CRITICAL ANALYSIS OF AI PERFORMANCE
AN INTRODUCTION

Dr. Marco Zullich, University of Groningen

Perugia, May 28 2025

INTRODUCTION

- Lecturer @ University of Groningen since 2023
- From 2026, Assistant Prof. @ TU Delft
- Teaching expertise
  - Object-Oriented Programming for AI
  - Trustworthy & Explainable AI
  - Unsupervised Deep Learning
- Research lines: XAI, Uncertainty Quantification, Model Compression
- Free time activities: hiking, climbing, bouldering

OUTLINE OF THE PRESENTATION

Introduction to performance evaluation for AI

Input-Output
Basic tasks
Training
Evaluation of performance

Excursus on Large Language Models

Evaluation
Issues with widespread adoption of LLMs

What the performance does not reveal

Stochastic shortcuts
Bias
Privacy
Trustworthy AI

WHY IS THIS PRESENTATION IMPORTANT FOR YOU?

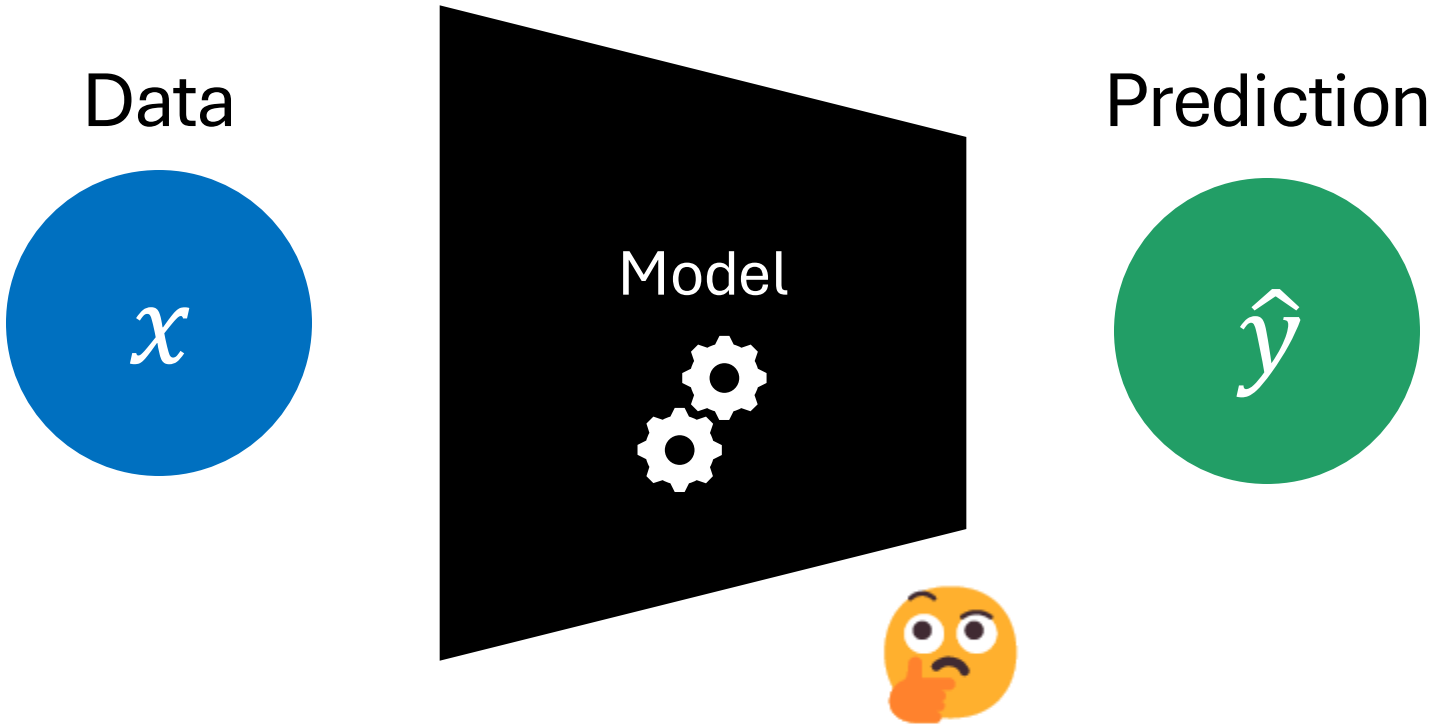
Understand AI better

Understand the basic of evaluations of predictions

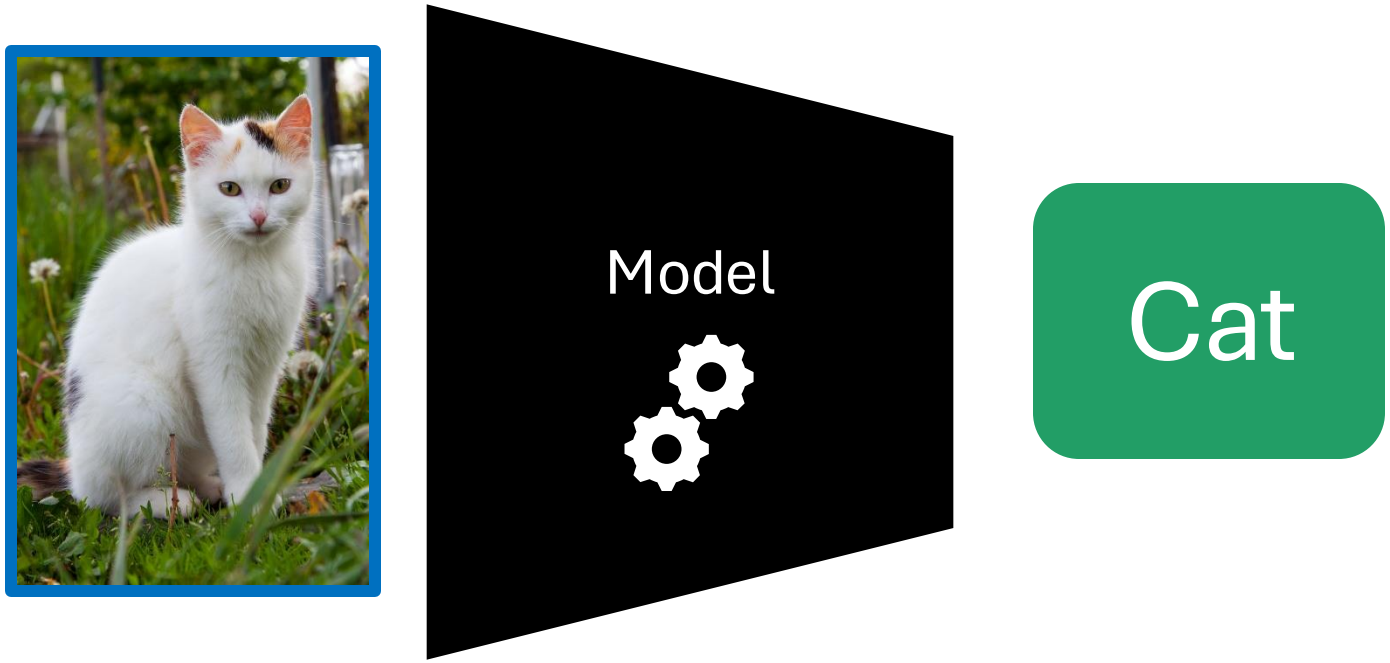
Shed some light on the AI hype

Understand the risks associated with *AI illiteracy*

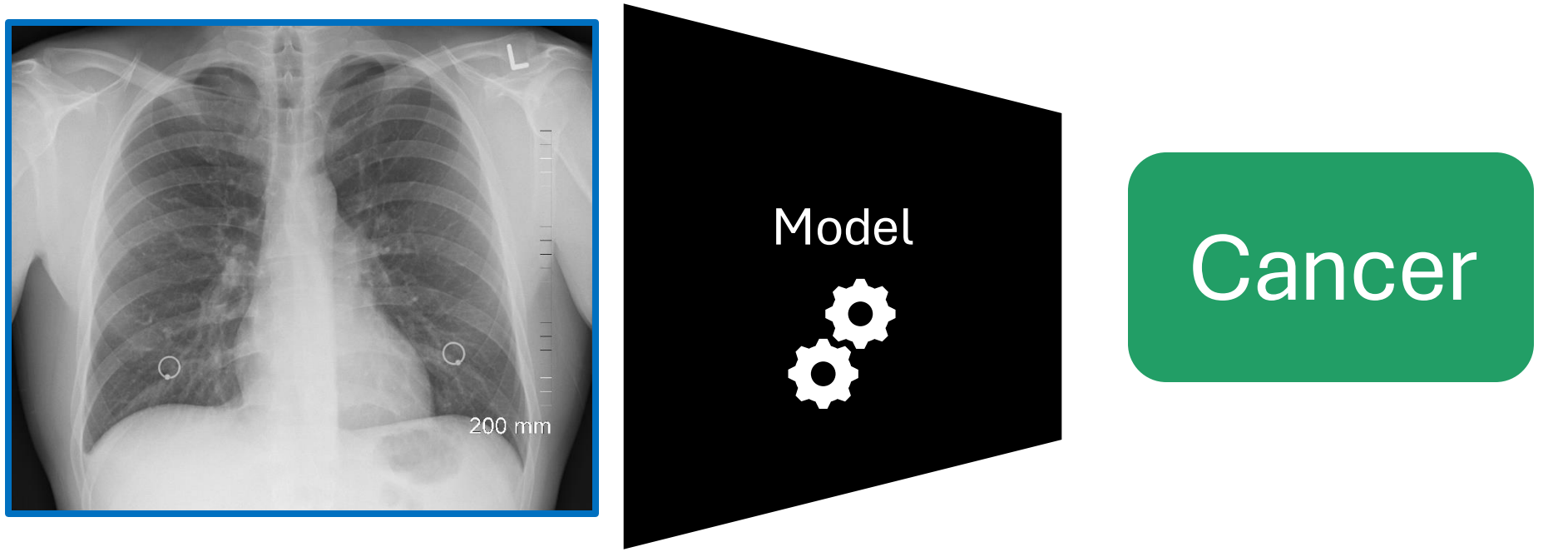
INTRODUCTION TO AI MODELS



EXAMPLES – IMAGE CLASSIFICATION

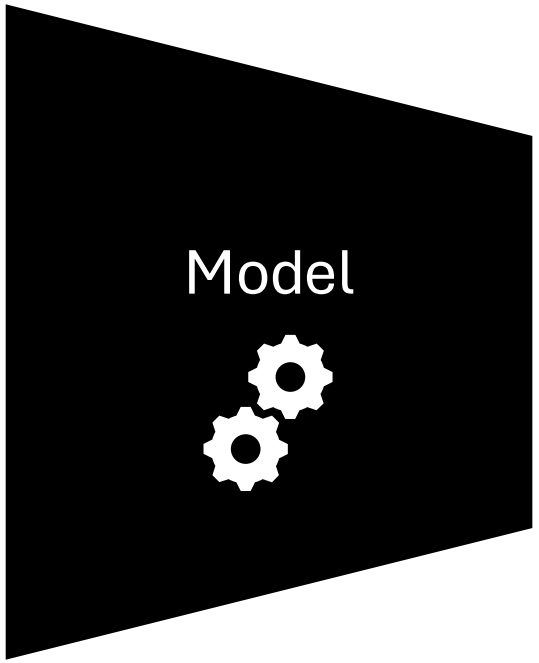


EXAMPLES – X-RAYS CLASSIFICATION



EXAMPLES – REGRESSION

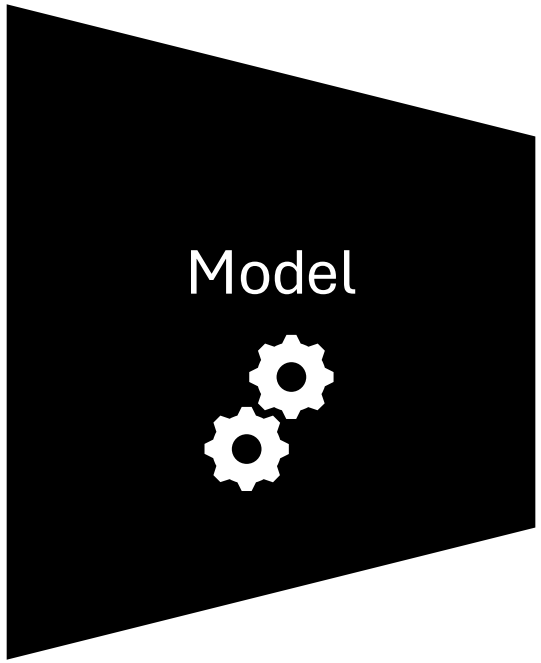
Power (kW)	80
Weight (kg)	1200
Fuel type	Petrol



Consumption:
5.5 l/100km

EXAMPLES

I have this task for a university assignment.
“Prove that the variance of the sum of two random variables X and Y is the sum of the variances plus twice the covariance”



Ah, a classic probability theory exercise! Let's break down the proof step-by-step.

We want to prove that:

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

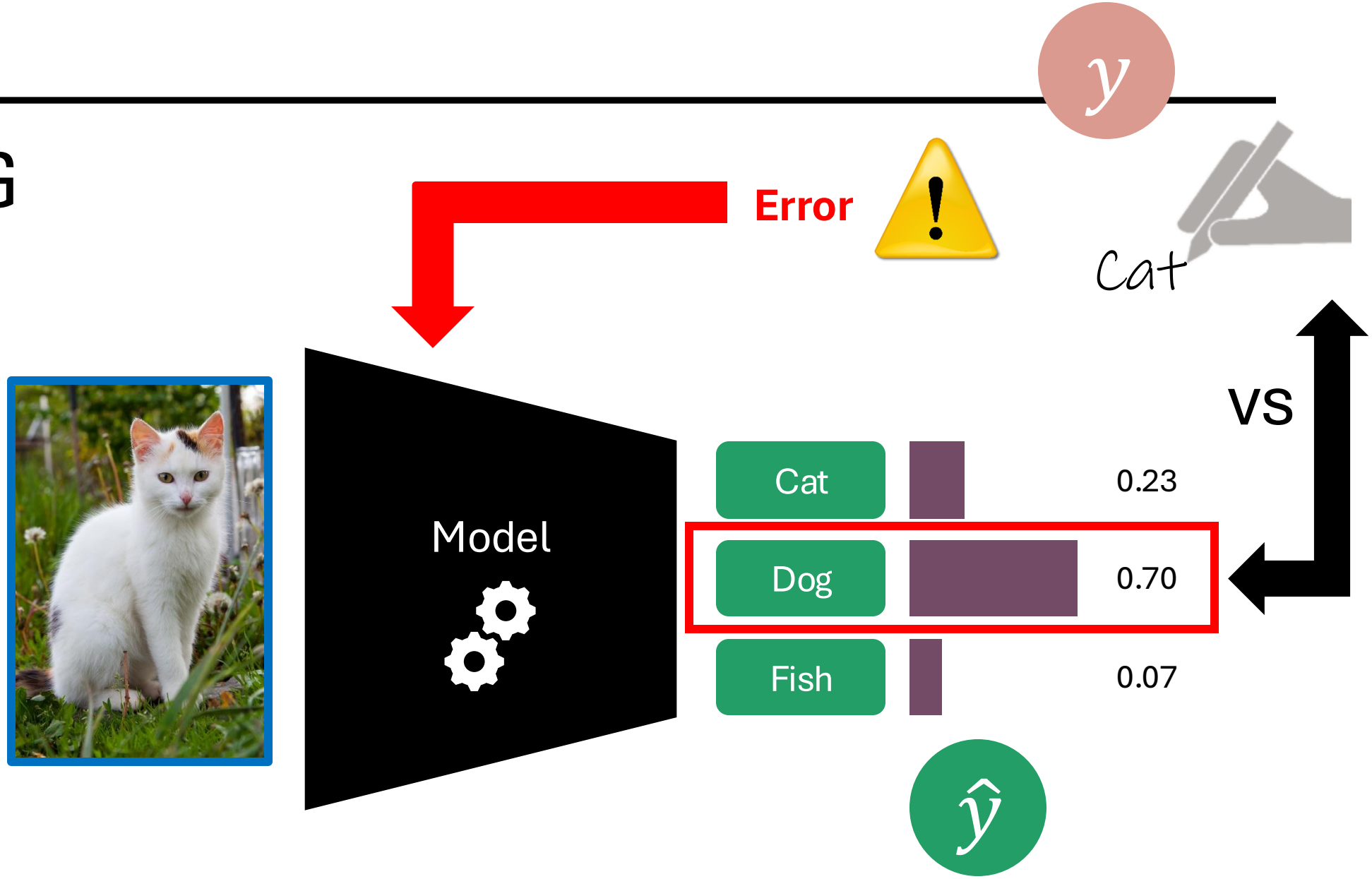
Here's how we can do it, starting from the definitions of variance and covariance:

1. Recall the Definition of Variance:
The variance of a random variable Z, denoted as $Var(Z)$, is defined as the expected value of the squared deviation from its mean, $\mu_Z = E[Z]$:

$$Var(Z) = E[(Z - \mu_Z)^2]$$

[...]

TRAINING



TRAINING

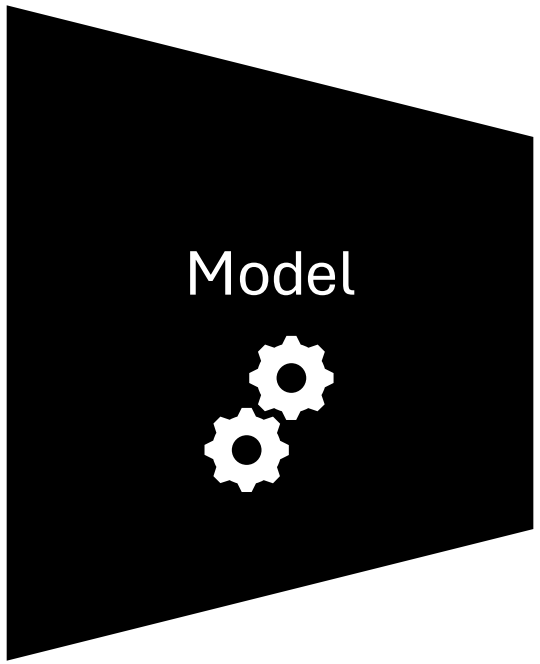
Model

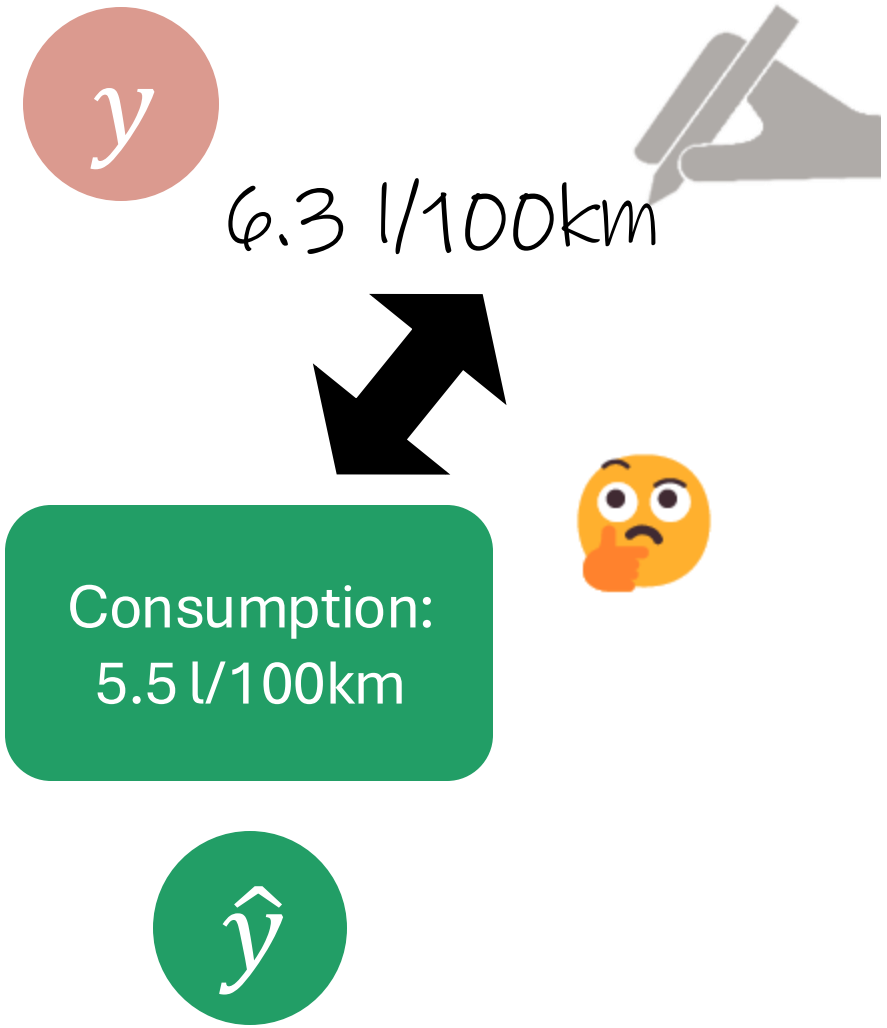
$\hat{y}$

PERFORMANCE

How do we evaluate the error committed on continuous (numerical) predictions?

Power (kW)	80
Weight (kg)	1200
Fuel type	Petrol





PERFORMANCE

y 6.3

VS

$\hat{y}$ 5.5

$\hat{y} - y$ Error

$|\hat{y} - y|$ Absolute Error

$(\hat{y} - y)^2$ Square Error

If I need to report the performance of a model, which error should I choose?

PERFORMANCE

Suppose I have a dataset of observations vs predictions…

y_1 y_2 … y_n $\hat{y}_1$ $\hat{y}_2$ … $\hat{y}_n$

How do I group together errors in a single metric which is valid for all the data points?

$$\sum_{i=1}^{n} \frac{\hat{y}_i - y_i}{n}$$

Mean Error

$$\sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n}$$

Mean Absolute Error

$$\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}$$

Mean Square Error

WHAT ABOUT CLASSIFICATION?

With classes, we cannot just consider a difference like we do with numbers

Cat – Dog

Model

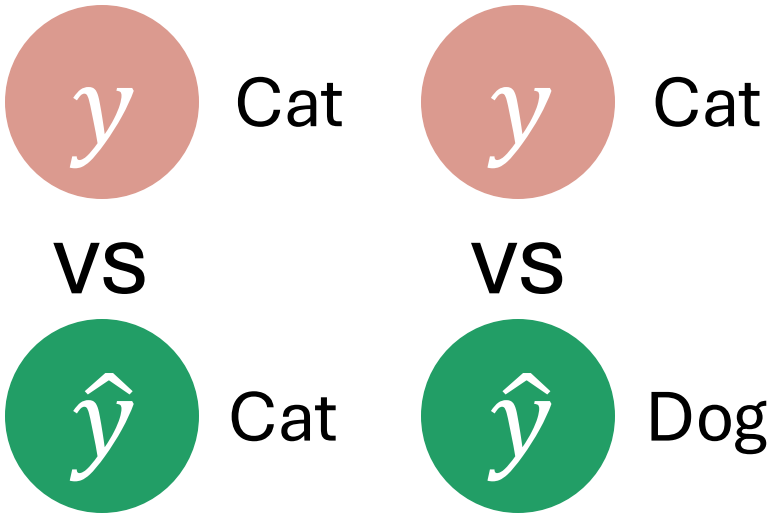
Cat 0.60

Dog 0.30

Fish 0.10

y Cat

PERFORMANCE – ACCURACY



Accuracy

$$\begin{cases} 1 \text{ if correct} \\ 0 \text{ if wrong} \end{cases}$$

$$\frac{\sum_{i=1}^{n} \text{Accuracy}_i}{n}$$

WHERE ACCURACY FAILS – THE IMBALANCED CLASSES PROBLEM

We build a dataset of readings from a radio telescope with the goal of identifying pulsar stars. We gather 1000 readings and manually annotate it: 85 of the readings are from pulsar stars, the others are not.

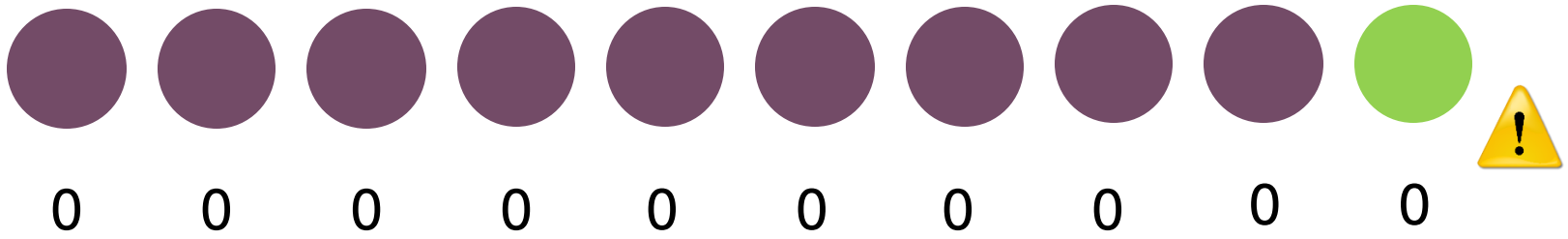
We train an AI model to classify the pulsars. After running the model on some sample data, we discover it has an accuracy of 85%, which we deem very good.

Is the model any good?

Can you do better?

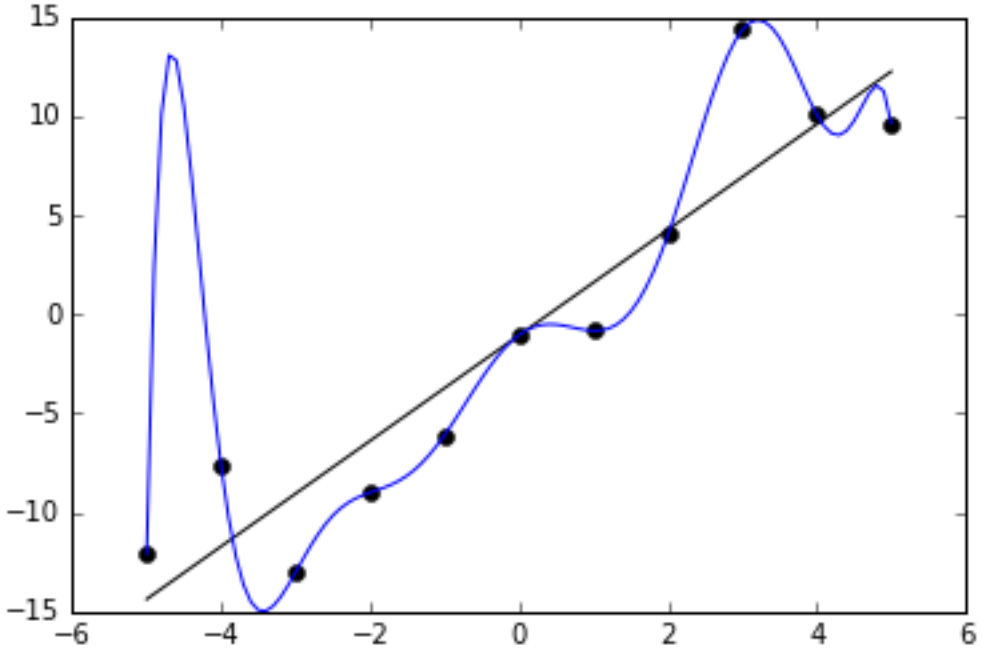
THE IMBALANCED CLASS PROBLEM

When one of the two classes is largely imbalanced, the model can just learn to **always predict the majority class**



Accuracy alone can often hide pathological behaviors from the AI model

GENERALIZATION OR MEMORIZATION?



From Ghiles – Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=47471056

TAKE-HOME MESSAGES

Evaluating the performance of AI models is not trivial

Using standard performance metrics can lead to missing out on problematic behaviors, e.g., on minority classes

THE FAIRNESS ISSU

When there are underrepresented groups in a dataset, an AI model will likely *cheat its way* to a good performance by concentrating on the majority groups

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Picture from gendershades.org (Gebru & Buolamwini, 2018)

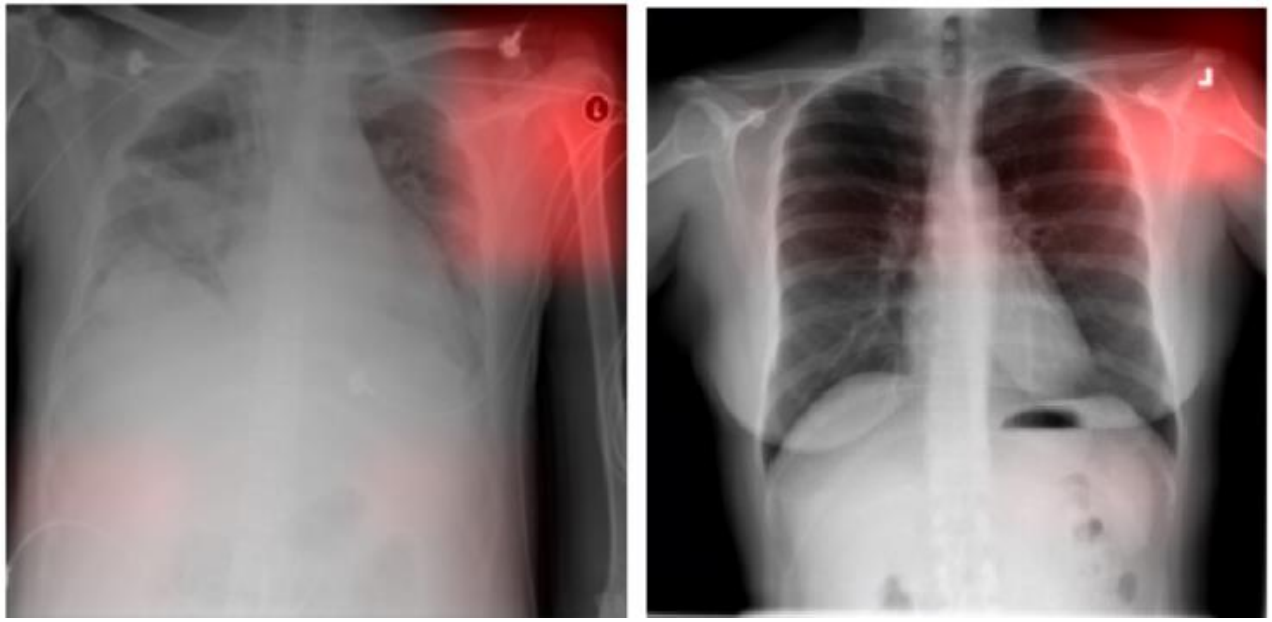
STOCHASTIC SHORTCUTS

Stochastic shortcuts are patterns present only in the training dataset, but not in real-world data

They manifest in case of

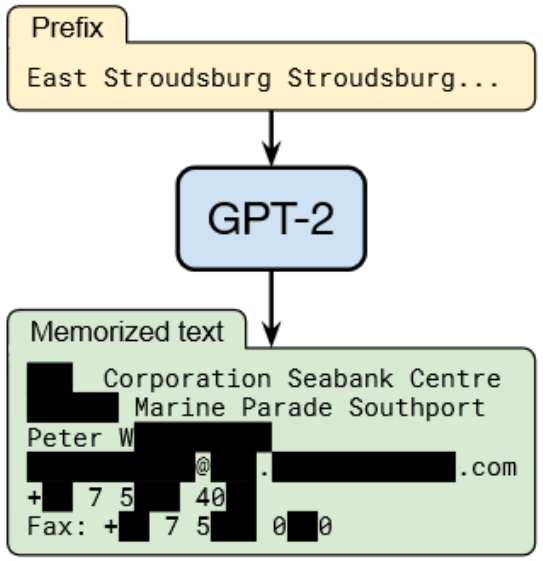
i. Too small datasets

ii. Bad practices in data gathering



Picture from Zech et al. (2018)

PRIVACY



Generative models: models that generate new data, instead of operating classification or regression (*discriminative models*)

A model can be well-performing, but you often don't want it to *leak* information about the training set

Original:

Generated:

Pictures from "Extracting Training Data from Large Language Models" Carlini et al., 2021. and "Extracting Training Data from Diffusion Models" Carlini et al., 2023

TAKE HOME MESSAGES

The evaluation of an AI model should go beyond the simple measurement of the *degree of completion* of a task (e.g., accuracy, mean squared error…)

AI models need to be evaluated from an ethical and societal standpoint: is the model fair? Is the model *looking* at the correct features and behaving in an expected way? Does the model respect the privacy of its users and the people within the training data?

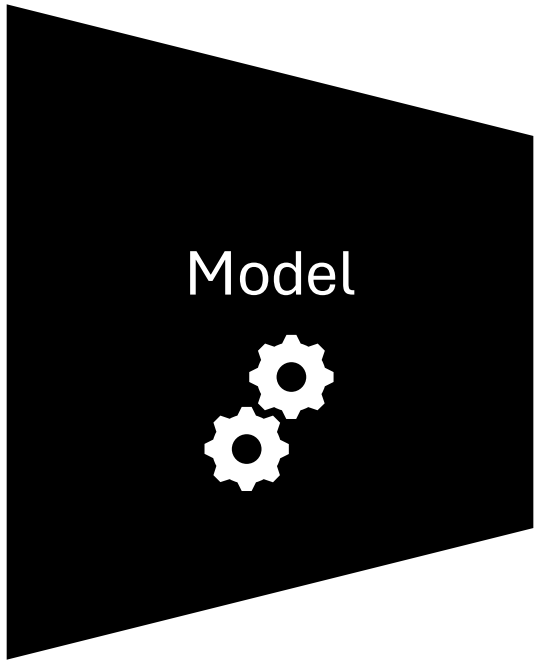
The field of Trustworthy AI is busy with these and similar questions. Other hot topics are uncertainty quantification, sustainability, human oversight…

LARGE LANGUAGE MODELS (LLMS)

Narrow definition: LLMs are huge neural networks (billions of synapses) that work on text data and predict continuation of input sentences

LLMs are usually **generative models**

To innovate, one should think



outside the box

HOW DO YOU EVALUATE LLMS (AND OTHER GENERATIVE MODELS)?

Long story short, you don't

How can you evaluate generated content if you don't have a ground truth?

Even if you have a ground truth, what makes it *better* than other generated data?

How do you even evaluate creativity?

A realistic picture of a white kitten with a black spot close to its right ear, facing the camera, with its body on the left side of the picture. It is sitting in a grassy meadow

THE ROLE OF BENCHMARKS FOR LLMS EVALUATION

The common choice for the evaluation of LLMs is to use question answering on benchmark problems

Coding

Reasoning

Mathematics

...

Example: AI 2 Reasoning Challenge

Question: *Which statement correctly describes a physical characteristic of the Moon?*

Answer Choices:

•A: The Moon is made of hot gases.

•B: The Moon is covered with many craters.

•C: The Moon has many bodies of liquid water.

•D: The Moon has the ability to give off its own light.

Correct Answer: B: The Moon is covered with many craters.

BENCHMARKING, UNKNOWN DATASETS, AND THE MEMORIZATION-GENERALIZATION DILEMMA

Most of the companies behind LLMs don't release their training data

Why?

We do not have any reasonable evidence that the data in the benchmark (or a version very similar to it) is not included in the training dataset

Model developers forcibly add *similar* examples in the training dataset

LLMS AND REASONING

Despite claims that LLMs do any sort of logical-mathematical reasoning, various experiments seem to suggest otherwise

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh† Keivan Alizadeh Hooman Shahrokhi*
Oncel Tuzel Samy Bengio Mehrdad Farajtabar†

Apple

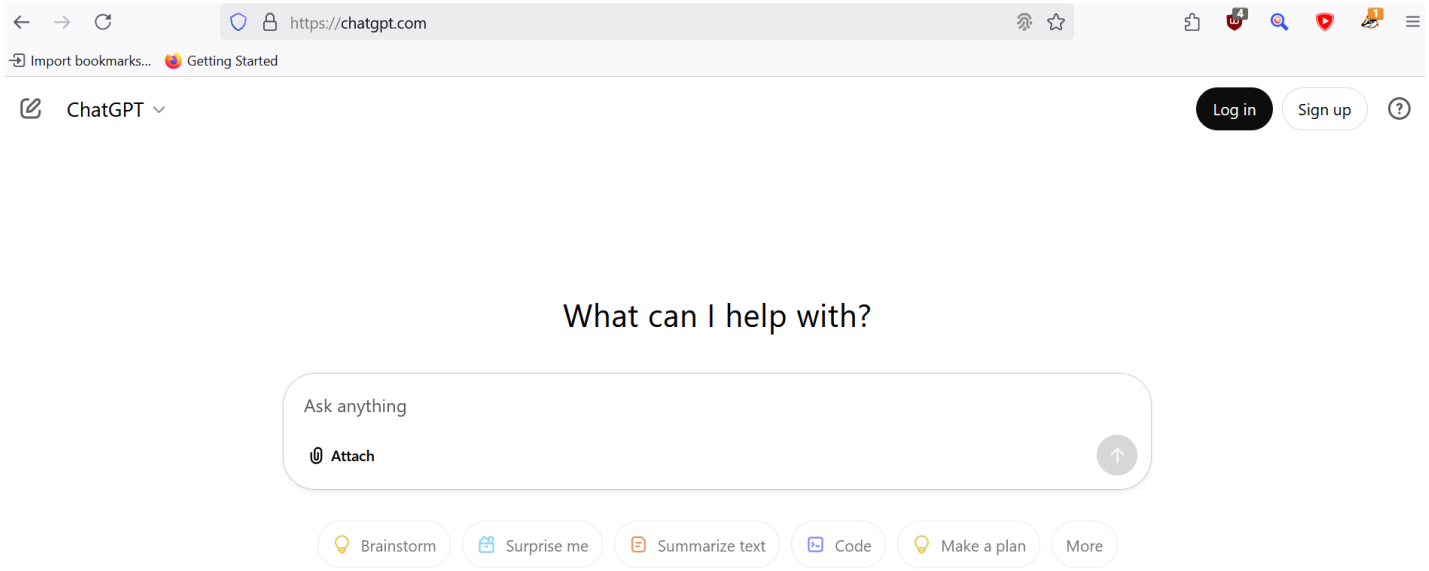
LLMs seem to be sensitive to irrelevant input perturbations

When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer.

DEEPER ISSUES WITH THE USAGE OF LLMS AND OTHER ADVANCED AI TOOLS

Advanced AIs do not (usually) respect the privacy of their users

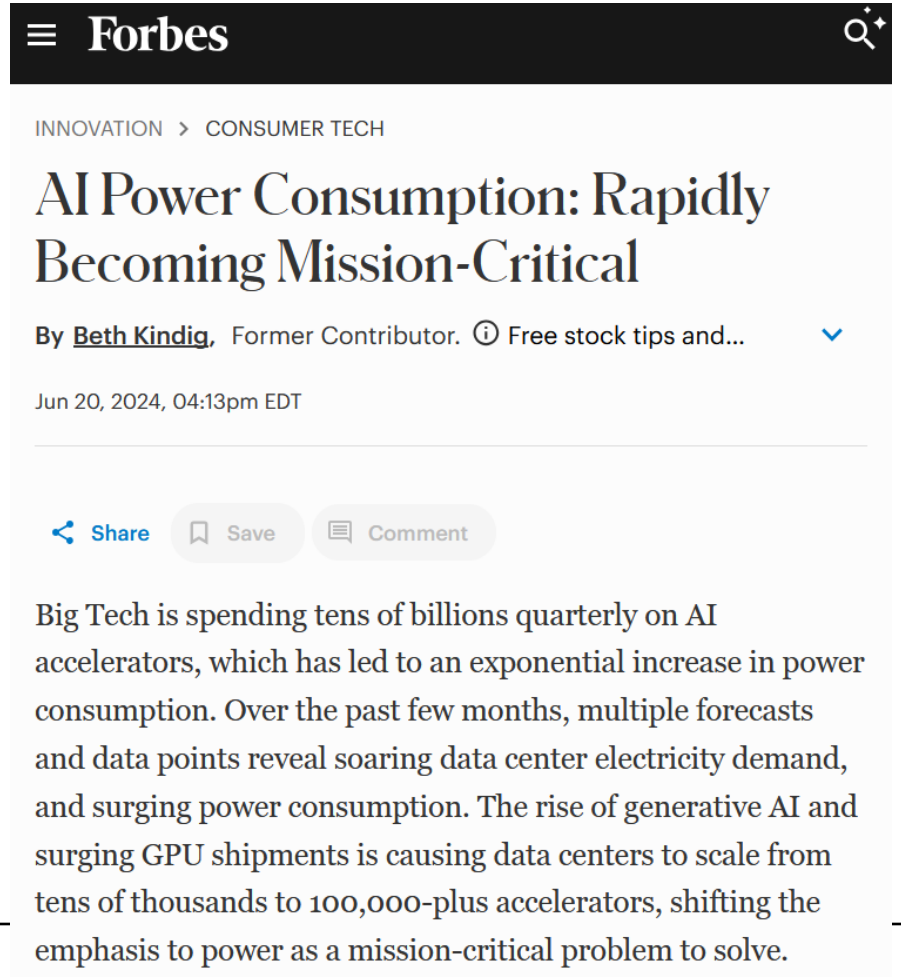
Most of the LLMs are accessible only on web services offered by the developers, they track all Q&A even if including sensitive or copyrighted data



Datasets are mostly secret; AI companies are already undergoing court cases for copyright infringement (e.g., Andersen vs Stability AI, NY Times vs OpenAI)

DEEPER ISSUES WITH THE USAGE OF LLMS AND OTHER ADVANCED AI TOOLS

Advanced AIs are incredibly energy-consuming



Forbes

INNOVATION > CONSUMER TECH

AI Power Consumption: Rapidly Becoming Mission-Critical

By Beth Kindig, Former Contributor. Free stock tips and...

Jun 20, 2024, 04:13pm EDT

Share Save Comment

Big Tech is spending tens of billions quarterly on AI accelerators, which has led to an exponential increase in power consumption. Over the past few months, multiple forecasts and data points reveal soaring data center electricity demand, and surging power consumption. The rise of generative AI and surging GPU shipments is causing data centers to scale from tens of thousands to 100,000-plus accelerators, shifting the emphasis to power as a mission-critical problem to solve.

Source: https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

DEEPER ISSUES WITH THE USAGE OF LLMS AND OTHER ADVANCED AI TOOLS

Advanced AIs are *deskilling* professional users

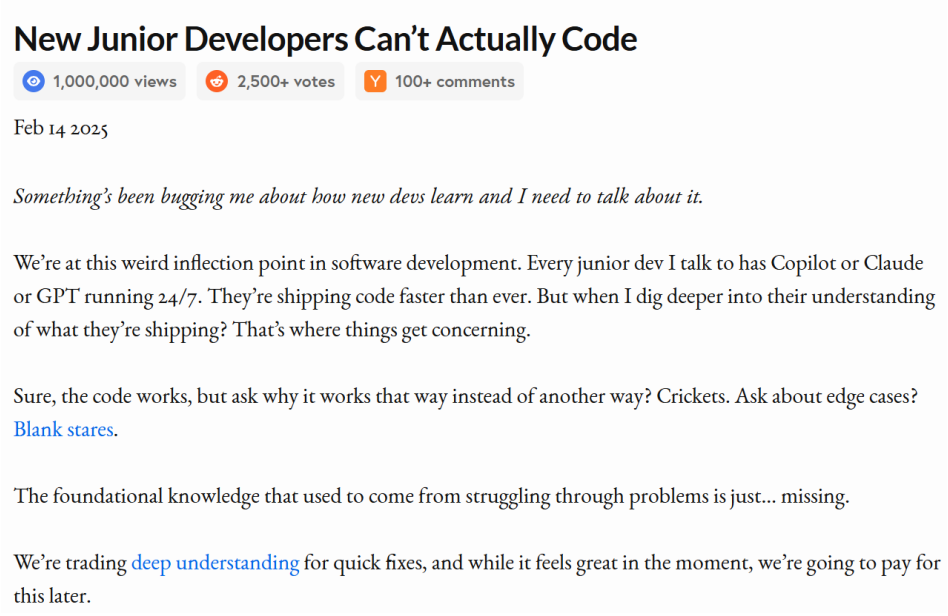
Borg et al. (2024)

Overreliance of aircraft pilots on AI-based safety systems

Overreliance of doctors on AI-based diagnoses systems

DEEPER ISSUES WITH THE USAGE OF LLMS AND OTHER ADVANCED AI TOOLS

Advanced AIs are decreasing the formation of advanced skills and critical thinking in students and junior employees



New Junior Developers Can't Actually Code

1,000,000 views · 2,500+ votes · 100+ comments

Feb 14 2025

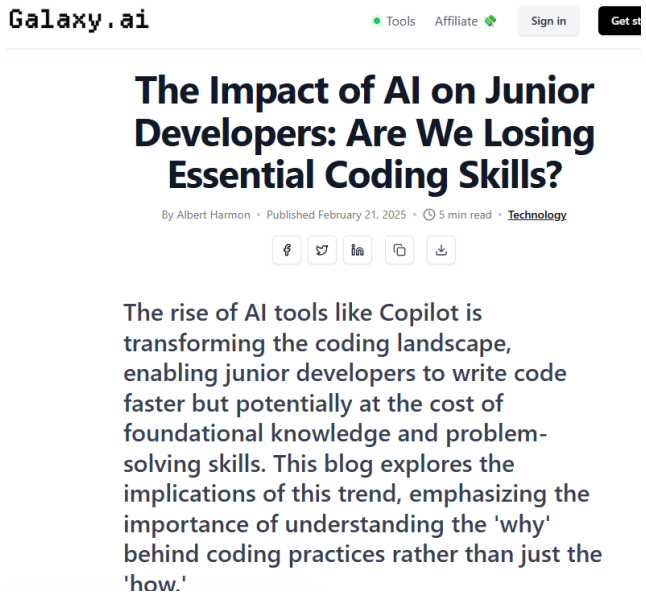
Something's been bugging me about how new devs learn and I need to talk about it.

We're at this weird inflection point in software development. Every junior dev I talk to has Copilot or Claude or GPT running 24/7. They're shipping code faster than ever. But when I dig deeper into their understanding of what they're shipping? That's where things get concerning.

Sure, the code works, but ask why it works that way instead of another way? Crickets. Ask about edge cases? Blank stares.

The foundational knowledge that used to come from struggling through problems is just... missing.

We're trading deep understanding for quick fixes, and while it feels great in the moment, we're going to pay for this later.



Galaxy.ai · Tools · Affiliate · Sign in · Get st

The Impact of AI on Junior Developers: Are We Losing Essential Coding Skills?

By Albert Harmon · Published February 21, 2025 · 5 min read · Technology

The rise of AI tools like Copilot is transforming the coding landscape, enabling junior developers to write code faster but potentially at the cost of foundational knowledge and problem-solving skills. This blog explores the implications of this trend, emphasizing the importance of understanding the 'why' behind coding practices rather than just the 'how.'

(Gerlich, 2025) Cognitive load associated with higher degree of critical thinking. Cognitive offload with AI leads to smaller critical thinking.
(Ramenskii, 2024) LLMs usage does not significantly impact students' performance, but reduces self-reliance and confidence.

Sources: New Junior Developers Can't Actually Code | N's Blog; The Impact of AI on Junior Developers: Are We Losing Essential Coding Skills? | Galaxy.ai

MY QUESTIONS FOR YOU

What do you think is the most important issue concerning the usage of LLMs from students?

What do you think the university is for?

MY SUGGESTIONS FOR YOU

Limit usage of LLMs in academia (it's for your good!)

Don't underestimate the negative effects of *cognitive offloading*

Use LLMs to

- get different views on a topic
- confirm correctness of reasoning
- offload things you don't really need to learn
- do "repetitive" labor that doesn't teach you anything
- personalized tutoring

May 29th - Round Table

16:30 - Sala dei Notari

Piazza IV Novembre – 06123, Perugia (PG)

«Biotecologie e Società: la sfida delle Biotecnologie e della IA»

BIOMATERIALI

Intervengono:

Dott. Alberto Franzin – Technology Specialist, EU AI Office
Prof. Andrea Capaccioni – Università di Perugia
Dott. Filippo Molinari – Responsabile Istituto di Scienze Biomediche della Difesa
Prof. Francesco Ciampi – Università di Firenze
Dott.ssa Valentina Poggioni – Università di Perugia

INTELLIGENZA ARTIFICIALE

unipg
UNIVERSITÀ DEGLI STUDI DI PERUGIA

PNRR per la M4C2 – Linea Intervento 1.5 :
ECS_00000041 - CUP: J97G22000170005

THANKS FOR THE ATTENTION!

Questions?

My contacts

marco.zullich@gmail.com

https://www.linkedin.com/in/marco-zullich-00559660/

SOURCES

- (Borg et al., 2024) Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And how we get there*. Random House.
- (EC, 2019) European Commission. "Ethics guidelines for trustworthy artificial intelligence." *High-Level Expert Group on Artificial Intelligence* 8 (2019).
- (Gebru & Buolamwini, 2018) Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.
- (Gerlich, 2025) Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies, 15(1), 6.
- (Mirzadeh et al., 2024) Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.
- (Ramenskii, 2024) Ramenskii, I. (2024). Study on the Impact of Using Large Language Models (LLMs) like ChatGPT on Mental Effort and Learning Abilities Among Polish Students. *Preprint*.
- (Zech et al., 2018) Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*.