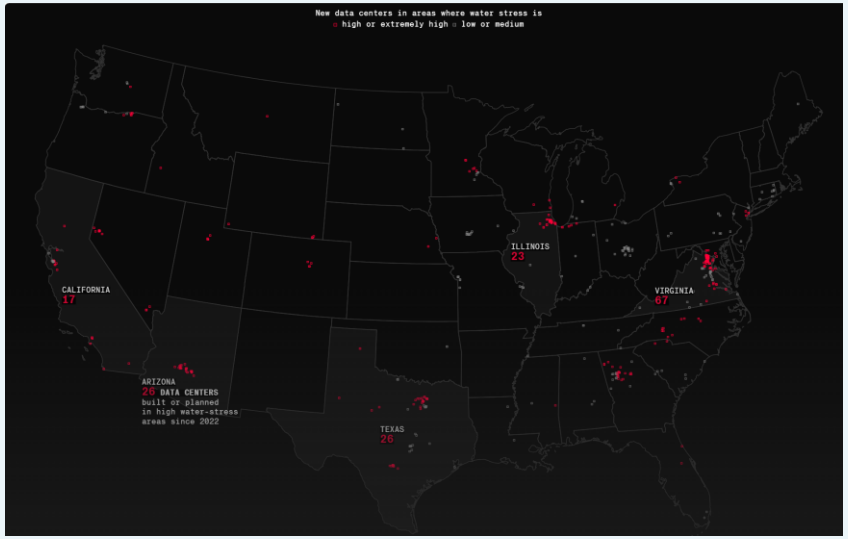
https://www.bbc.co.uk/news/articles/c5y4jpg922qo

> Resource use

Both training and inference require **many calculations at once**.

GPUs are particularly suitable for that work.

These need 🔌 **electricity** to run and (often) 💧 **fresh water** to cool.



*Exact impact depends on location – in the USA, ~2/3 of data centres built since 2022 are in a **water-stressed** region.*

https://www.bloomberg.com/graphics/2025-ai-impacts-data-centers-water-data/

> Exploitation of workers

Processes like data labelling and RLHF are **very labour-intensive** and tend to be outsourced to (mostly) East African data firms.



Made in Africa: Exploiting the bodies and minds behind Artificial Intelligence

Mophat Okinyi, Richard Mathenge, Mohammad Amir Anwar

School of Social and Political Science

Research output: Contribution to journal › Article › peer-review

Overview Fingerprint Projects (1)

Abstract

As automation threatens to displace human across many sectors, artificial intelligence has simultaneously created new development prospects through the provisions of data work. For over a decade, major Silicon Valley technology firms have outsourced this work through complex global value chains to different parts of the globe, including East Africa. Drawing on survey and in-depth interview with data workers in Kenya, complemented by an auto-ethnographic account from two former data workers-turned labour activists at major data enrichment firm in Nairobi, this paper examines how data work value chains shape workers' lives and livelihoods. It argues that data work value chains constitute the latest space for algorithmic exploitation, in which poorly paid and precarious labour form the hidden underbelly of the Silicon Valley. This comes at significant costs to workers' physical and psychological well-being. The paper makes two scholarly contributions. First, it redirects data work value chains scholarship towards the embodied and psychological toll of this work. Second, it advances, AI for development literature on Africa by demonstrating that East Africa's integration into AI production networks does not primarily represent development opportunity but a new site of dependency and extraction reproducing long-standing inequalities in the global political economy.

Other files and links

https://www.tandfonline.com/journals/rjea20

Embargoed Document

OkinyiEtalJEAS2026MadeinAfrica

Accepted author manuscript, 629 KB

Licence: Creative Commons: Attribution (CC-BY)

"poorly paid and precarious labour form the hidden belly of the Silicon Valley"

https://www.research.ed.ac.uk/en/publications/made-in-africa-exploiting-the-bodies-and-minds-behind-artificial-/

> Incorrect output

Why do models generate **incorrect output** (“hallucinations”)?

→ Mistakes in training data

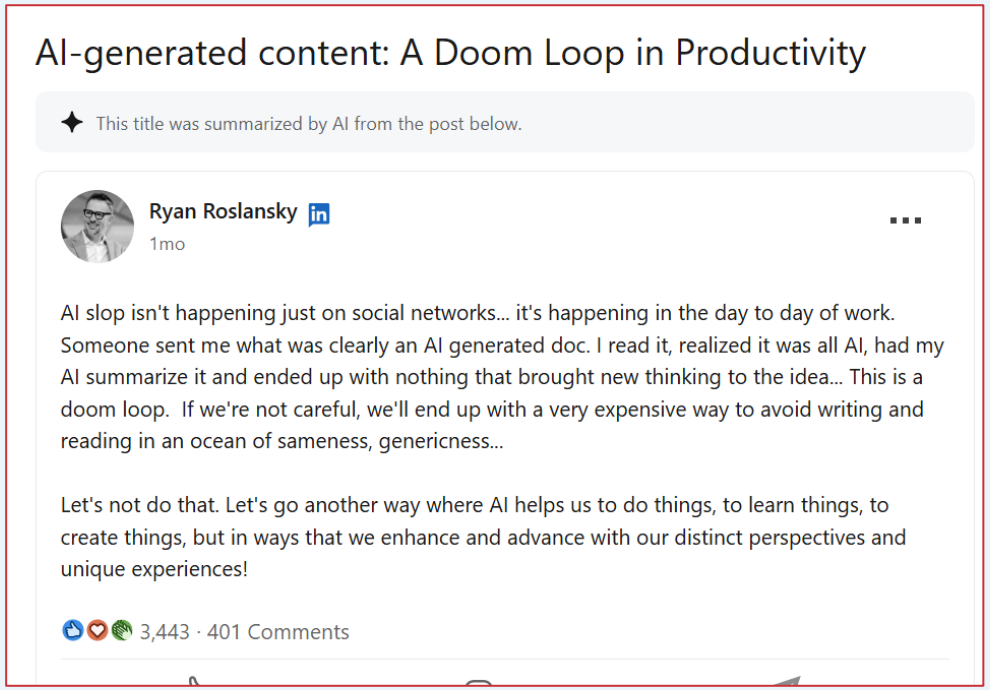
→ Random sampling during inference

→ Fine-tuned per personal preference

Plus: even if a model accurately represents our world, **do we want that?**

> Cognitive offloading

What is the impact on **yourself** when you outsource your thinking / reading / writing to a machine?



https://www.linkedin.com/posts/ryanroslansky_ai-slop-isnt-happening-just-on-social-networks-share-7496622258124152832-VBXc/
via
https://www.theguardian.com/technology/2026/oct/13/ai-companies-rushed-workers-fallout

> Abuse

Elon Musk's Grok AI appears to have made child sexual imagery, says charity



Chris Vallance
Senior technology reporter

7 January 2026
Updated 8 January 2026

The Internet Watch Foundation (IWF) charity says its analysts have discovered "criminal imagery" of girls aged between 11 and 13 which "appears to have been created" using Grok.

https://www.bbc.co.uk/news/articles/cvg1mzlryxeo

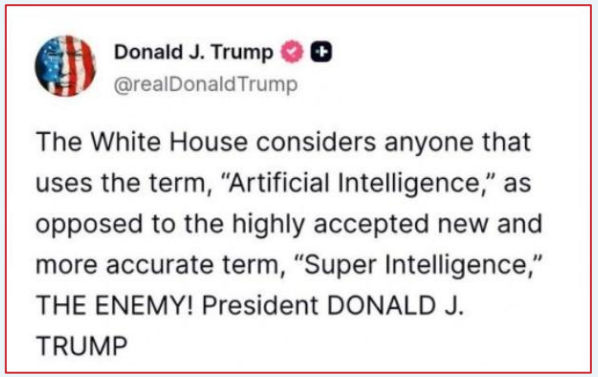
In a model with dozens of billions of parameters, where we are fundamentally incapable of understanding what each one does…

*…is it ever **actually possible** to create a model that is fully 'safe'?*

*And **who decides** what is safe and what isn't?*

> AGI and the end of the world





Well then, is AI going to take over the world?

Can unrestricted agents cause real harm?

Yes, absolutely, and LLM use needs to be strictly regulated.

Is AI going to develop consciousness and decide to kill us all?

LLMs take input and predict the next most likely token.

They have no intent, subjective experience, or consciousness.

> Summary #2

🤔 When evaluating claims about AI, ask...

- What kind of AI do you mean?
- How is it applied?
- Why this type?

- **Where do they get their money from?**

> Can we do this responsibly, then?

Can we not build models that...

- ...are only trained on public-domain data?
- ...properly compensate all human workers involved?
- ...only use clean energy and limited fresh water?
- ...are finetuned for safety as best as we can?

To some extent...

> Models on ELM

Commercial models

- Zero data retention → companies do not see your queries
- Still use their data & resources
- Not all functionality (e.g. image creation)

Local models

- Also zero data retention
- Run on University infrastructure: clean power + no water use!

> Should you use LLMs...

...never? ‘Within reason?’ Only through ELM? As much as you want?

This is ultimately a **personal choice**.

However, in a world that is increasingly artificial...

- ✍️ **Develop your writing & thinking** to stand out
- 🧠 Understand your **brain as a muscle** that needs practice
- 👩‍🎨 Find what **you enjoy** and do it for the love of it
- 🎤 Develop **your own voice** and unique perspective on the world