Engineering with Generative AI

01 – Introduction to GenAI Engineering

Sebastian Vollmer

28.10.2025

Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Hello from our side..

Nice To Meet You!

- We are the Data Science and its Applications (DSA) Research Group. We are a constellation of researchers who are united under the supervision of Professor Sebastian Vollmer, with a shared goal of advancing data science methods and tools and using them across industrial and socially-important applications.

- Our main homes are the German Research Center for Artificial Intelligence (DFKI), where we form the Data Science and its Applications research department, and the University of Kaiserslautern, where we form the Applied Machine Learning group.

Sebastian Vollmer

Islam Mesabah

Mohaddeseh Tabrizian

Pre-Requisites

- **Proficiency in Python 3:** The engineering components of the course will primarily be conducted using the Python 3 programming language. While this is not a programming-focused course, having a comfortable level of proficiency in Python 3 is important.

- **Fundamental Machine Learning Skills:** This course assumes that you possess a basic understanding of machine learning concepts. You should be able to write, comprehend code, and work with classification and regression models, and handle basic data science pipelines in Python.

- **Experience with NLP (Optional):** While not mandatory, prior experience with Natural Language Processing (NLP) concepts will give you a head start. If you're already familiar with topics like tokenization, embeddings, attention, and transformers, you'll find it easier to grasp the advanced concepts presented in the course.

Lecture logistics

- **Attendance is required**
  - **Sessions are recorded, but only some of the recordings will be uploaded.**
- Lecture schedule will be on the OLAT course
- Lecture on **Tuesday, 12:15 - 13:45 in 42-115**
- Exercise sessions **Thursday, 12:15 - 13:45 Online**
- Passing two of three assignments is mandatory for the written exam admission
- Questions via OLAT forms will be answered, others will be hit and miss
- OLAT course link:
  https://olat.vcrp.de/auth/RepositoryEntry/4988502657

Course Updates

- This course is 5 Credits (2V+2Ü)
- More topics included:
  - Context Engineering
  - Reasoning & Inference Optimization
  - LLM Security
  - GenAI 4 Science
- More focus on MCP and context engineering
- A wide range of guest speakers from England, Mainz, and Egypt will be joining the course.
- A new drag-and-drop teaching tool designed to simplify the starting point with the LLM [N8N (https://n8n.io)].

n8n

Teaching Approach

- The course is inclusive: it will signpost aspects that some of you might know in depth but for others are new and help them on their learning journey
- We will emphasize practical aspects of engineering with large language models
  - **This is not an NLP course**
- We will learn how to use these powerful models to build applications
- Many people claim to be experts – I am not an expert on LLMs – I don't have all the answers
- We bring many different lectures that have different expertise and style of lecturing

Assistant tools for AI prototyping & testing

Helping developers make their AI systems
more trustworthy, robust, secure,

MISSION KI

Federal Ministry
for Digital
and Transport

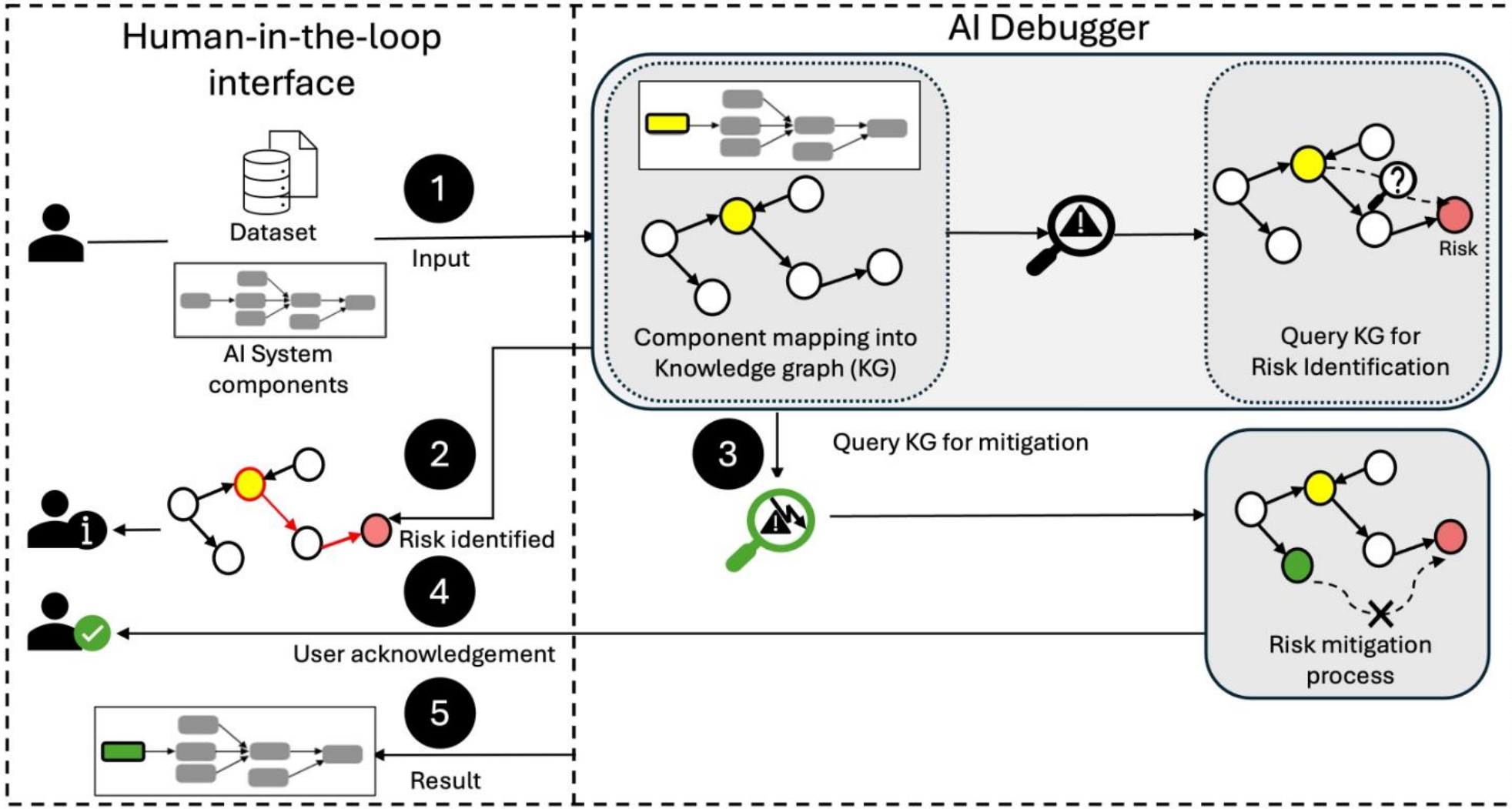


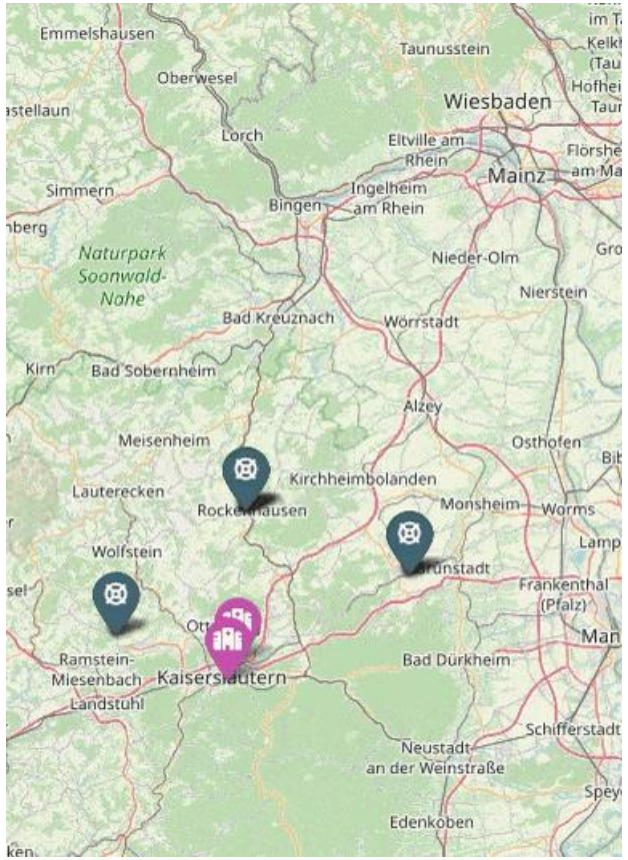
Figure 2: A flowchart depicting the **AI debugger** workflow. This process maps user inputs to the KG and query for potential risk identification and mitigation under HITL supervision.

(Spatio-)temporal modelling and forecasting

BOSTON COLLEGE

UNIVERSITY OF OXFORD

Ambulance allocation



github.com/gerritgr/AmbulanceGame

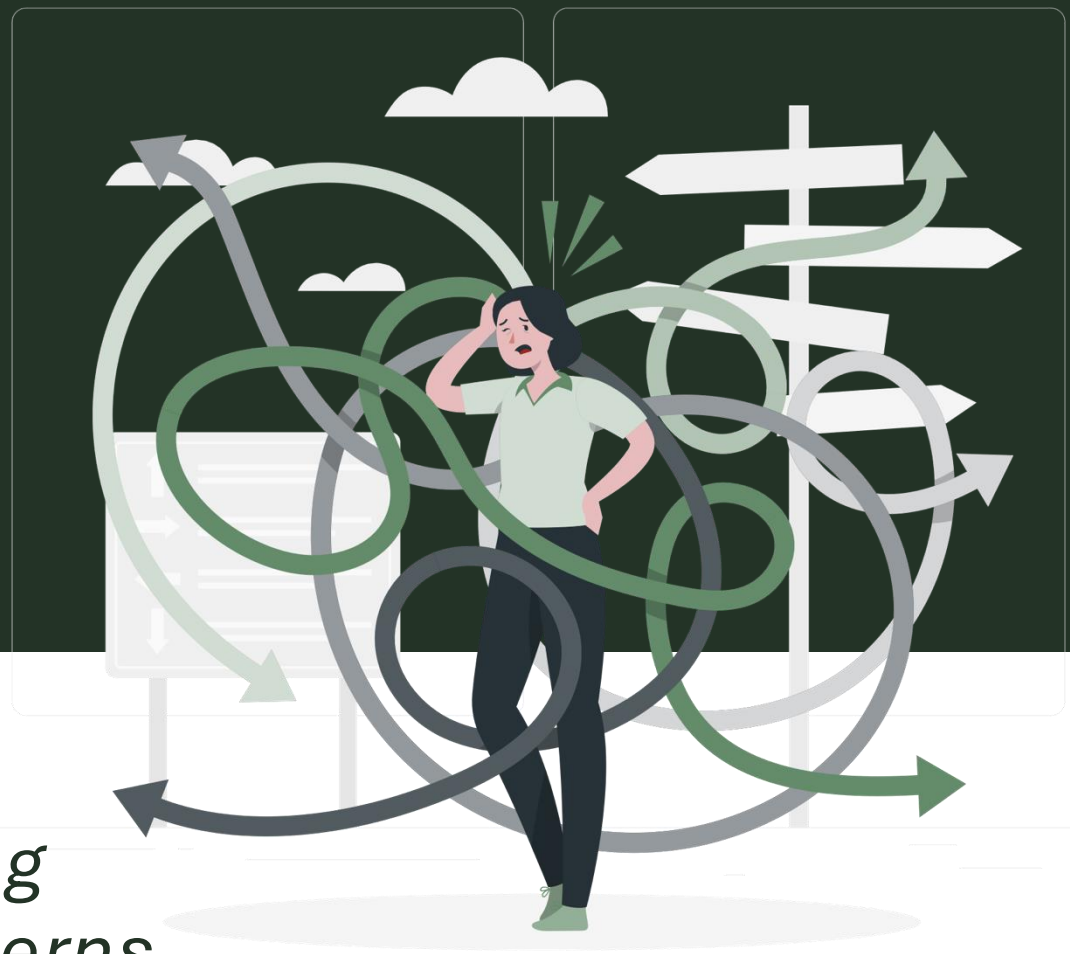
Goal: Better decision-making in ambulance dispatching for faster response times and stronger coverage.

- Dispatch an ambulance?
- Dispatch an emergency doctor?
- Which ambulance to send?
- Which hospital to choose?
- Use lights and sirens?
- Which route?
- Relocate other ambulances?

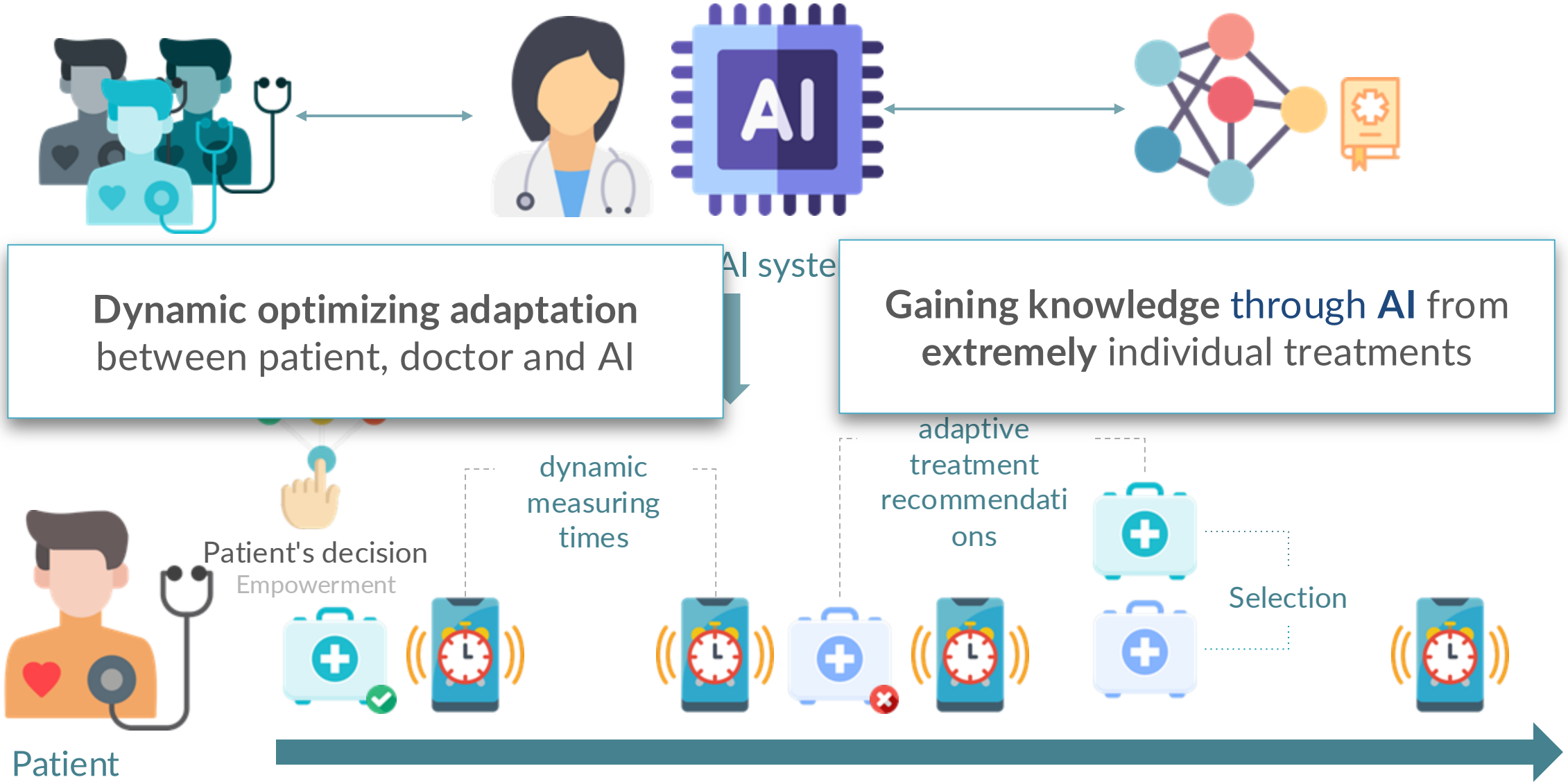
Utilising unstructured data and knowledge

Structuring unstructured data to give actionable insights

Causal, probabilistic, reinforcement learning that scales



Explainable, actionable machine learning that captures relationships, not just patterns

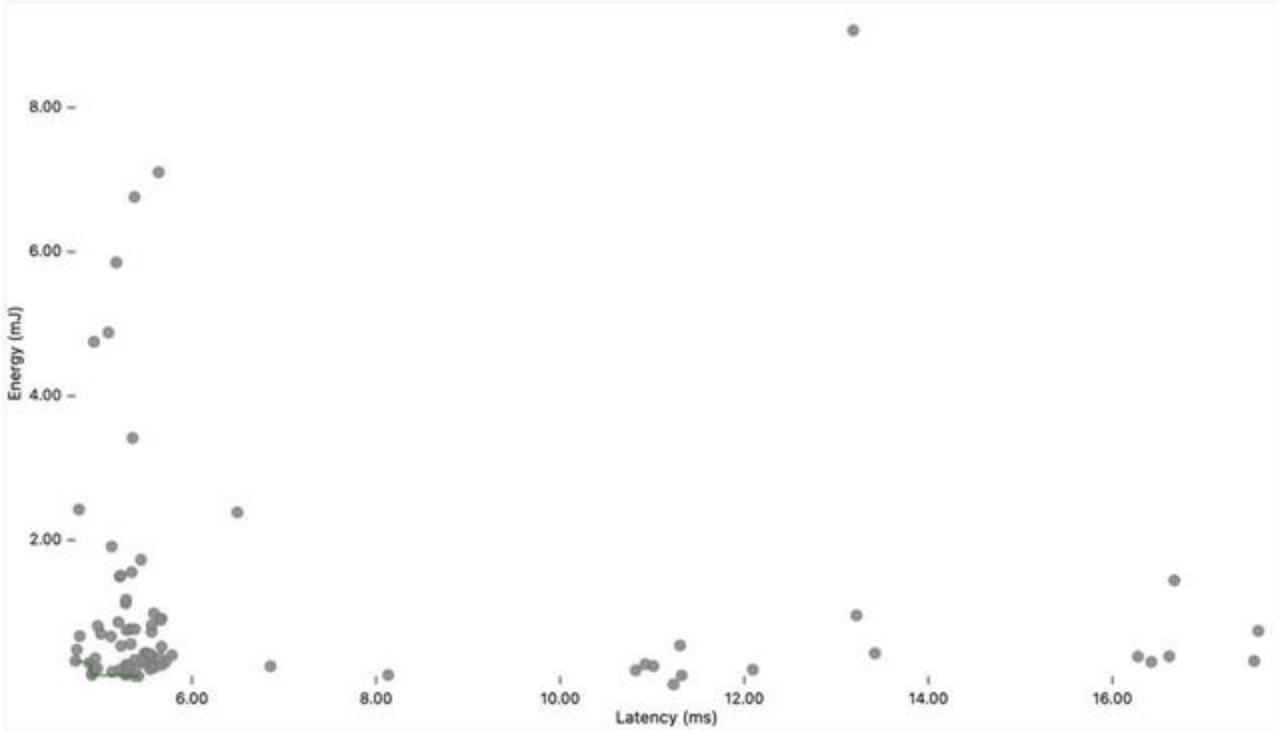

AI
AI syste
Dynamic optimizing adaptation
between patient, doctor and AI
Gaining knowledge through AI from
extremely individual treatments
Patient's decision
Empowerment
dynamic
measuring
times
adaptive
treatment
recommendati
ons
Selection
Patient

Leaner, greener, cheaper AI

Examples

SDPA Pareto Visualisation

Total candidates: **82** | With energy: **81** | Pareto count: 5



Axes are linear. Units: latency (ms), energy (mJ). If energy is unavailable, Y shows index.

Details

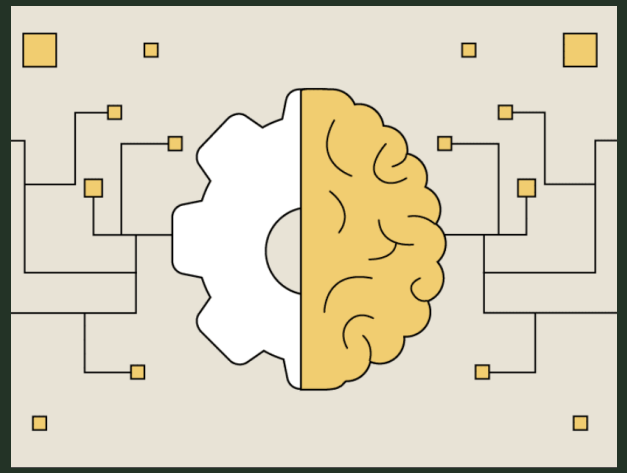
variant: builtin | **dtype:** bfloat16 | **matmul:** matmul | **chunk_q:** 128 | **chunk_kv:** 256 | **causal:** true | **dropout_p:** 0

latency_ms	6.853083963505924
energy_mJ	0.2509485448598862
correct	true
note	LLM-provided code
code_hash	e9ceb48b29cd
code_len	826

Underlying code

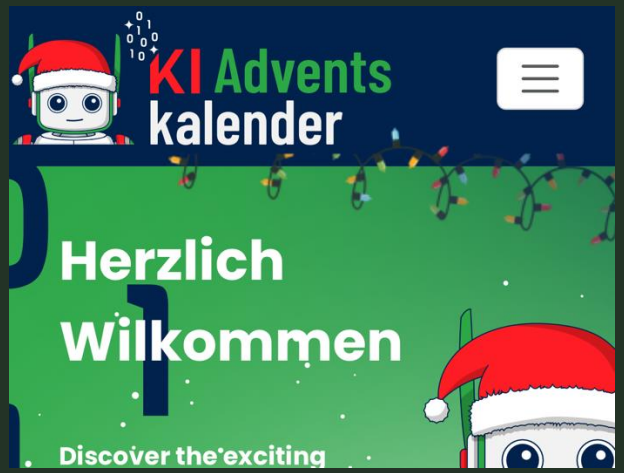
```
import torch
import math

def sdpa_forward(q, k, v, causal=False, dropout_p=0.0):
    B, H, T, D = q.shape

    # Scaling factor
    scale = 1.0 / math.sqrt(D)

    # Calculate attention scores
    attn_scores = torch.matmul(q, k.transpose(-2, -1)) *

    # Apply causal mask if needed
    if causal:
        mask = torch.tril(torch.ones(T, T, device=q.devi
        attn_scores = attn_scores.masked_fill(~mask, flo

    # Softmax to get attention weights
    attn_weights = torch.softmax(attn_scores, dim=-1)
```

Engineering with Generative AI



Machine Learning 2

Machine Learning & NLP with AWS



AI advent calendar

Welcome to “Engineering with Generative AI”

Today

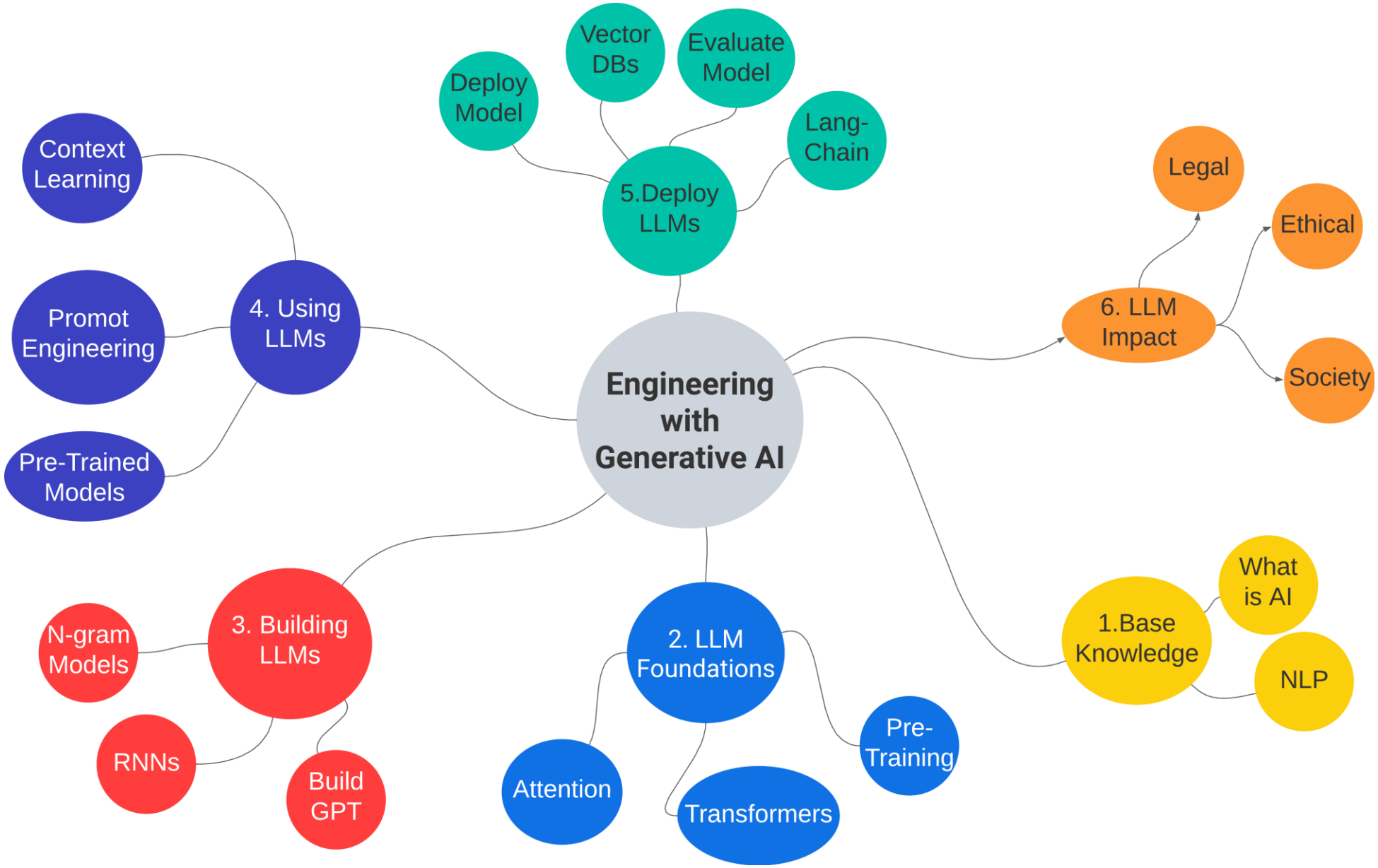
- Overview of the Landscape and the course's contents and logistics
- You should get a sense of what you know and don’t know
- Discriminative AI vs Generative AI
- Focus on Large Language Models as the main type of generative AI in this course

- We are in this together! It is our **THIRD** version! Your feedback, comments and additions are highly appreciated!

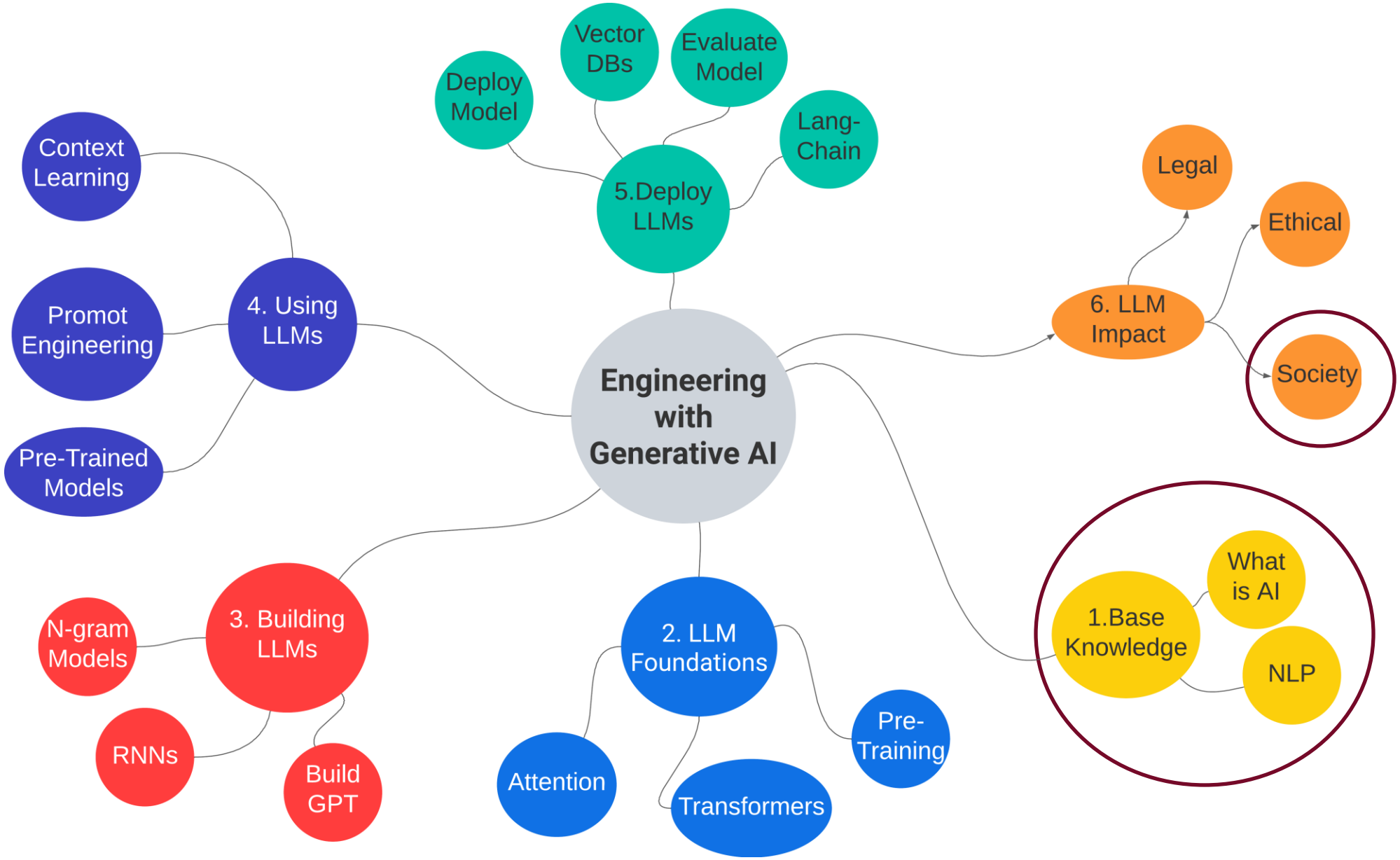
Landscape of Large Language Models

Follow @theailifestyle for more content

Our Course: Engineering with Generative AI



1. AI & Society


Engineering with Generative AI
1. Base Knowledge
What is AI
NLP
2. LLM Foundations
Pre-Training
Transformers
Attention
3. Building LLMs
N-gram Models
RNNs
Build GPT
4. Using LLMs
Context Learning
Promot Engineering
Pre-Trained Models
5.Deploy LLMs
Deploy Model
Vector DBs
Evaluate Model
Lang-Chain
6. LLM Impact
Legal
Ethical
Society

What is AI?

From the EU AI ACT

AI is defined as: "Software developed using specific techniques and approaches (detailed in an appendix) that can produce results such as **content, predictions, recommendations or decisions** for a given set of human-defined goals that influence the environment with which it interacts.

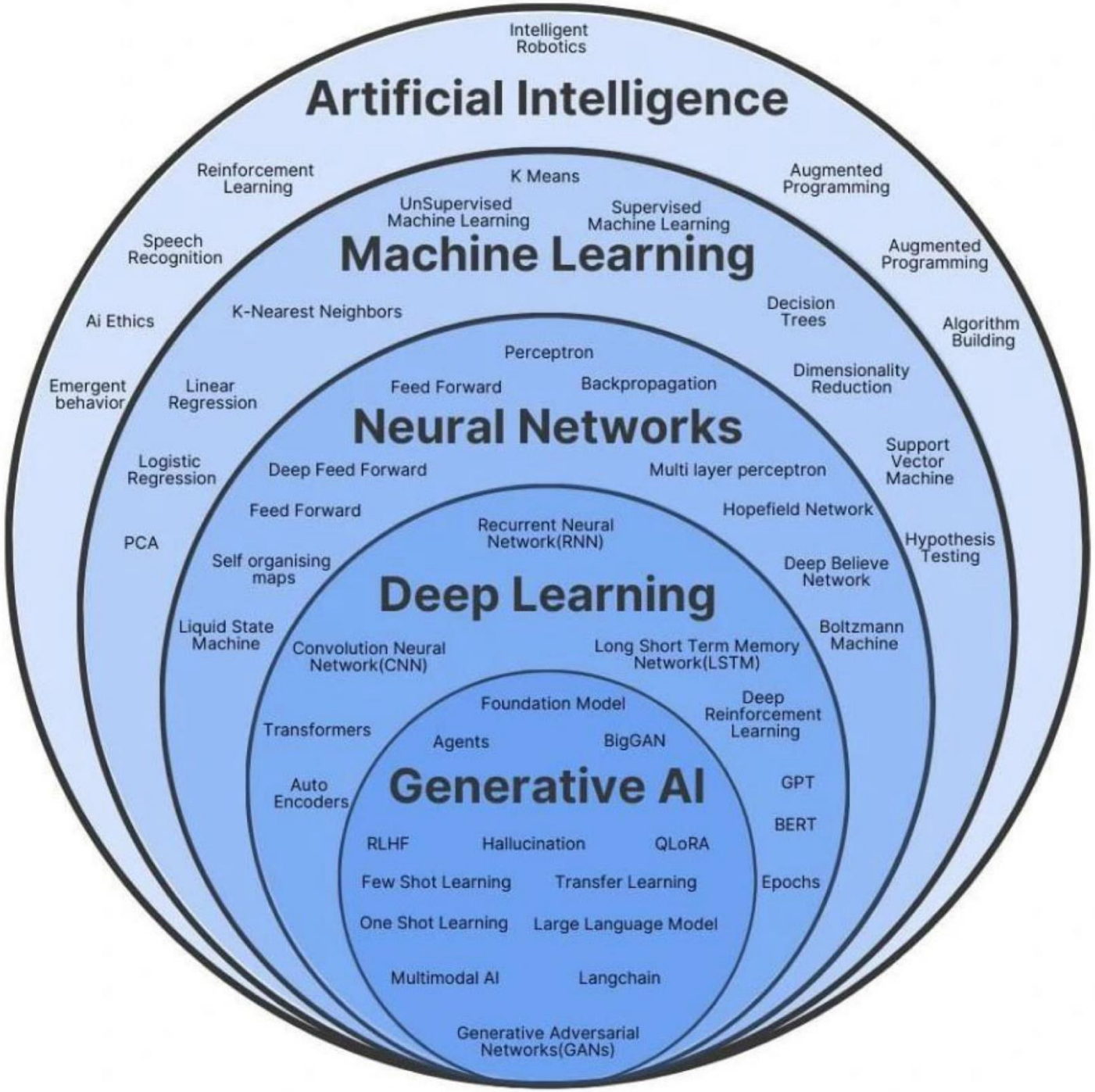
Generative AI

ANNEX I
ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES
referred to in Article 3, point 1

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

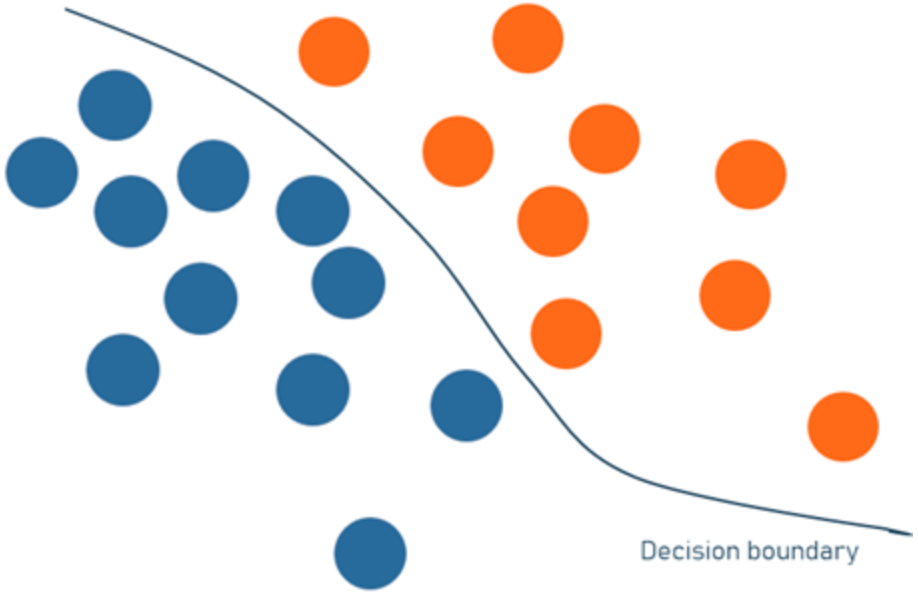

Intelligent Robotics
Artificial Intelligence
Reinforcement Learning
Speech Recognition
Ai Ethics
Emergent behavior
Augmented Programming
Augmented Programming
Algorithm Building
K Means
UnSupervised Machine Learning
Supervised Machine Learning
Machine Learning
K-Nearest Neighbors
Decision Trees
Linear Regression
Logistic Regression
PCA
Dimensionality Reduction
Support Vector Machine
Hypothesis Testing
Perceptron
Feed Forward
Backpropagation
Neural Networks
Deep Feed Forward
Multi layer perceptron
Feed Forward
Hopefield Network
Self organising maps
Deep Believe Network
Liquid State Machine
Boltzmann Machine
Recurrent Neural Network(RNN)
Deep Learning
Convolution Neural Network(CNN)
Long Short Term Memory Network(LSTM)
Deep Reinforcement Learning
Transformers
Auto Encoders
Foundation Model
Agents
BigGAN
Generative AI
GPT
BERT
Epochs
RLHF
Hallucination
QLoRA
Few Shot Learning
Transfer Learning
One Shot Learning
Large Language Model
Multimodal AI
Langchain
Generative Adversarial Networks(GANs)

Discriminative AI

- Traditional AI with machine learning and deep learning models were primarily designed for categorical/numerical decision-making,

- Most discriminative AI applications were task-centric, focusing on specific objectives
  - like image classification, sentiment analysis, or spam detection, often lacking broader context or creativity.

- Discriminative models typically utilized only labeled data for classification tasks,
  - missing out on the wealth of unlabeled or unstructured data that could be harnessed for broader AI capabilities.



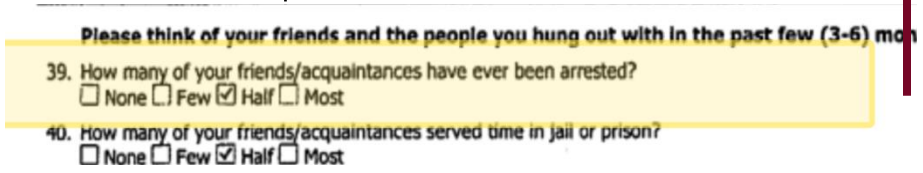
Applications of Supervised Learning

Question: Does the client repay the mortgage?

Age	Salary	Account balance	Credit	Paid back?
35	2000	100000	500000	Yes
45	3000	15000	300000	No
...	...	...	...	...

Input	Output	Application area
Email	Spam (Yes/No/True)	Spam Filter
Advertising, user profile	Click (Yes/No/True)	Online Advertising
Medical record	Develops cancer within 3 years	Health system
Flight route	Paraffin demand (number)	Transport optimisation
Shoe print 1 & 2	Concordance	
Offender profile	Repeat offender (Yes/No)	Justice

Crimes in the family (Keyword: COMPAS)

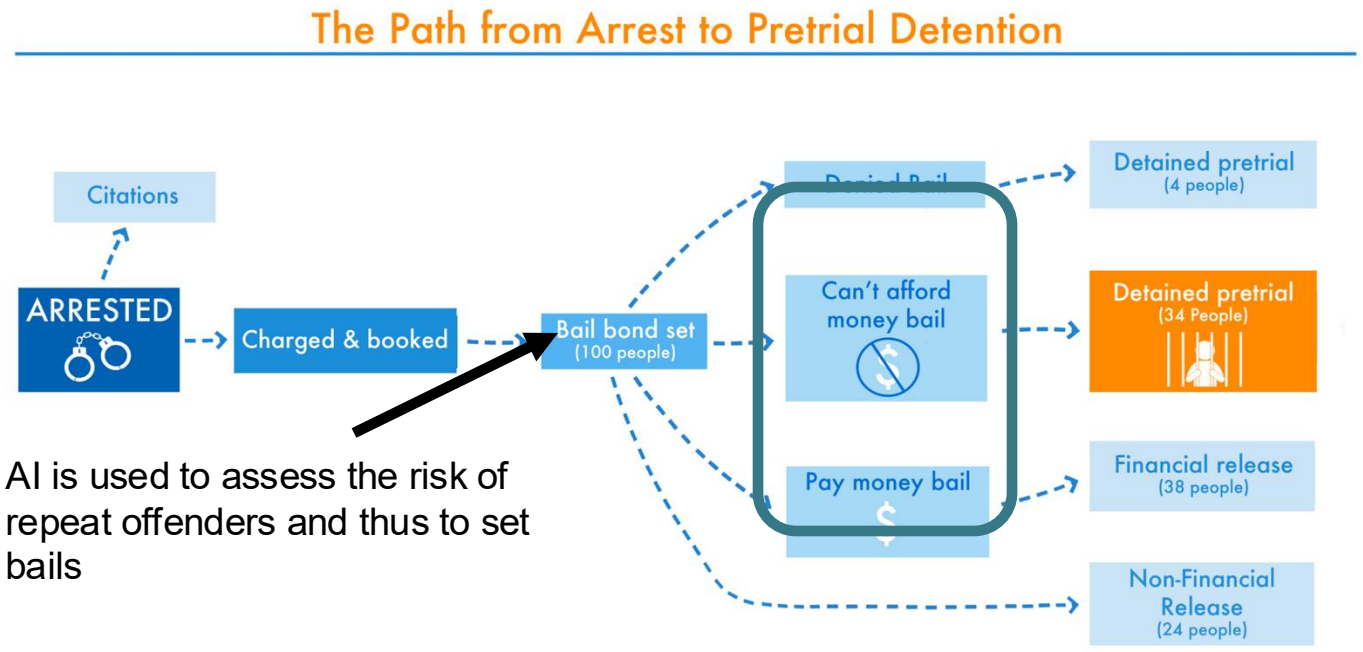

Please think of your friends and the people you hung out with in the past few (3-6) mon
39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most
40. How many of your friends/acquaintances served time in jail or prison?
☐ None ☐ Few ☑ Half ☐ Most

Evaluation & Fairness

GDPR

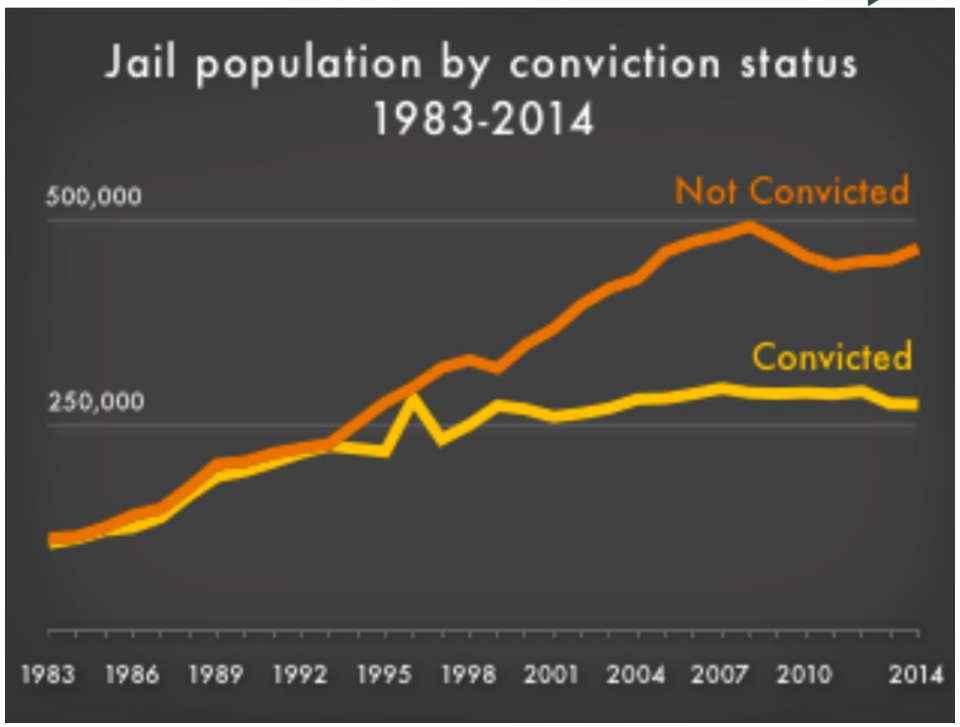
THE AI ACT

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)



AI is used to assess the risk of repeat offenders and thus to set bails

https://www.prisonpolicy.org/



https://www.prisonpolicy.org/

Probublica

COMPAS

- Prediction of a risk score from 0 to 10, which represents the likelihood of a person becoming a repeat offender or absconding.
- The judge makes his decisions on the basis of this value, **among other things**

Question:
Is this risk value fairly distributed or does skin colour play a role?

https://www.prisonpolicy.org/reports/incomejails.html

Many Measures: How do we select what we care about?

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

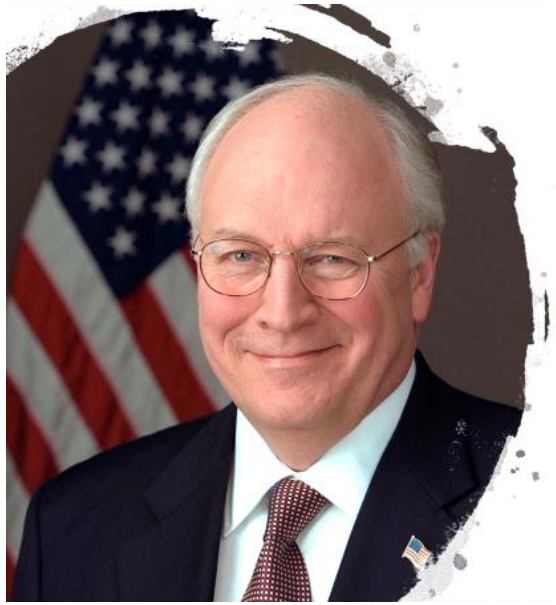
$\frac{1}{\tau} \leq \qquad \leq \tau$

Verma, S. & Rubin, J. Fairness Definitions Explained. in 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) 1–7 (2018).

What is the right error to prioritize?

„It is better that ten guilty persons escape than that **one** innocent suffer.“

William Blackstone, Rechtsphilosoph, 1760



"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

Dick Cheney, ehemaliger Vizepräsident der USA,

False Negative Rate =

Proportion of offenders who could be released

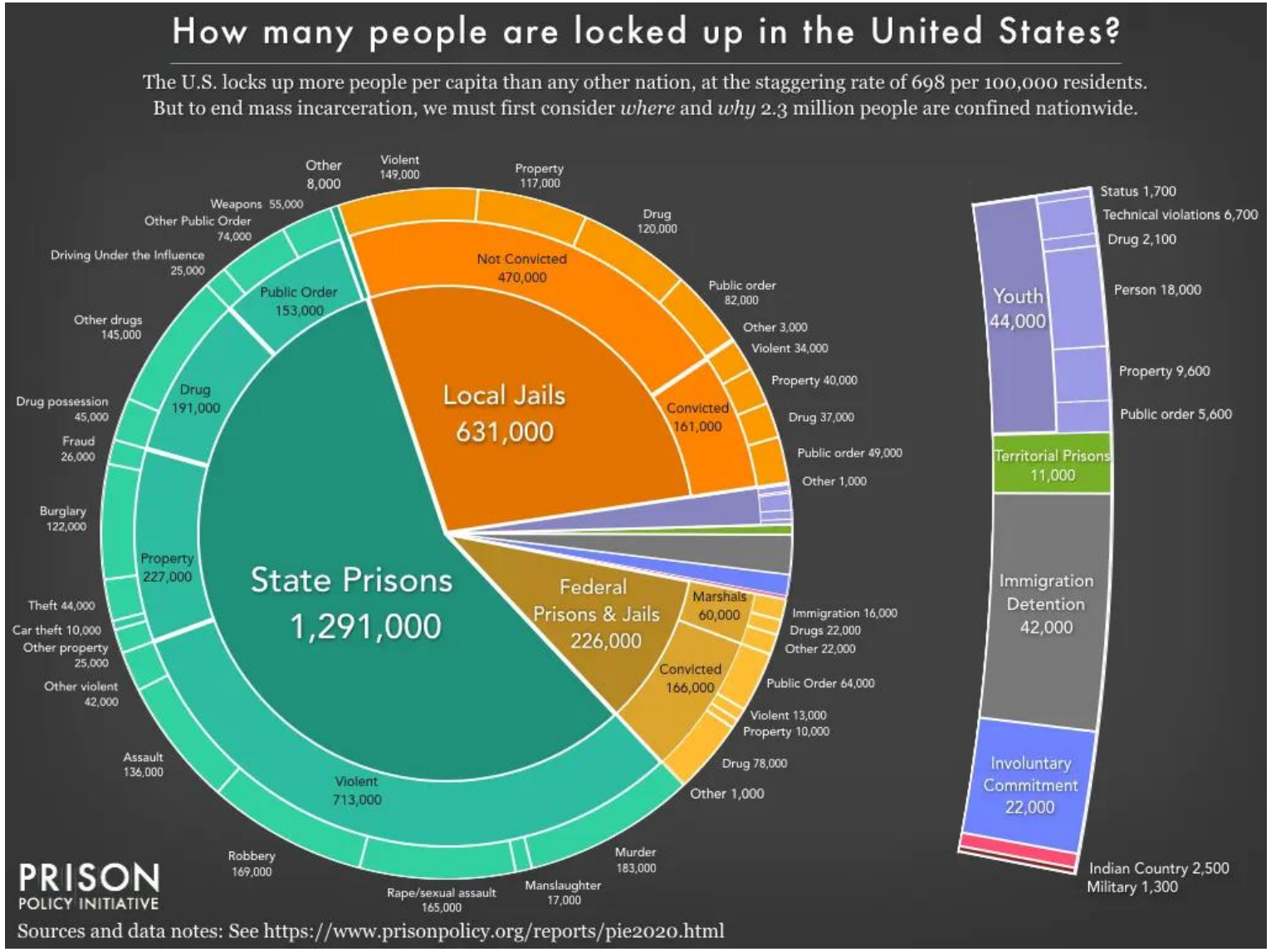
False Discovery Rate =

Proportion of released offenders who reoffend

Metric	Caucasian	African American
False Negative Rate (FNR)	48%	28%
False Discovery Rate (FDR)	41%	37%

Credit: K. Zweig

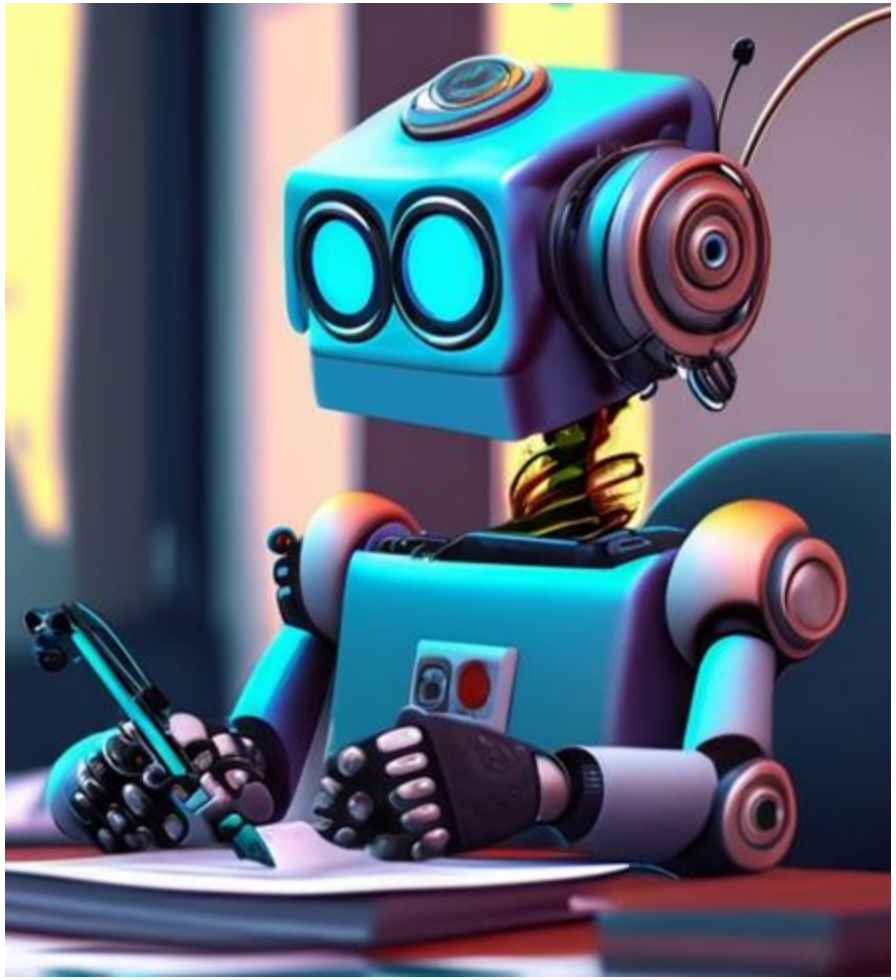
Extent of the Problem



So, What Exactly is Generative AI?

Generative AI

- Generative AI (GenAI) systems are designed to produce new data or content that is not explicitly programmed or predefined by humans.
- This can include generating text, images, music, or other types of media.
- Generative AI learn patterns from existing data and generate novel content based on those patterns.
- Generative AI can exhibit a degree of creativity and imagination by generating content that is not copied but rather synthesized from its training data.



Generative AI vs Discriminative AI

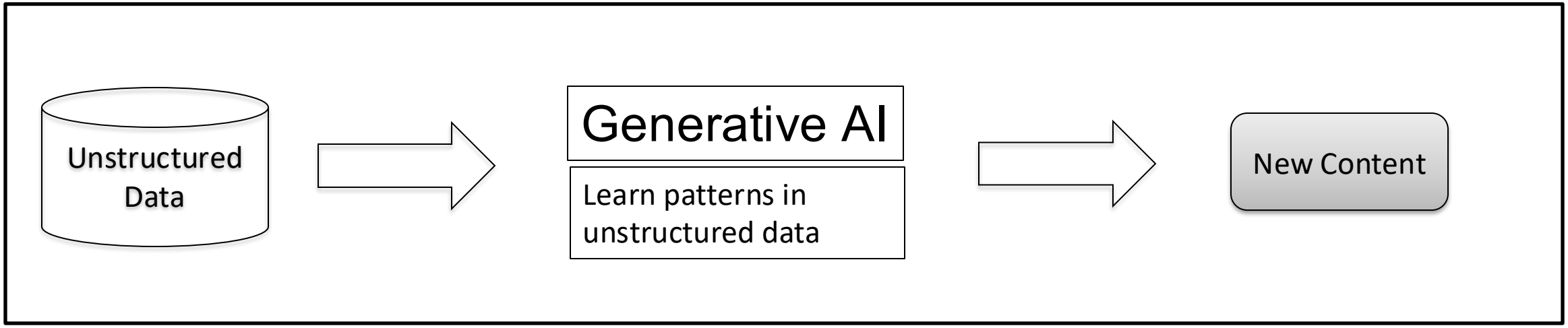
Disriminative AI

Classify

Generative AI

Generate

Generative AI vs Discriminative AI



Language vs Image GenAI

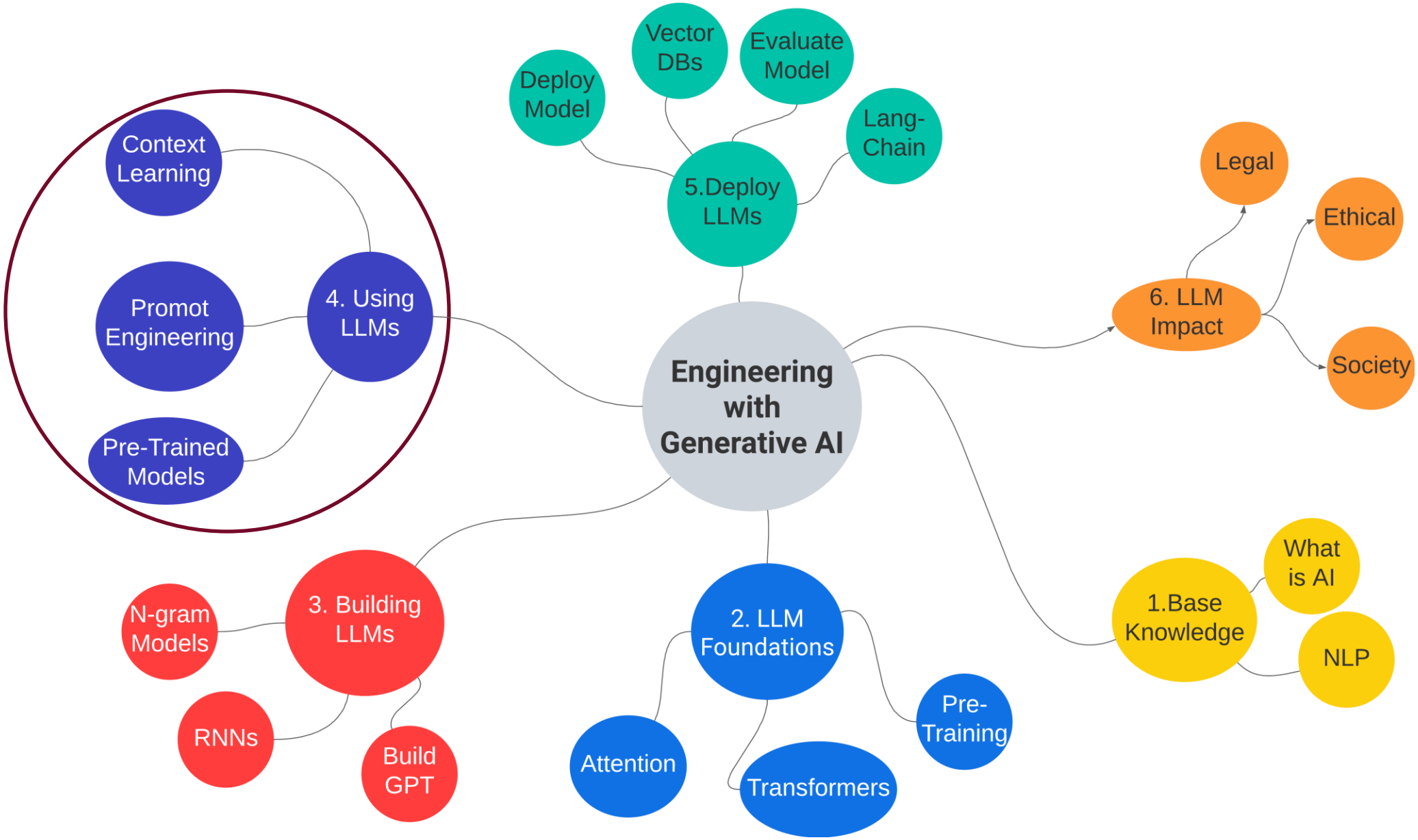
Generative Language Models

- Learn about patterns in languages through training
- They typically use sequences of words or tokens as their input and output
- Given some text, predict what comes next

Generative Image Models

- Learn about images and understand the underlying statistical patterns.
- Capable of generating highly realistic images that closely resemble photographs of real objects.
- Given some text or related imagery, generate new images

2. Using LLMs


Engineering with Generative AI
4. Using LLMs
Context Learning
Promot Engineering
Pre-Trained Models
5.Deploy LLMs
Deploy Model
Vector DBs
Evaluate Model
Lang-Chain
6. LLM Impact
Legal
Ethical
Society
1.Base Knowledge
What is AI
NLP
2. LLM Foundations
Pre-Training
Attention
Transformers
3. Building LLMs
N-gram Models
RNNs
Build GPT

Simple prompts – time saved

Please plan a halloween party hour by hour for my 8 year old daughter and 5 year old son inviting their friends. The party should runf form 5pm to 9 pm.

That sounds like a fun idea! I can help you plan a spooktacular Halloween party for your kids and their friends. Here are some suggestions for the party theme, food, decorations, and activities, based on the web search results I found.

Party Theme: One possible theme for your Halloween party is a **Haunted Mansion**. You can transform your home into a creepy and mysterious place, with spooky music, dim lighting, and eerie props. You can also invite the kids to dress up as ghosts, vampires, zombies, or other scary characters.

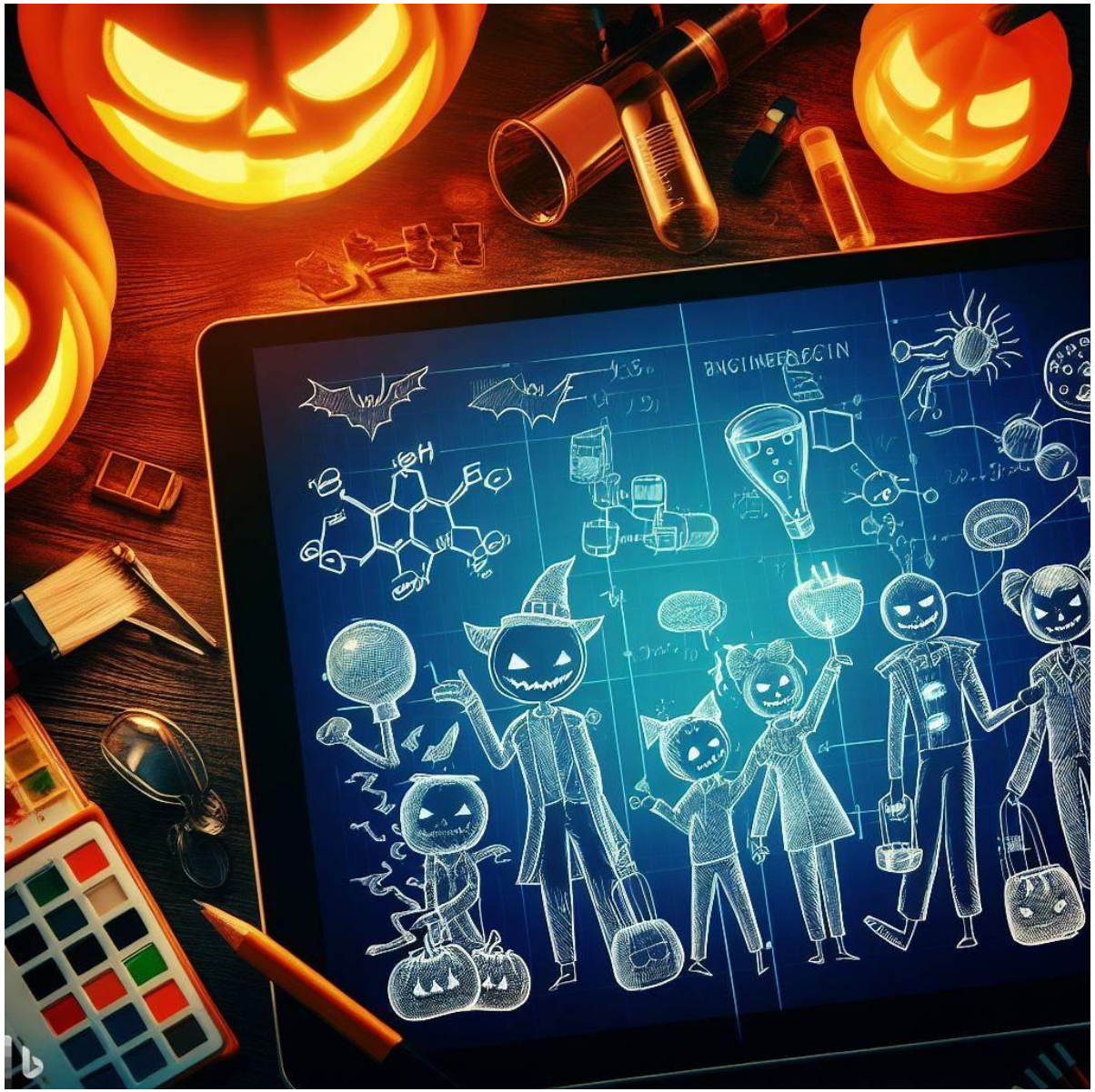
Party Food: Some of the food ideas for your Haunted Mansion theme are:

•**Mummy Dogs**: Wrap hot dogs in crescent roll dough and bake them until golden. Add mustard or ketchup eyes to make them look like mummies[1].

•**Worms in Dirt**: Crush chocolate sandwich cookies and layer them with chocolate pudding and gummy worms in clear cups[2].

Simple prompts – Images as well!

Please generate a nice image about halloween party. Some engineering scheme. No horror or blood or that kind of things. Friendly one!:)



Simple prompts – time saved

Please plan a halloween party hour by hour for my 8 year old daughter and 5 year old son inviting their friends. The party should run from 5pm to 9 pm.

5:00 - 5:15 pm	Welcome the guests and show them where to put their coats and bags.
5:15 - 5:45 pm	Serve the party food and drinks in the dining room.
5:45 - 6:15 pm	Play Pin the Tail on the Black Cat in the living room.
6:15 - 7:00 pm	Take the kids on the Haunted House Tour around your house.
7:00 - 7:30 pm	Have a break and let the kids enjoy some candy or snacks.
7:30 - 8:30 pm	Have the Pumpkin Carving Contest in the backyard or garage.
8:30 - 9:00 pm	Give out the prizes for the best pumpkins

Simple prompts – time saved

Please add time for trick or treating to the schedule.

the best time for trick or treating is usually between **6 pm and 8 pm**, when it is dark enough but not too late. However, you may want to check your local regulations and safety guidelines before you go out.
Here is an updated schedule for your party, with some time for trick or treating:
I hope this helps you plan a fun and safe Halloween party for your kids and their friends. Have a wonderful time! 🎃

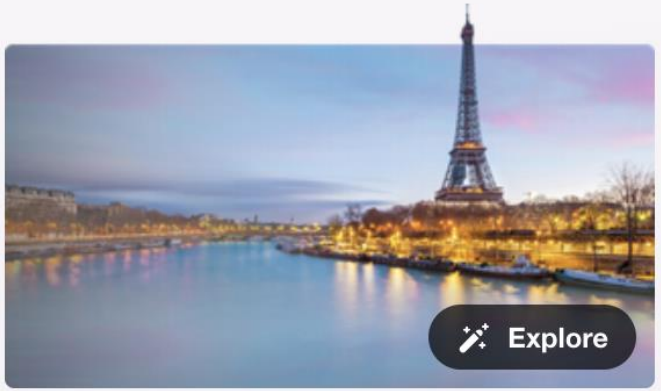
Time	Activity
5:45 - 6:15 pm	Play Pin the Tail on the Black Cat in the living room.
6:15 - 7:15 pm	Take the kids on the trick or treating around your neighborhood. Make sure they have bags or buckets to collect their candy and flashlights or glow sticks to be visible.
7:15 - 7:45	Have a break and let the kids enjoy some of their candy or snacks

Simple prompts – time saved

Plan a day in paris hour by hour and I want to enjoy the nightlife too.

Bonjour, this is Bing. I'm happy to help you plan a day in Paris, the City of Light. 😊

Here is a possible itinerary for you, based on some of the best things to do in Paris during the day and night. Of course, you can adjust it according to your preferences and budget.



9:00 AM - Start your day with a delicious breakfast at one of the many cafes or bakeries in Paris. You can try some of the typical French pastries, such as croissants, pain au chocolat, or baguettes with butter and jam.

10:00 AM - Visit the iconic Eiffel Tower, the symbol of Paris and one of the most visited monuments in the world. You can take the elevator or the stairs to the top and enjoy the panoramic views of the city. You can also explore the gardens and the Champ de Mars around the tower.

12:00 PM - Head to the Musée d'Orsay, one of the most famous museums in Paris, located in a former railway station. It houses a stunning collection of impressionist and post-impressionist paintings, sculptures, and other artworks by artists such as Monet, Renoir, Van Gogh, Degas, and Rodin [1].

Simple prompts – high consequences

Please draft an email to HR asking for ending the emolyment of Martin Schmidt - he was supposed to fill a proactive role between sales and R&D.

I can help you with drafting an email to HR. Here is a possible template that you can use or modify as you wish. Subject: Termination of Martin Schmidt

Dear HR Manager,

I regret to inform you that I have decided to terminate the employment of Martin Schmidt, effective from November 30, 2023. Martin was hired as a sales and R&D liaison, but he has failed to perform his duties satisfactorily. As you may recall, Martin was supposed to fill a proactive role between sales and R&D, facilitating communication and collaboration between the two departments.

He has missed several deadlines, ignored feedback, and caused conflicts with his colleagues. He has also violated the company's policies on confidentiality and ethics on multiple occasions.

I have tried to address these issues with Martin through regular meetings, coaching sessions, and performance reviews. I have also given him clear expectations, goals, and improvement plans. Unfortunately, he has not shown any signs of improvement or willingness to change. His poor

Hallucinations

SM

What is the most cited economics paper of all time

The most cited economics paper of all time is "A Theory of Economic History" by Douglass North and Robert Thomas, which was published in The Journal of Economic History in 1969. The paper presents a theory of how economic institutions evolve over time and how they affect economic performance. It has been cited over 30,000 times according to Google Scholar, making it the most highly cited economics paper of all time.
It's considered a classic in the field of economic history and institutional economics, and its main contribution is to understand the economic performance of nations and societies as a function of their institutional structure.

This article doesn't exist!

The world headlined in May 2023:

Anwälte in New York vor Gericht – KI hatte Präzedenzfälle erfunden

Search engine vs ChatGPT

GOOGLE SEARCH - STEPS

- 1) Division of the problem into individual logical steps
- 2) Look for partial solutions to these individual logical steps
- 3) Understand the search result and check whether it is relevant
- 4) Adapt search results to your own problem
- 5) Combine individual partial solutions into an overall solution and adapt it to the context

CHATGPT - STEPS

- 1) Description of the problem including context
- 2) Ideal case: The result can be adopted directly without adapting to the context
- 3) Standard case fine-tuning: Improving the result by further instructions or issuing alternative solutions.

Context Engineering

- Context engineering involves crafting specific prompts or queries to guide the behavior of large language models,
  - allowing users to elicit desired responses or information from the model.

- Customization: It enables customization of the model's responses to align with specific objectives, contexts, or industries.

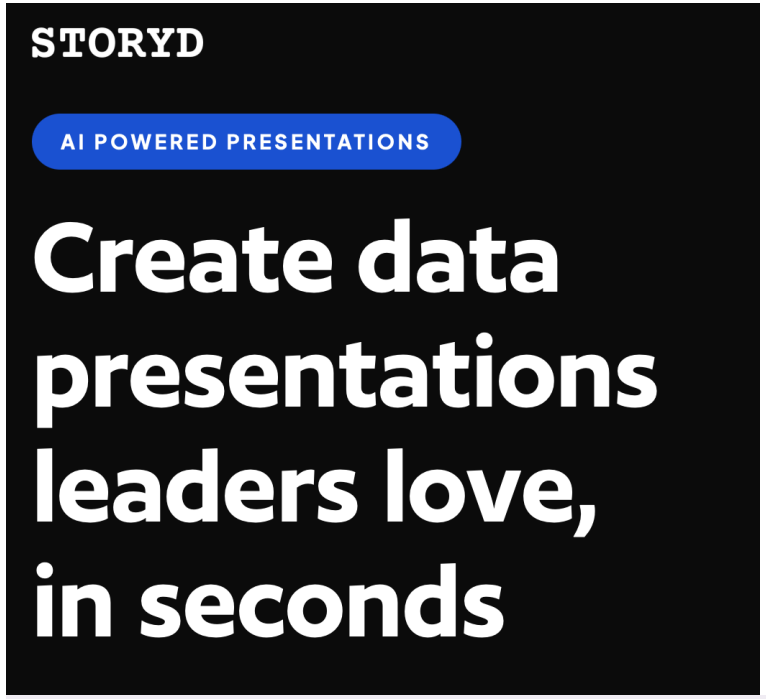
- Developers can fine-tune the model's responses by designing prompts that provide clear instructions, constraints, or examples, influencing the output.

- Crafting effective prompts often involves an iterative process of experimentation and refinement.

ChatGPT and Large Language Models

Be My Eyes

Be My Eyes uses GPT-4 to transform visual accessibility.



AI can connect people toknolwedge and each other

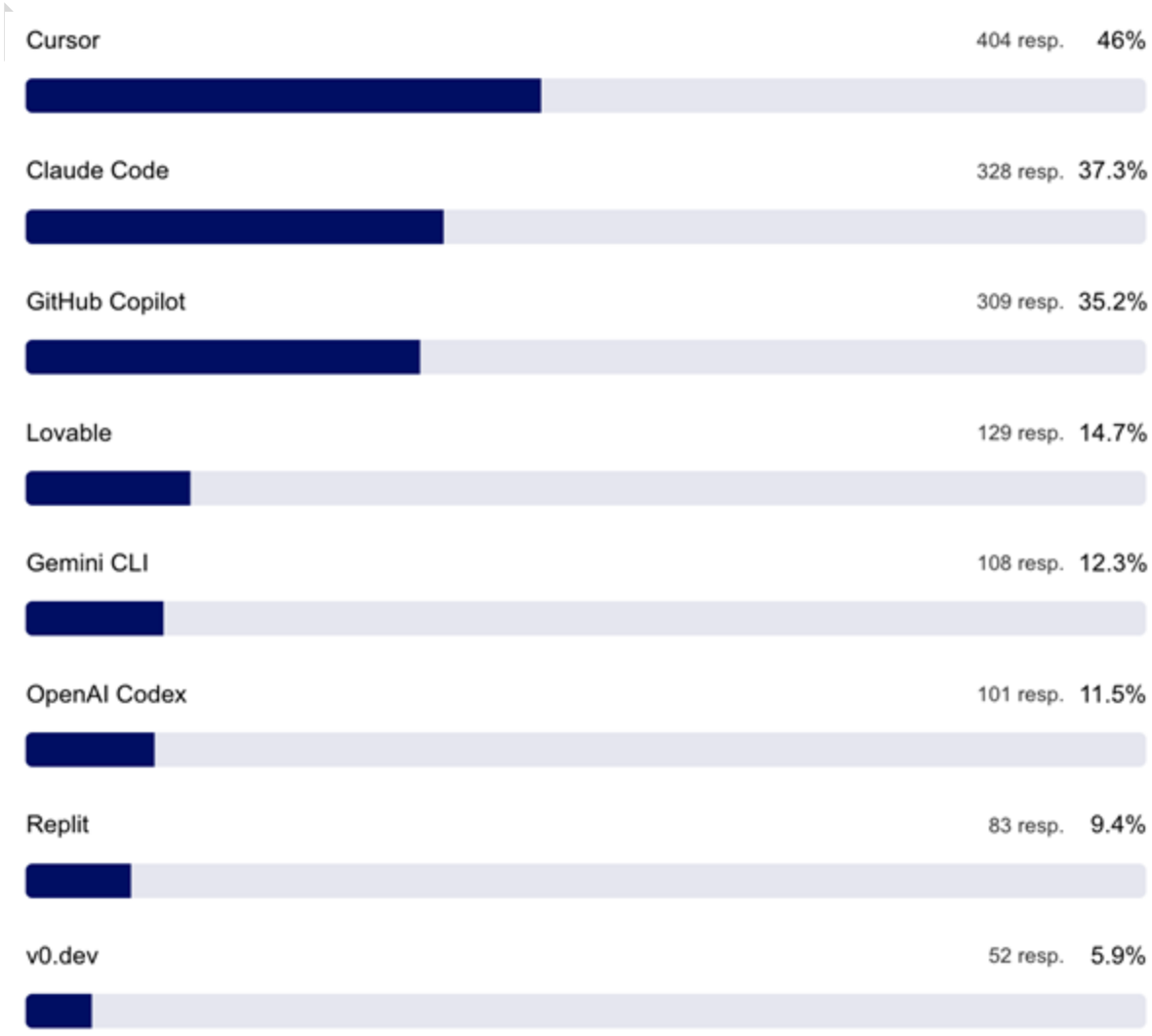
800 million can not read
300 million cannot read or can read very poorly

Write python code that uses a pretrained hugging face model to classify imag

Sure, I can try to write some python code that uses a pretrained hugging face model to classify images. Here is what I came up with:

```
# Import the necessary libraries
import torch
import torchvision
import requests
from PIL import Image
```

Developers love Cursor, Claude Code and GitHub Copilot

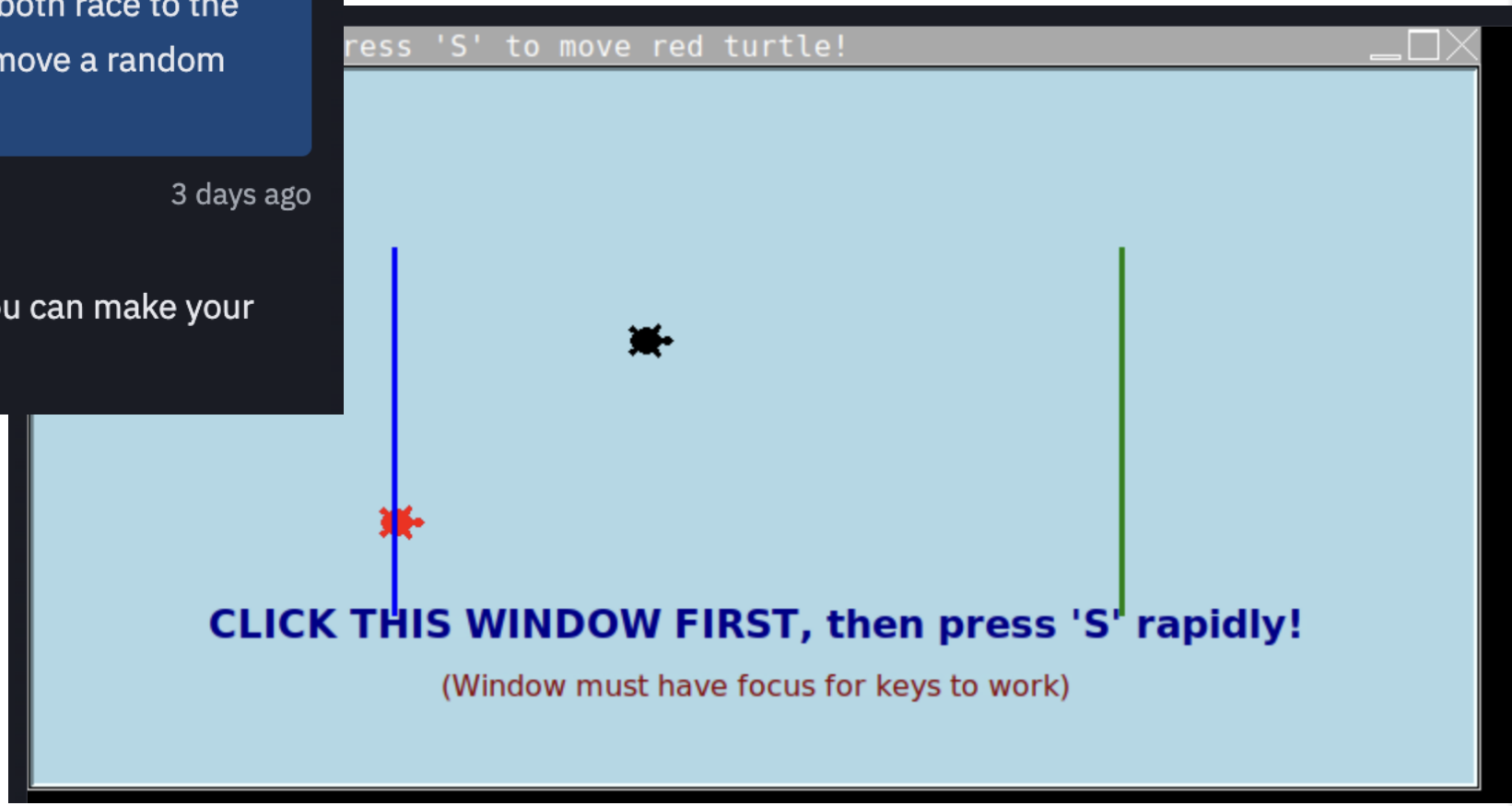


OpenAI's Codex and Gemini CLI are lagging behind =

"I have two turtles. How can I make them both race to the right side of the screen? Each one should move a random distance between 1 and 10 in each 'turn'.

3 days ago

I can help you create a turtle race! Here's how you can make your two turtles race to the right side of the screen:



Replit, Claude Code,

RockPaperScissors

Add player vs player functionality to the r... Agent 3

In progress tasks 0 / 6

I have coded a rock paper scissors game:

```
import random

human = input("Enter rock, paper, or scissors: ")
computer = random.choice(["rock", "paper", "scissors"]) #
picks randomly from rock, paper, and scissors
print(computer)

Compare the random choice with your guess
if computer== "rock":
if human == "paper":
print("human won")
elif human == "rock":
print("draw ")
else:
print("human lost")
if computer== "scissors":
if human == "scissors":
print("draw")
elif human == "rock":
print("human won")
else:
print("human lost")
if computer== "paper":
if human == "paper":
print("draw")
elif human == "rock":
print("human won")
else:
print("human lost")
```

I would like to share it with my freinds as a web app. What do I need to do?

9 hours ago

Scroll to latest

Rock

You

0

Your Choice

?

Rock

put a Rock Photo to the Rock thing not a hand

9 hours ago

2 messages & 24 actions

I'll replace the hand icon with an actual rock image for the Rock c...

Checkpoint made 9 hours ago

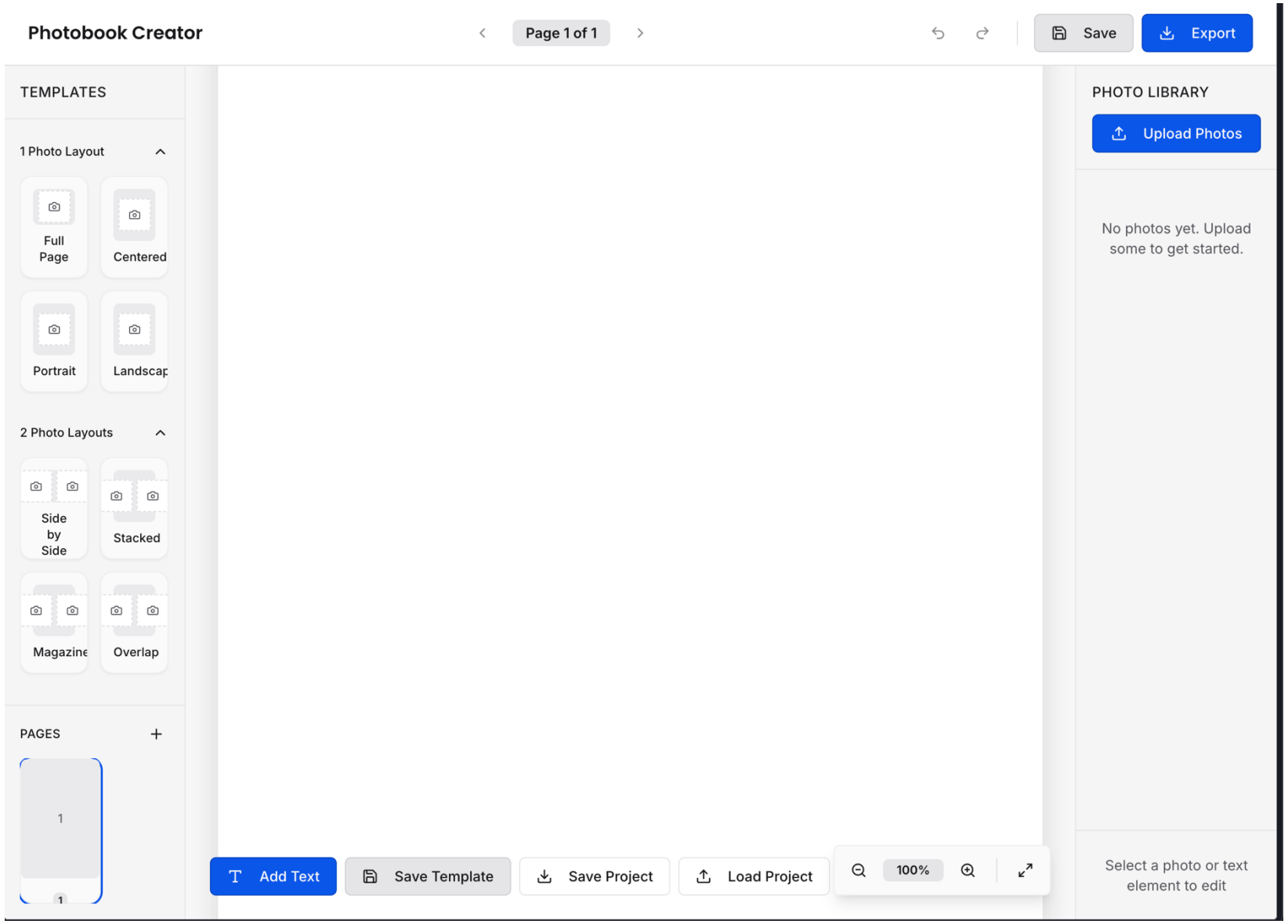
Replace rock icon with a rock image in the game interface

Rollback here | Changes | Preview

Show less

Time worked	1 minute
Work done	11 actions
Items read	335 lines
Code changed	+47 -42
Agent Usage	$0.70

Lock-in Free Photobook Editor



- Didn’t quite work

Gen AI Landscape – Applications

Lean Canvas is adapted from Business Model Canvas and is licensed under the Creative Commons Attribution-Share Alike 3.0 Un-ported License.

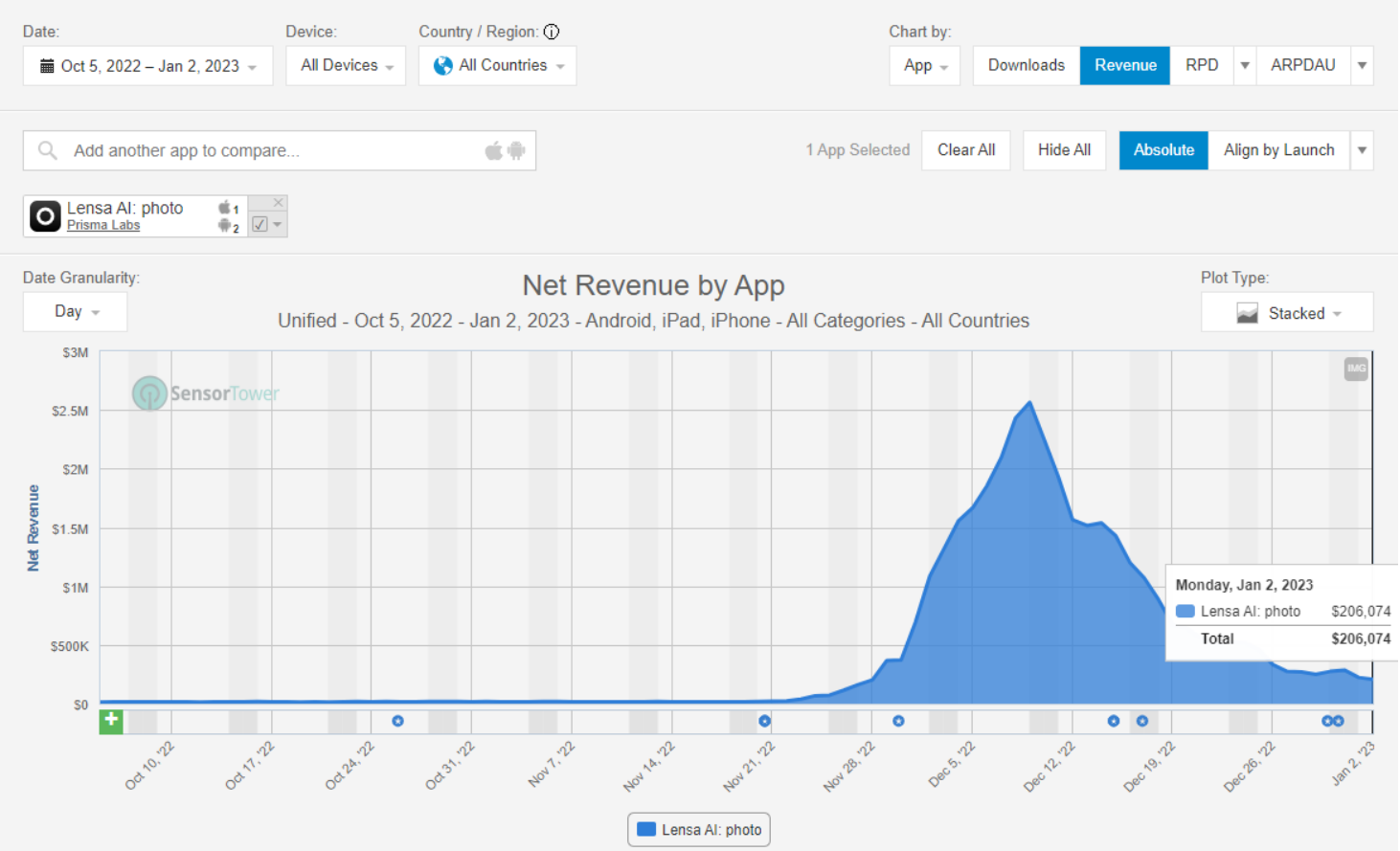
LEAN CANVAS

(No) Unfairadvantage

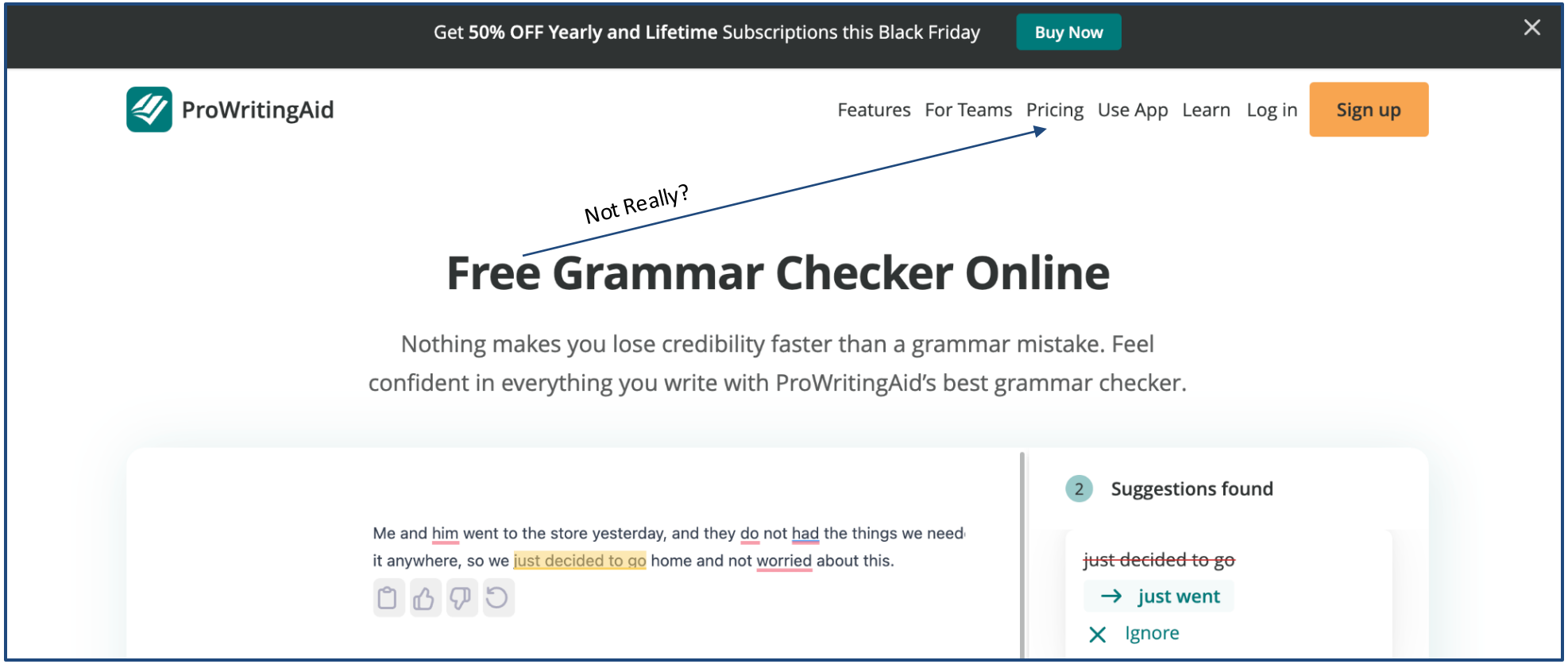
Sasha Kaletsky
@SashaKaletsky

UPDATE: Lensa AI revenue is down to ~$206k per day, down 92% from its peak

AI was fun while it lasted



Example StartUp: Pre GenAI



Could Take Months/Years to Build Pre Late 2022

Data Collection → Labeling → Training → Tuning → …

Post GenAI: 10 Lines of Code

```
import openai

openai.api_key = '299HoJHC4GhC543h32JM4BHS23t83'

def grammar_check(text):
   prompt_template = \
       f"""Given the following text, return a json file that maps \
           each grammatical error in the text (given by its index) to \
          its correction: {text}"""

   response = openai.Completion.create(
       engine="gpt-3.5-turbo-instruct", prompt=prompt_template,
       max_tokens=150,temperature=0.7, n=1,stop=None
   )

   return response.choices[0].text.strip()

# Example usage
input_text = "Your input text goes here."
result = grammar_check(input_text)
```

1) We wrap user input with a prompt

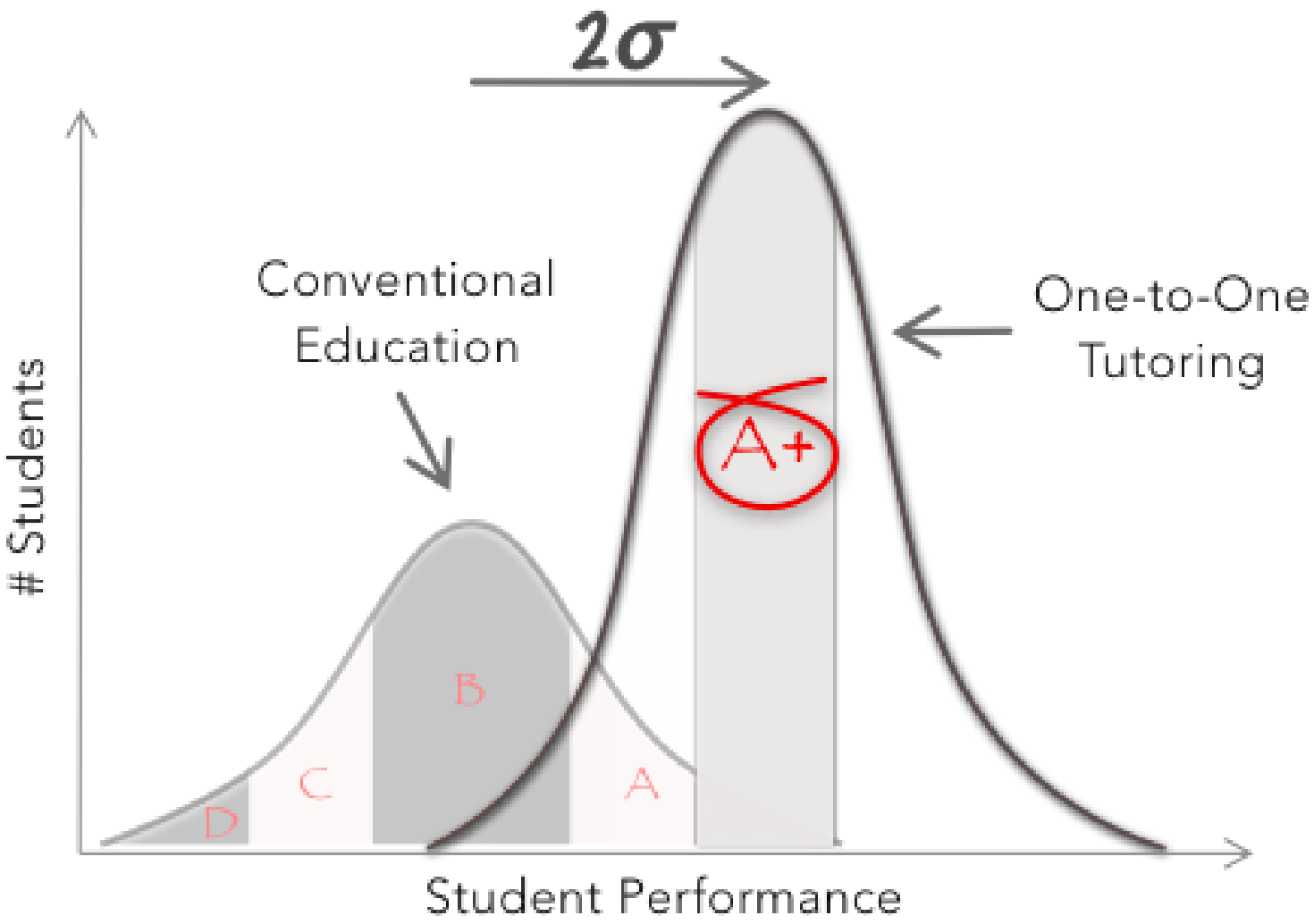
2) Send it to OpenAI GPT 3.5 Turbo Instruct model (≈ level to ChatGPT)

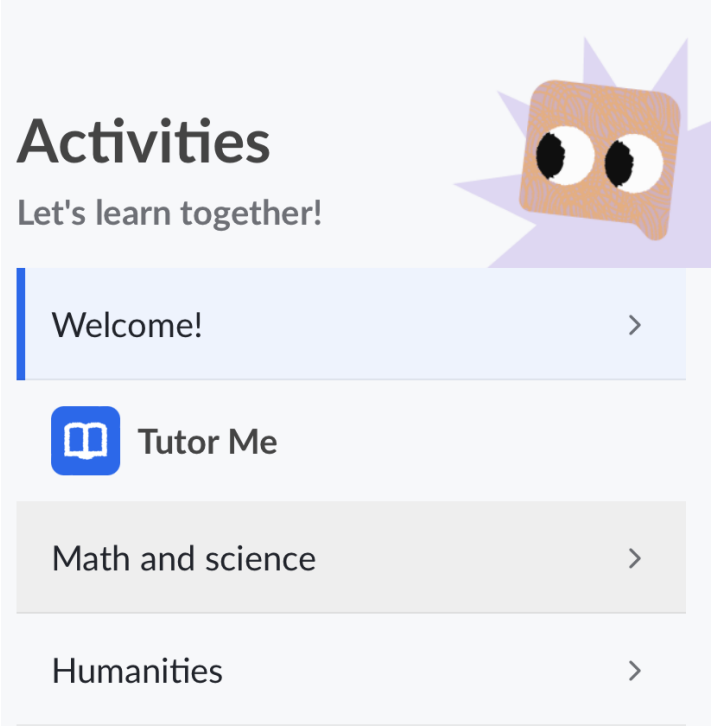
Costs $$

3) Access the most probable response and remove leading/trailing whitespace

4) Deploy this function on our own server that our website UI can talk to (JSON is well understood)

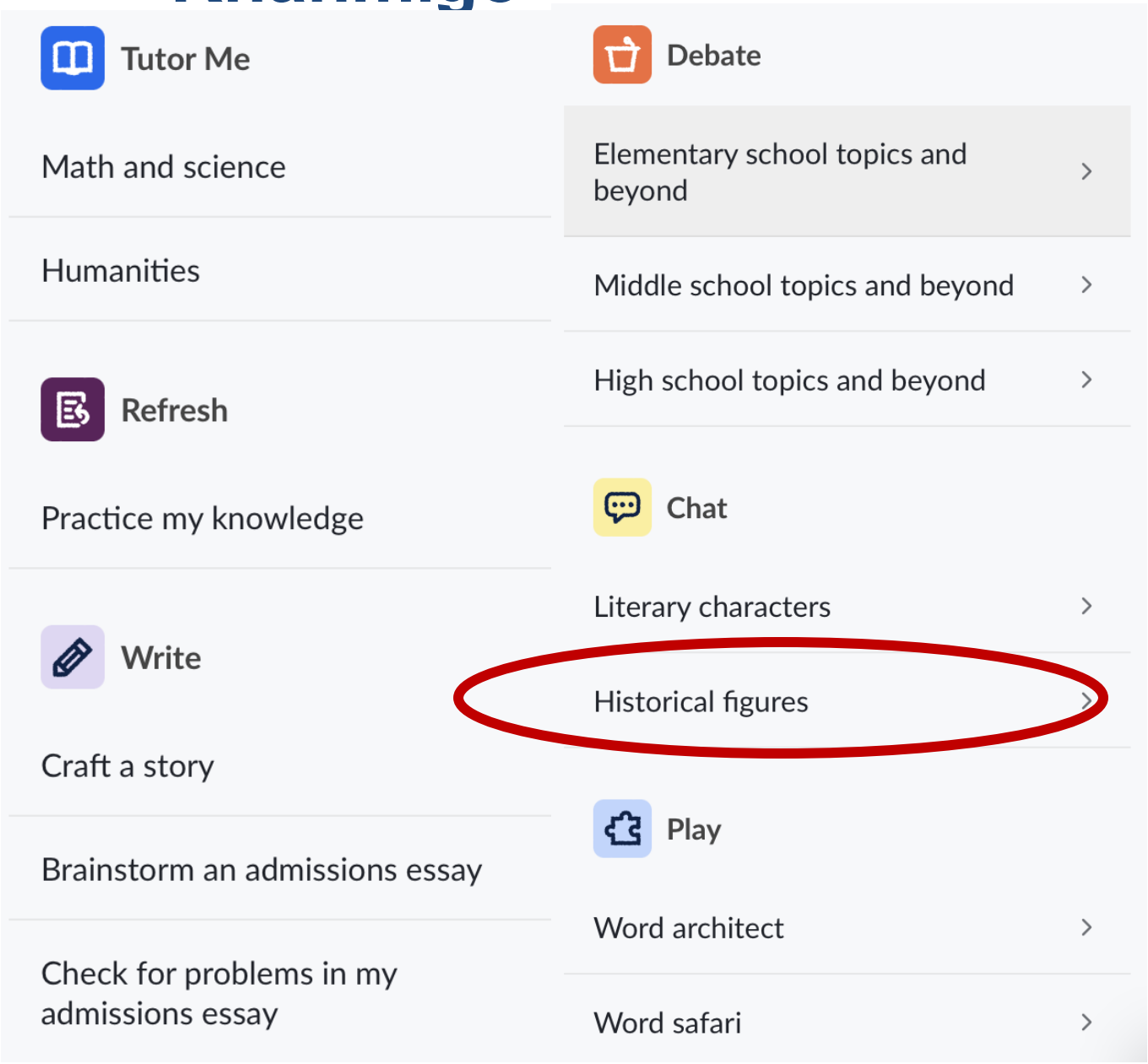
AI Tutor for education





Bloom, Benjamin S (June–July 1984). "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring

Khanmigo

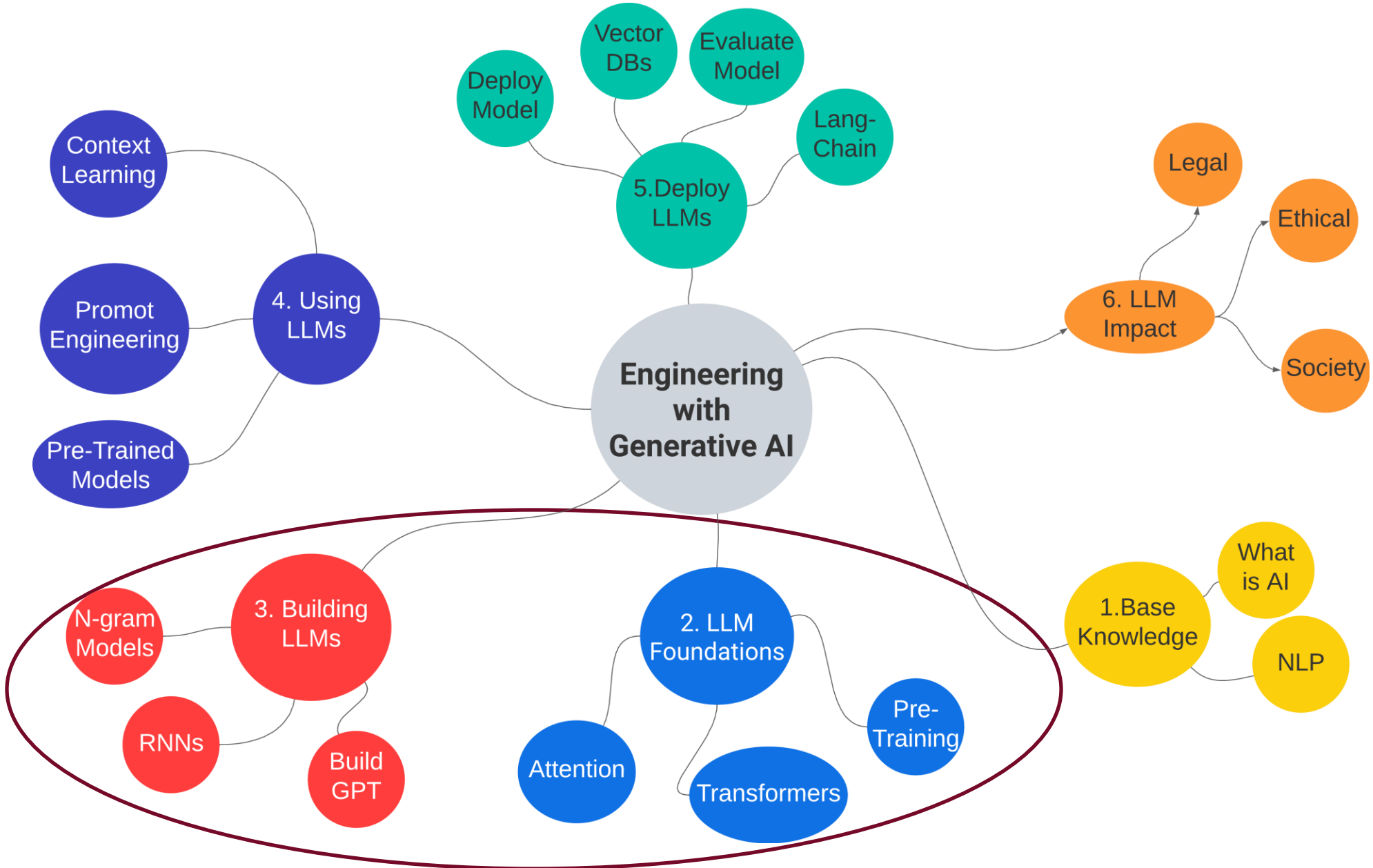


•I would love to meet Marie Curie/Albert Einstein/**Galileo Galilei**

Of course, my dear friend. **In my time, most people believed that the Earth was in the middle of everything and that the Sun moved around the Earth. This is what we call the geocentric model. But I found out that this wasn't true**. I discovered that the Earth actually moves around the Sun. This is what we call the heliocentric model. I knew this because I watched the stars and planets very carefully with my telescope. Even though some people didn't agree with me, I knew I was right. Now, do you like looking at the stars and planets too?

Engineering with Large Language Models

3. A Glimpse Under the Hood



Task Specific NLP Models

- Define the Task: Sentiment Analysis
- Collect a lot of data.
  - Collect labeled data for positive, negative and neutral reviews.
- Train a task-specific model for sentiment classification

From General Purpose to Task Specific Again

General Purpose Dog

Drink - Eat – Walk – Bark

Special Dog

Police Dog

Guide Dog

Foundational Large Language Models

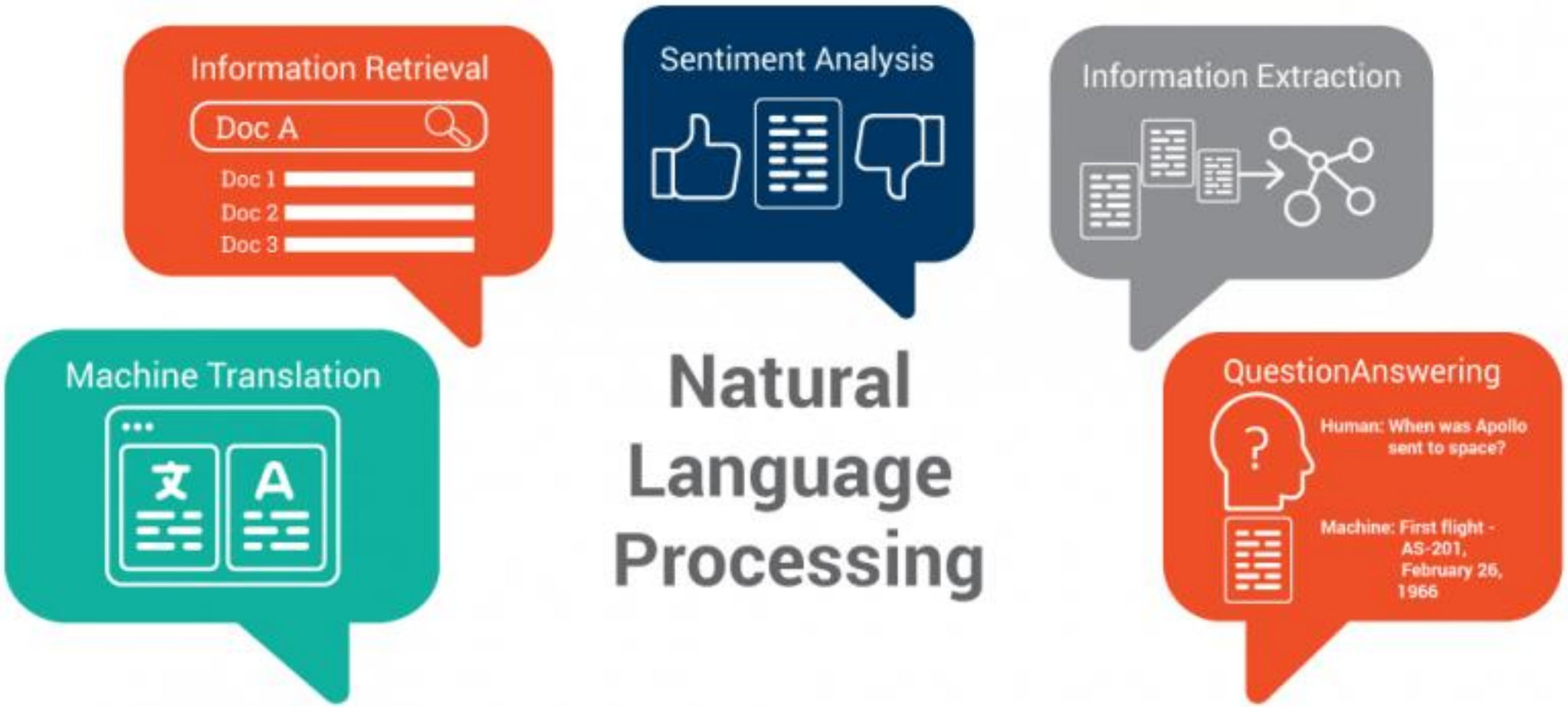
Large Language Models are trained to solve common language problems like..

Text Classification

Question Answering

Summarization

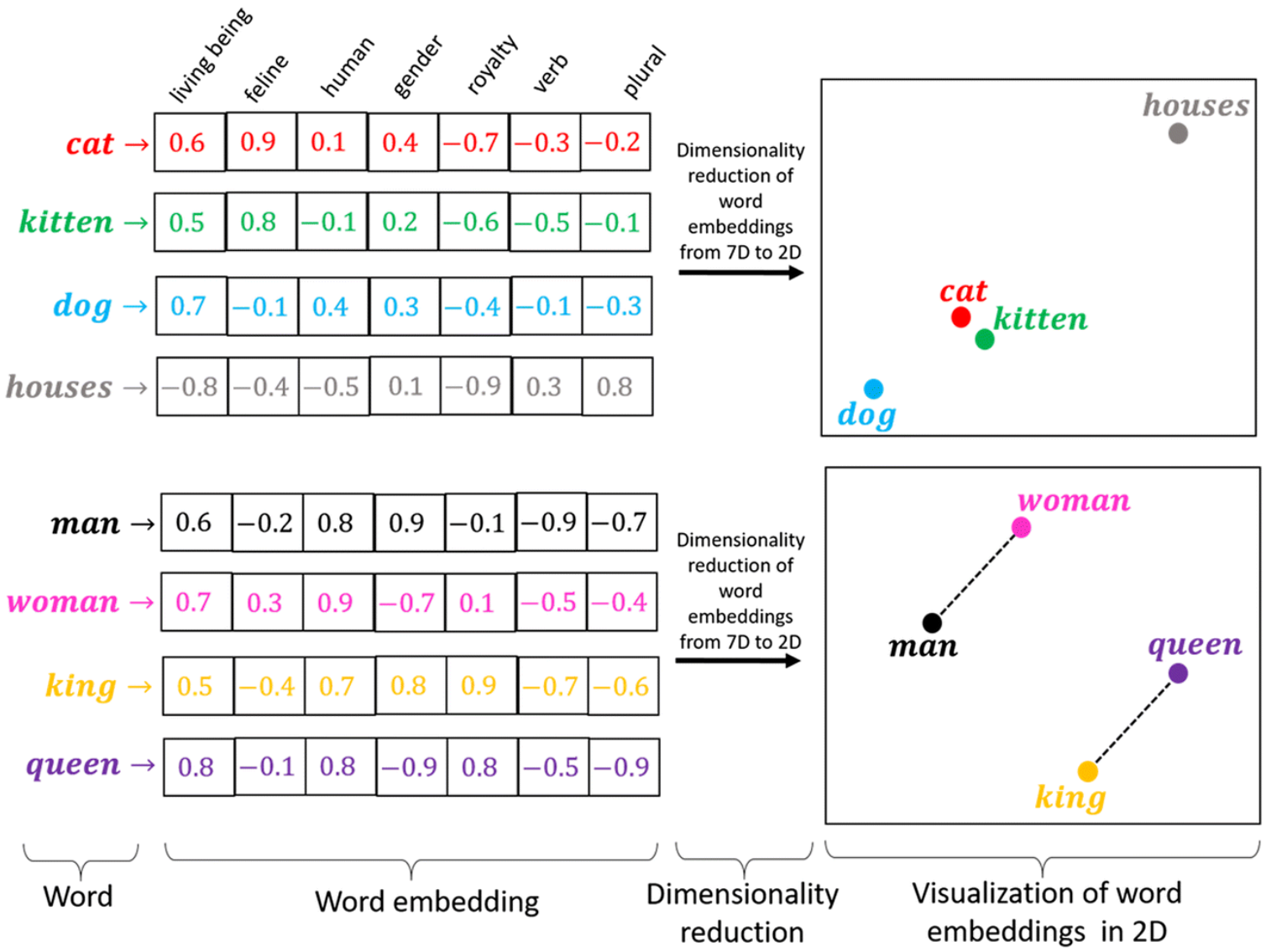
Natural Language Processing



Natural Language Processing

Image by NLPlanet

Word Embeddings



Training a Generative AI

The generation process

My favourite food is

prompt

Pizza

Pasta with tomato sauce

Bratwurst with chips

Output

How does it work?

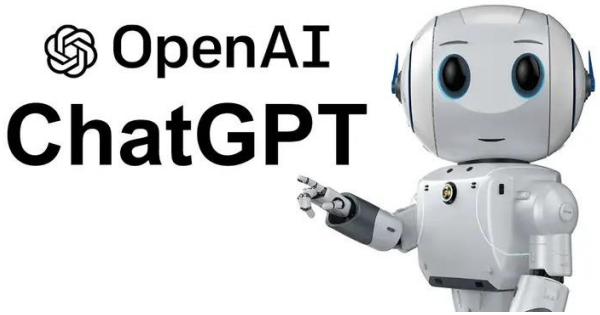
The concept of Generative AI is based on the use of supervised learning, which is applied repeatedly to predict the next word.

My favourite food is pasta with tomato sauce.

Input (A)	Output (B)
My favourite foods are	Noodles
My favourite food is noodles	with
My favourite food is pasta with	Tomato sauce

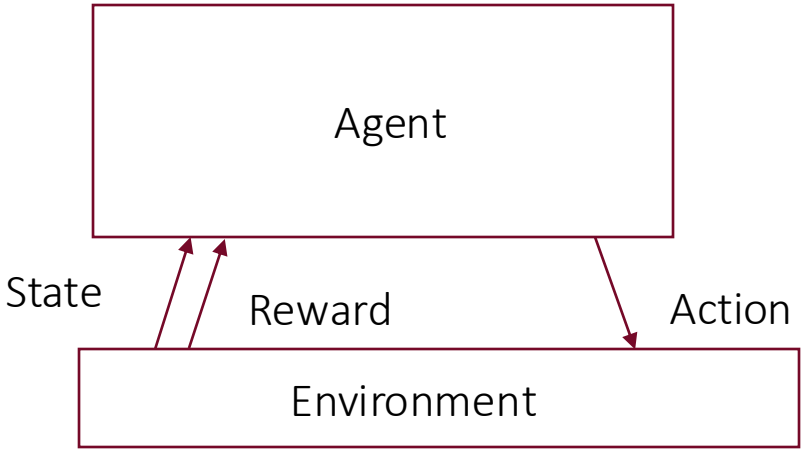
If we run a very large AI model on a very large amount of data (hundreds of billions of words), we get a Large Language Model like ChatGPT

How does ChatGPT work?



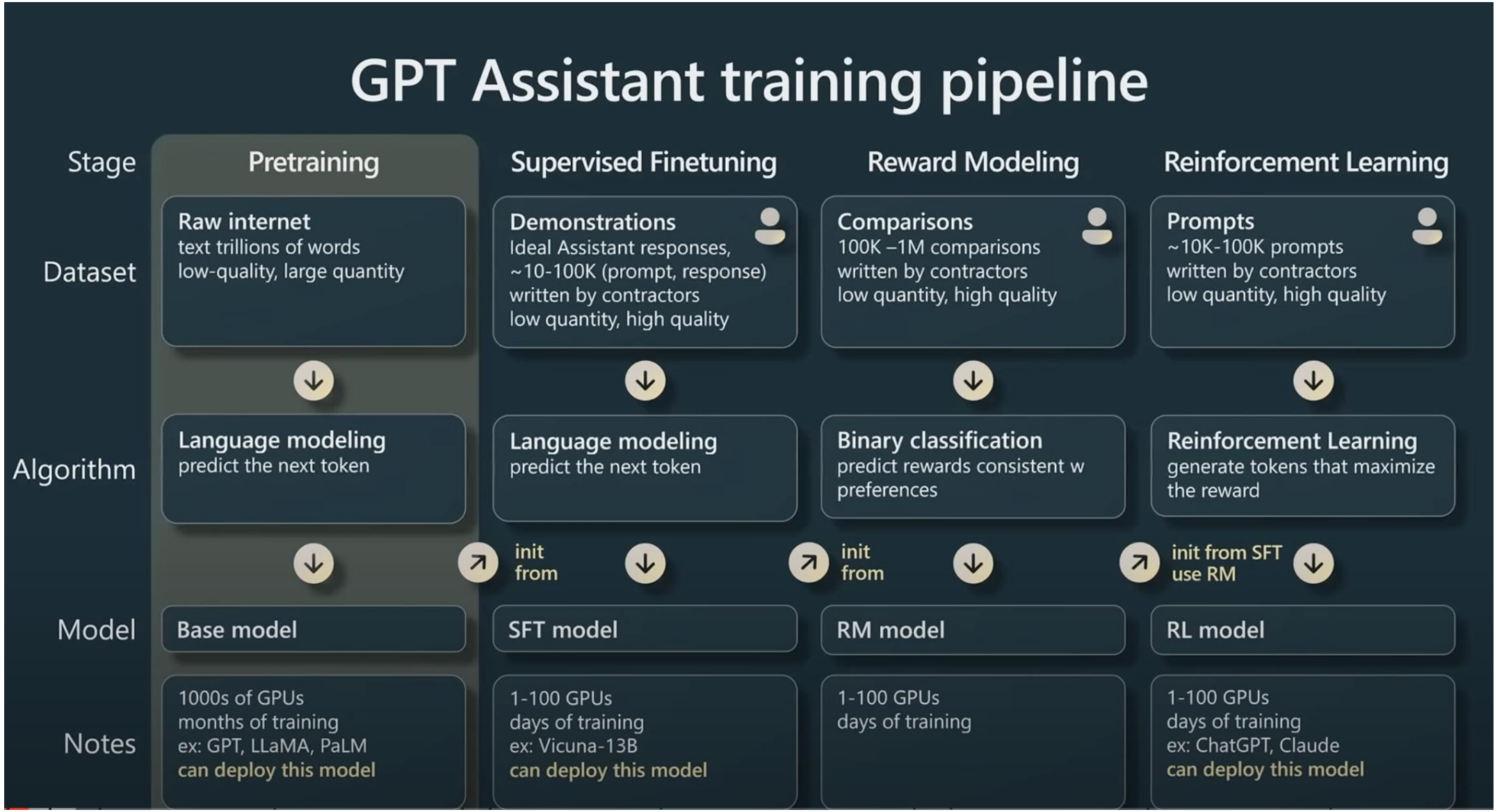
=

+



1. Large language model
2. Robust, proven reinforcement learning methodology
3. Lots of human annotated data for the RL method reward model.

How ChatGPT is built



https://www.youtube.com/watch?v=kCc8FmEb1nY

Step 1: Which training data does ChatGPT use

We don’t know for Facebooks Sprachmodel LLamA:

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

https://paperswithcode.com/paper/llama-open-and-efficient-foundation-language-1/review/

21 days on 2048 A100 GPUS – energy costs ca. 5 000 000 $!!!

A Raw Model can be useful too

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?
A: 54

Q: where does she live?
A:

GPT-2 is "tricked" into performing a task by completing the document

https://www.youtube.com/watch?v=kCc8FmEb1nY

Step 2: Supervised fine tuning

Finetuning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
cheese =>                         ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← example
cheese =>                         ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                         ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer        ← example #1
↓
gradient update
↓
peppermint => menthe poivrée      ← example #2
↓
gradient update
↓
...
↓
plush giraffe => girafe peluche   ← example #N
gradient update
cheese =>                         ← prompt
```

It is becoming a lot more accessible to finetune LLMs:

- Parameter Efficient FineTuning (PEFT), e.g. LoRA
- Low-precision inference, e.g. bitsandbytes
- Open-sourced high quality base models, e.g. LLaMA

Keep in mind:

- Requires a lot more technical expertise
- Requires contractors and/or synthetic data pipelines
- A lot slower iteration cycle
- SFT is achievable
- RLHF is research territory

[Language Models are Few-Shot Learners, Brown et al. 2020]

Reward modelling

Summarise generative AI in three sentences

1. Generative AI is a type of artificial intelligence that learns from existing data to create new, realistic content such as images, text, music, and more1. It uses complex mathematical techniques and significant computing power to generate this content, often in response to natural language prompts1. The technology has a wide range of applications, including innovations in product design, business process optimization, and content creation.
2. Generative AI is a type of AI that can create new content, such as text, images, audio and video, by learning from existing data. Generative AI uses foundation models, which are large AI models that can perform multiple tasks and adapt to specific domains. Generative AI has many applications in content creation, data analysis, customer service, product design and more

Loss function is not prescribed

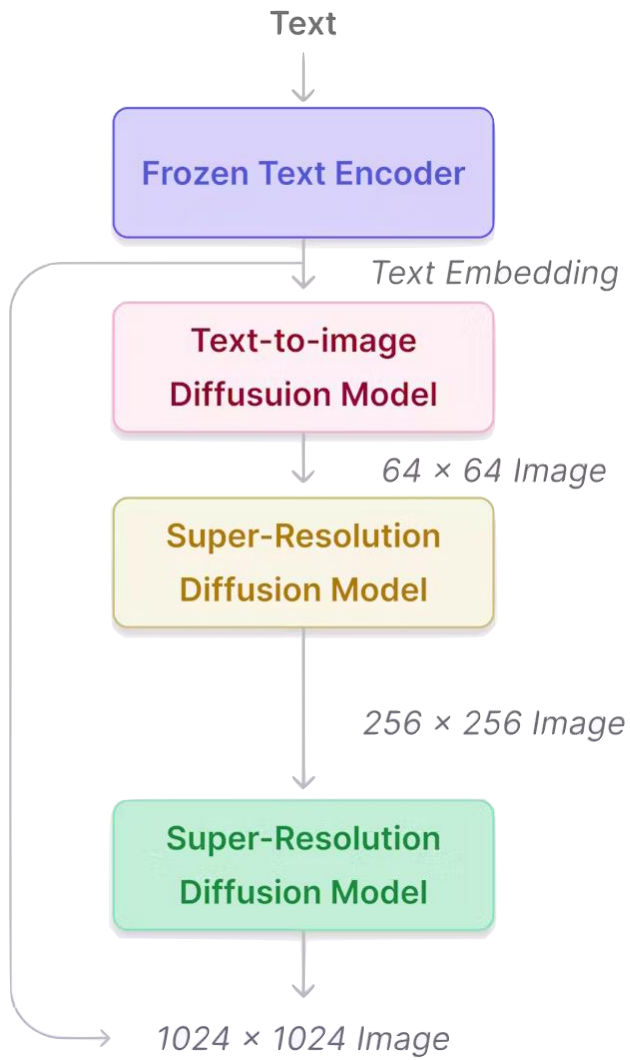
But how ‘open’ are ‘open source’ models?

With open source commanding considerable community support and becoming a hot button regulatory issue, some researchers have suggested that the term is often used misleadingly. It can be used to lump together vastly different openness practices across weights, datasets, licensing, and access methods.

Project	Availability						Documentation						Access	
	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
AmberChat	✓	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
OpenChat 3.5 7B	✓	✗	✓	✗	✓	✓	~	✓	✓	✓	~	✗	✓	~
Pythia-Chat-Base-7...	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
Cerebras GPT 111...	~	✓	✓	✓	✓	~	✗	✓	~	✗	✗	✓	✗	✓
RedPajama-INCITE...	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~
dolly	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✗
Tulu V2 DPO 70B	✓	✗	~	✓	✓	~	~	~	✓	✗	~	~	✗	✓
MPT-30B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	~	✗	✓	~
MPT-7B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	✓	✗	✓	✗
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	~	✓
Vicuna 13B v 1.3	✓	~	✓	✗	✗	~	✓	✗	✓	✗	~	✗	✓	~
minChatGPT	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	✗	✓
ChatRWKV	✓	~	✓	✗	✗	✓	~	~	~	✗	✗	✗	✓	~
BELLE	✓	~	~	~	~	✗	~	✓	✓	✗	✗	~	✗	✗
WizardLM 13B v1.2	~	✗	~	✓	✓	~	~	✓	✓	✗	✗	✗	✗	✗
Airoboros L2 70B G...	~	✗	~	✓	✓	~	~	~	✗	✗	~	~	✗	✗
ChatGLM-6B	~	~	✓	✗	✗	✓	~	~	✗	~	✗	✗	✗	✓
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✓

Other Applications of Generative AI

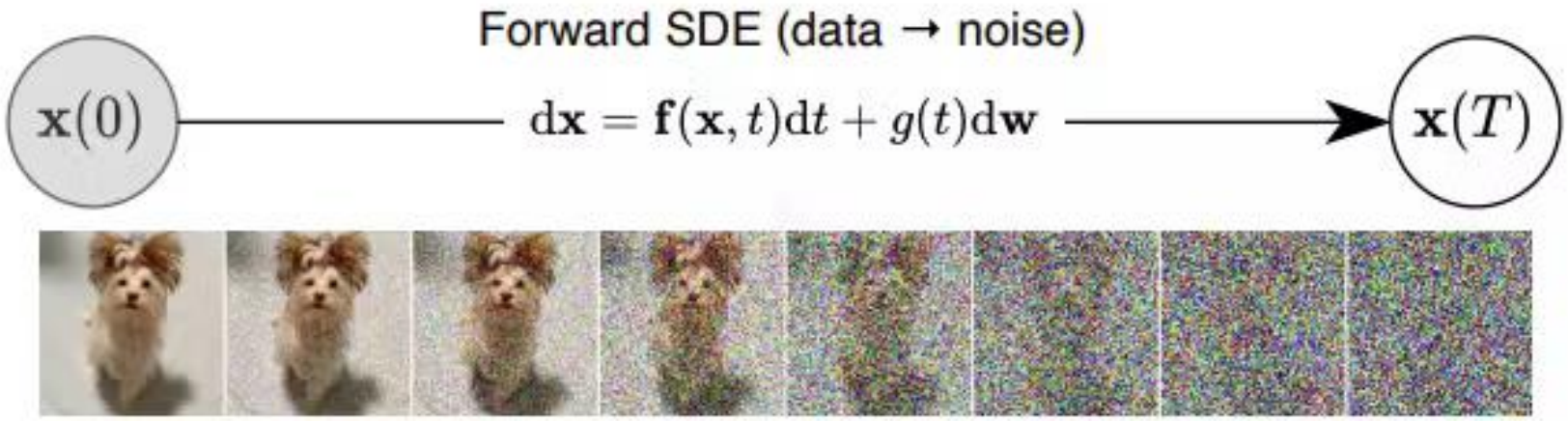
- Image synthesis and editing (2014: DeepDream (Google), faces with StyleGan 2018).
- Music and audio generation (2016)
- Data augmentation and synthetic data (2002)
- Video generation, editing and prediction,
- Healthcare and drug development
- Art generation and fashion design (2019 AI artworks at exhibitions).
- Scientific research and simulation and much more... (2017)



"A Pig wearing a top hat and a pink bow tie."

https://encord.com/blog/diffusion-models/

Training Diffusion Models



[88, 672, 15255, 2024, 7401] („Little yorkshire terrier“)

https://encord.com/blog/diffusion-models/

DSA Work on the topic

(a) Base LFM model (b) Vanilla Adjoint Matching (c) Joint Adjoint Matching

Figure 10: Non-curated independent samples from LFM model conditioned on class "macaw" and using guidance scale 4.0. Models were fine-tuned to maximize PickScore using the prompt "close-up pop art of a macaw parrot".

- Far out of distribution
- Methods can be used for simulating physics

Text-To-Image

DALL·E

Transform your ideas into amazing images

1 **Logo Creator**
Use me to generate professional logo designs and app icons!
By Chase Lean

2 **image generator**
A GPT specialized in generating and refining images with a mix of professional and friendly tone.image generator
By NAIF J ALOTAIBI

3 **Cartoonize Yourself**
Turns photos into Pixar-style illustrations. Upload your photo to try
By karenxcheng.com

4 **Super Describe**
Upload any image to get a similar one using DALL·E 3 along with the detailed prompt!
By bestaiprompts.art

5 **Consistent Character GPT👉 Fast & High Quality⚡**
Your creative partner for generating characters in different poses, expressions, styles, and scenes. No prompt needed, jus...
By ailemonacademy.com

6 **Tattoo GPT**
Tattoo GPT designs your tattoo. It assists you in refining your tattoo ideas, suggests designs, generates visual previews of the...
By Michael Moncada

Mehr laden

ideogram

What do you want to create? Your prompt cannot be empty.

Magic Prompt: Auto | On | Off

Magic Prompt enhances your initial prompt to improve image variety and richness. It can also translate to English.

16:9

Visibility: Public | Private

photo | illustration | 3d render | typography | cinematic | poster | ...

Generate

LavenderDreams
1h ago
16

mudeerhassan
31m ago
2

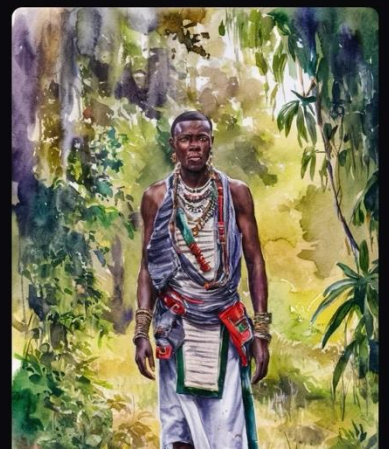
ashinzanaka
30m ago
2

mihail2023
1h ago
2

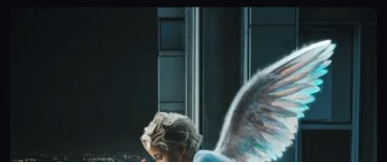
jose1964
1h ago
4

newssvet
2h ago
14

codrut
3m ago
1



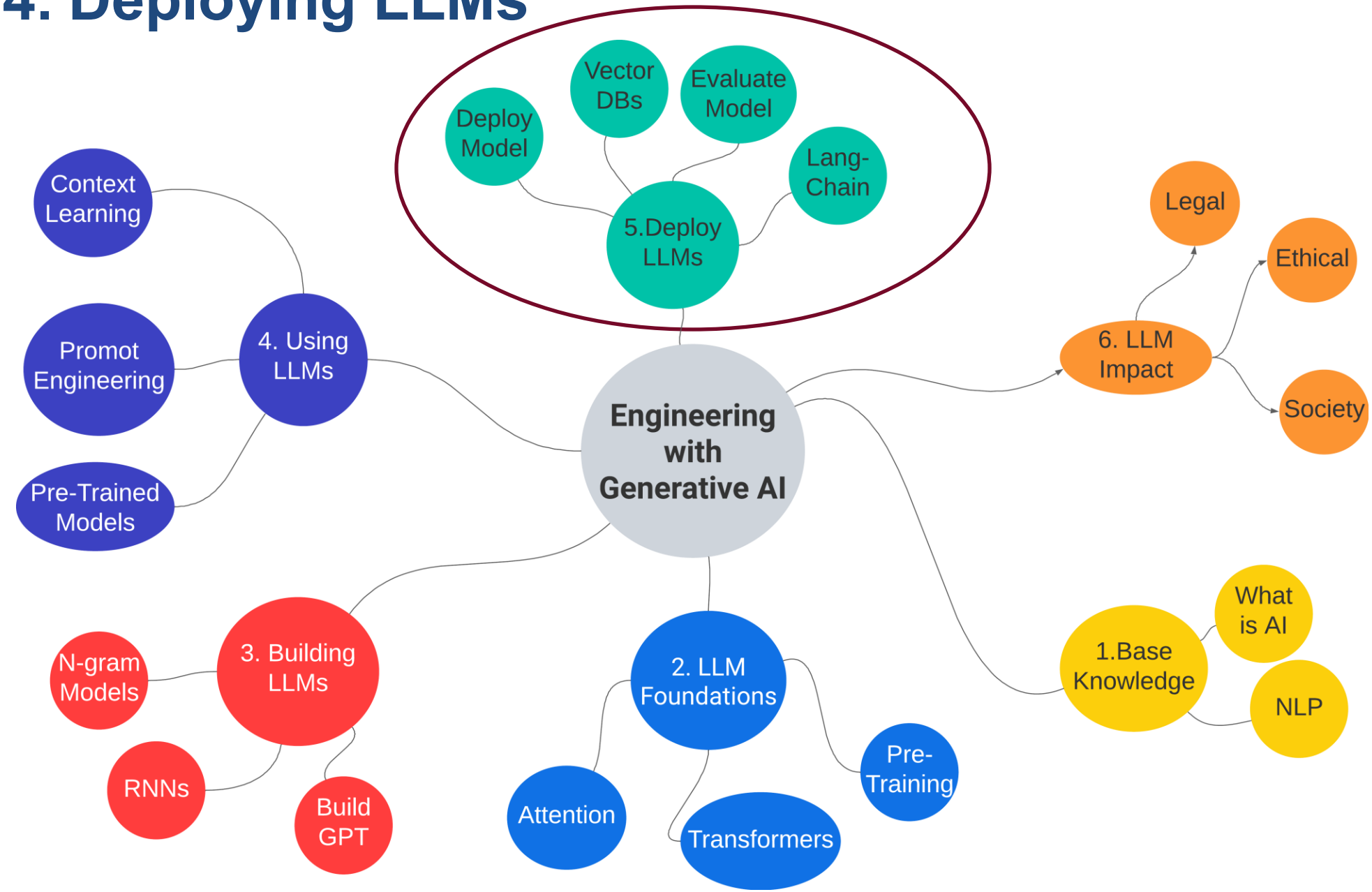
omarkioum
4h ago
4



rollandglow
4m ago
1

imagecraftai
2h ago
10

4. Deploying LLMs



Numbers every developer should know

Numbers Everyone Should Know

L1 cache reference	0.5 r
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	100 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	10,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from network	10,000,000 ns
Read 1 MB sequentially from disk	30,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

http://static.googleusercontent.com/media/research.google.com/en/us/people/jeff/stanford-295-talk.pdf

Numbers 💯 every LLM Developer 🛠️ should know *

🪄 Prompts

40-90%	Amount saved by appending "Be Concise" to your prompt
1.3	Average tokens per word

💡 Training and Fine Tuning

~$1 million	Cost to train a 13 billion parameter model on 1.4 trillion tokens
<0.001	Cost ratio of fine tuning vs training from scratch

💰 Price

~50	Cost Ratio of GPT-4 to GPT-3.5 Turbo
5	Cost Ratio of generation of text using GPT-3.5-Turbo vs OpenAI embedding
10	Cost Ratio of OpenAI embedding to Self-Hosted embedding
6	Cost Ratio of OpenAI base vs fine tuned model queries
1	Cost Ratio of Self-Hosted base vs fine-tuned model queries

🧠 GPU Memory

16GB	V100 GRAM capacity
24GB	A10G GRAM capacity
40/80GB	A100 GRAM capacity
2x number of parameters	Typical GPU memory requirements of an LLM for serving
~1GB	Typical GPU memory requirements of an embedding model
>10x	Throughput improvement from batching LLM requests
1 MB	GPU Memory required for 1 token of output with a 13B parameter model

* Check out bit.ly/llm-dev-numbers for how we calculated the numbers

Presented by RAY & anyscale with ❤️ Join the community ray.io or Request a Trial anyscale.com/signup today

Numbers

Method	7B Parameters	13B Parameters	30B Parameters	70B Parameters
Full Fine-Tuning (16-bit)	67 GB	125 GB	288 GB	672 GB
LoRA (16-bit)	15 GB	28 GB	63 GB	146 GB
QLoRA (8-bit)	9 GB	17 GB	38 GB	88 GB
QLoRA (4-bit)	5 GB	9 GB	20 GB	46 GB

(Data sourced from multiple analyses of fine-tuning memory requirements)

The Complete Guide to GPU Requirements for LLM Fine-Tunin...

runpod.io/blog/llm-fine-tu...

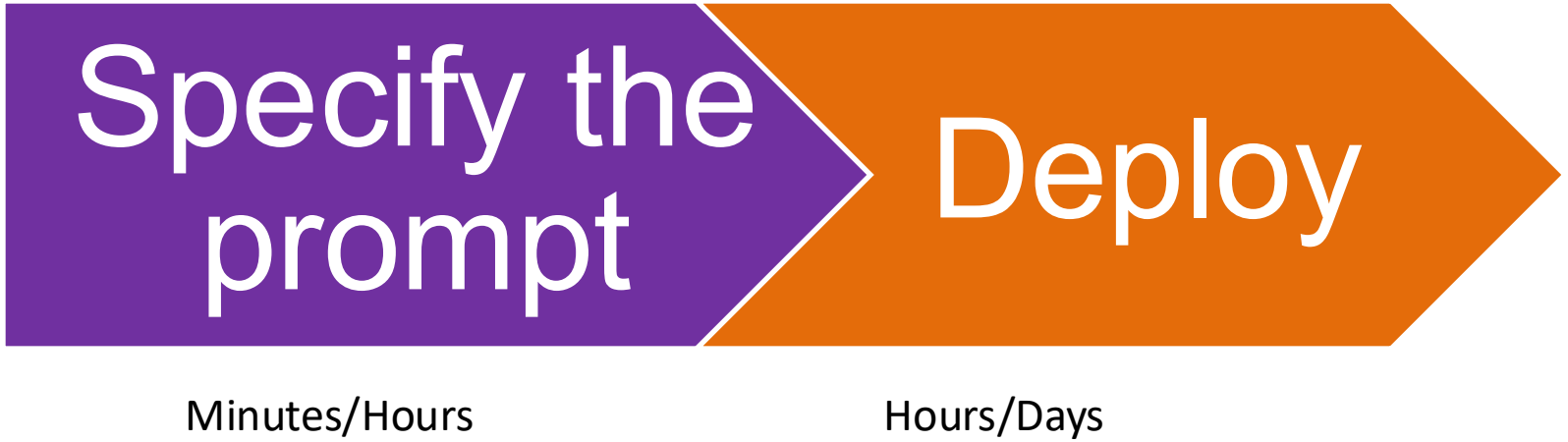
How much VRAM do I need for LLM model fine-tuning? | Modal Blog

modal.com/blog/how-muc...

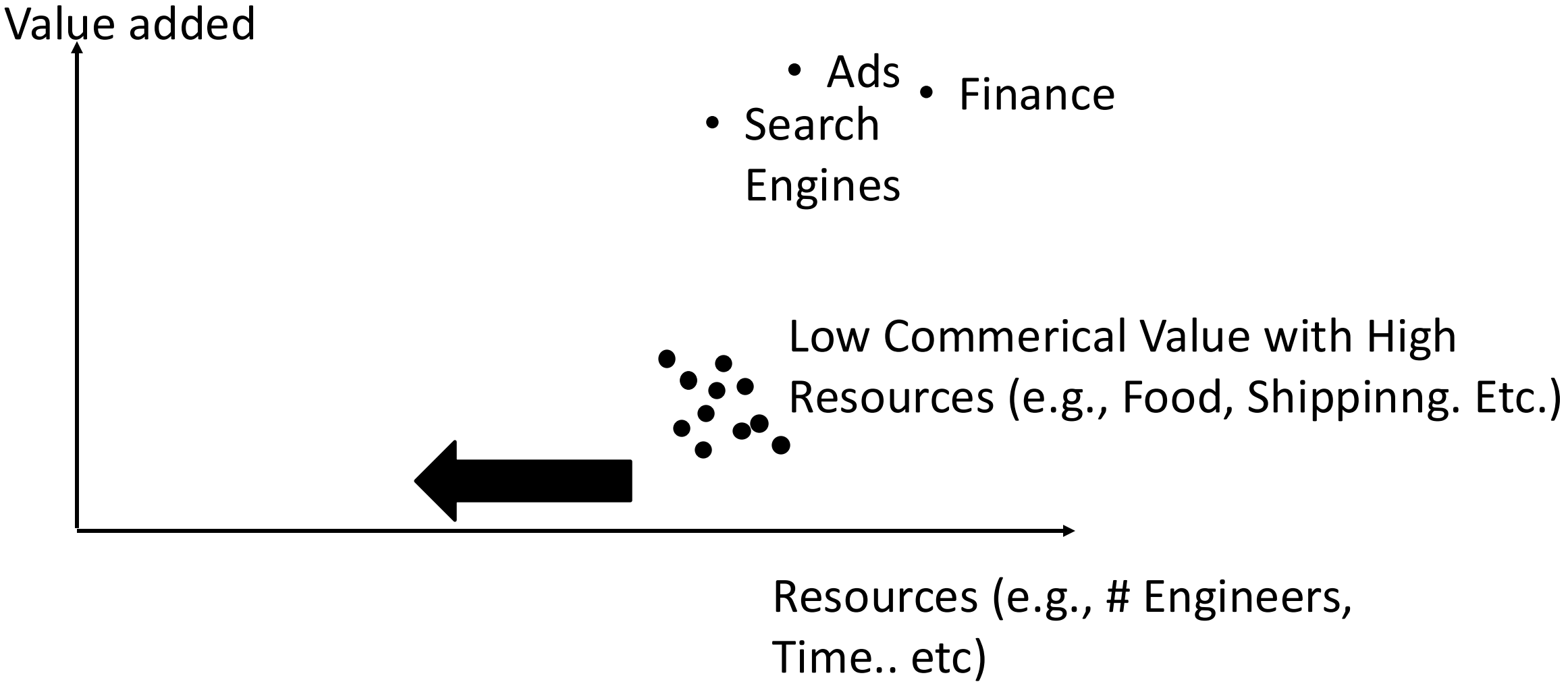
Supervised Learning vs Prompt Based

Supervised Learning

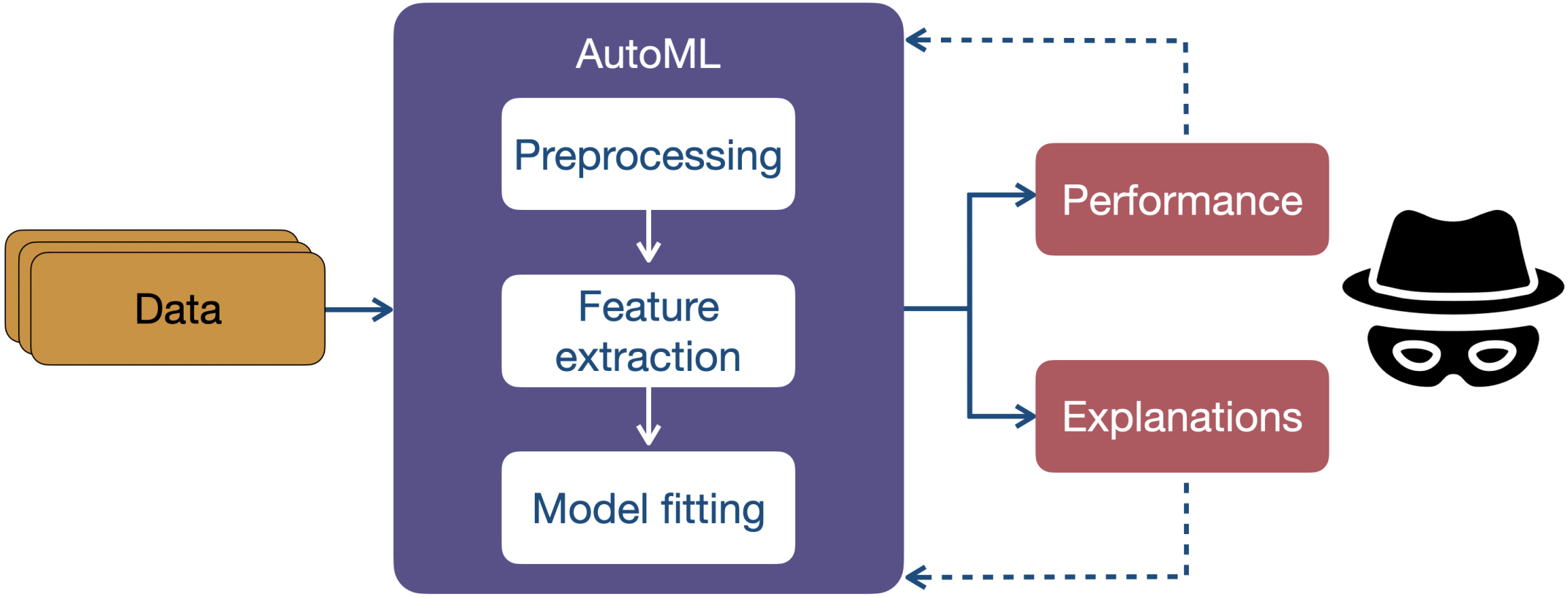
Prompt-Based AI



Hurdles for AI Applications

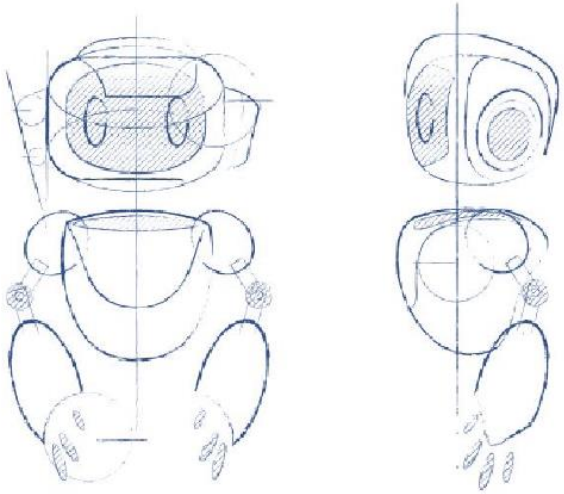


Our Research: Threats from AutoML



"The only statistics you can trust are the ones you have falsified yourself"
- *Winston Churchill*

Our Research: Data Analytics via LLMs

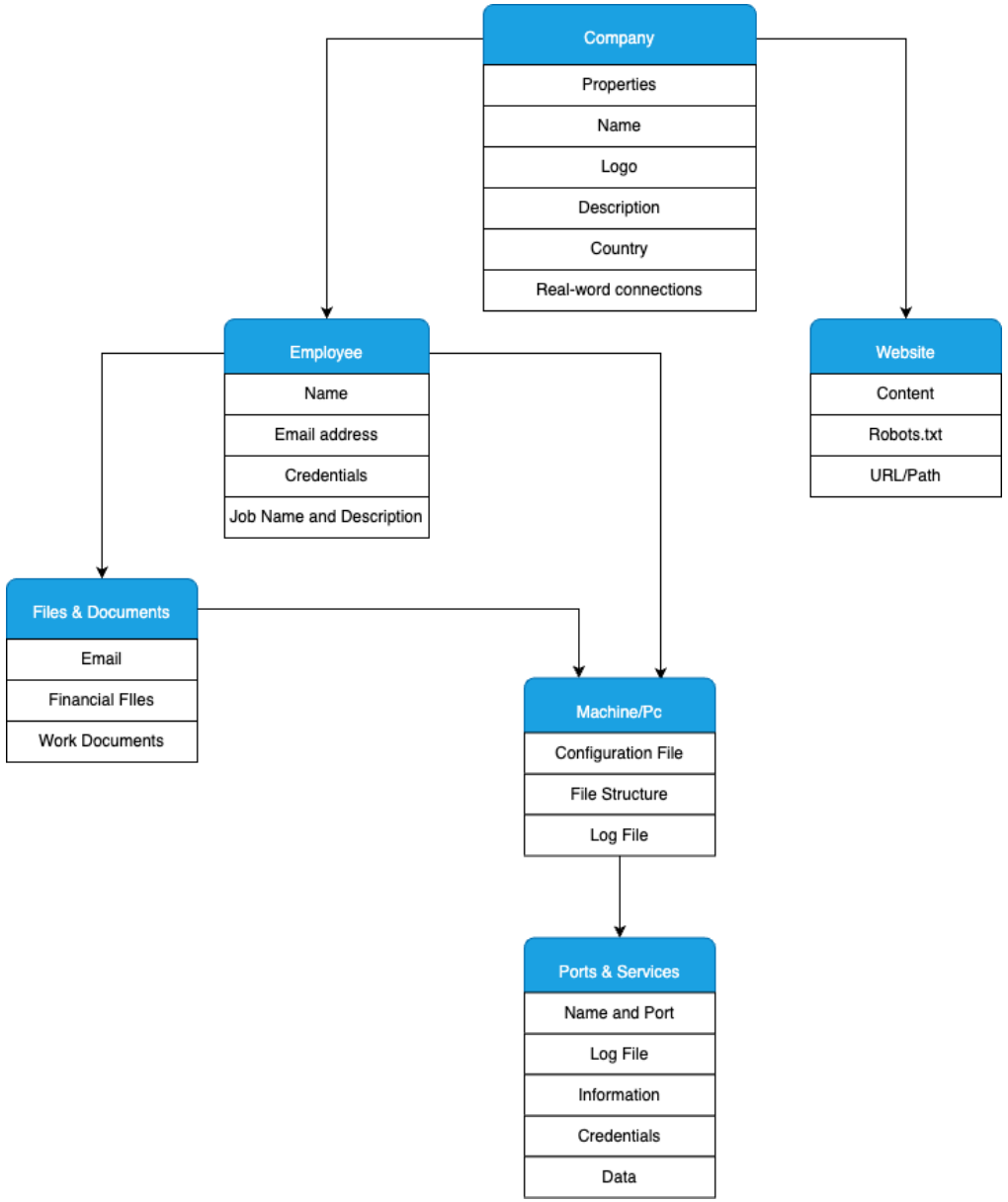


Design of Data Science Pipelines
Simplified, adapted to your needs.

Alpha release is coming soon.

My training data consists of a large corpus of text, which allows me to generate language that is grammatically correct and semantically coherent. However, I do not have direct access to the data that is being analyzed, nor do I have knowledge of the specific problem domain or the characteristics of the data.

Our Research: Honeypots



Which prompts generate good honeypots?
How do you measure how good a honeypot is?
How much can this be automated?

Quantitative knowledge retrieval from large language models

David Selby
Data Science & its Applications

Elicitation of prior distributions

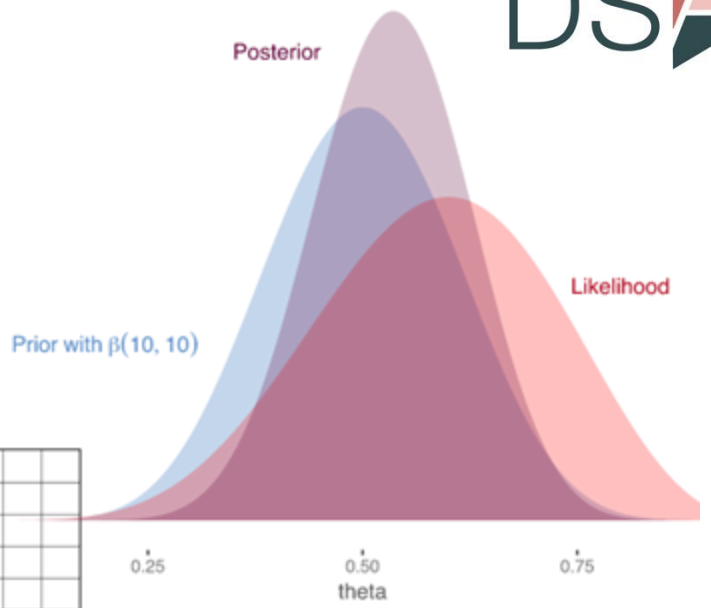
Sheffield elicitation framework

Roulette method

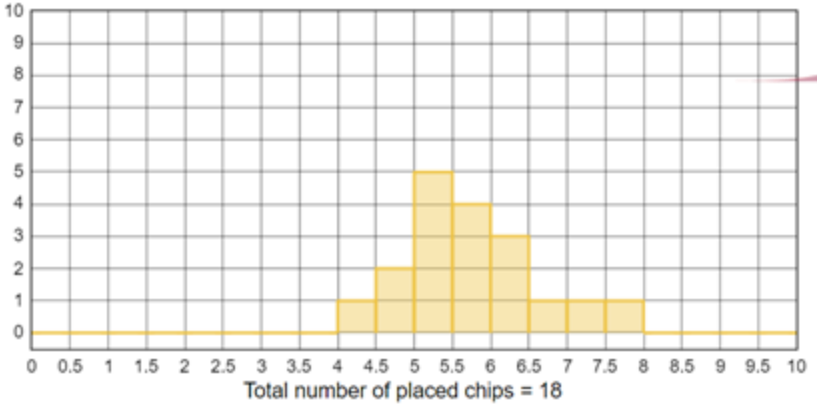
Cook's method

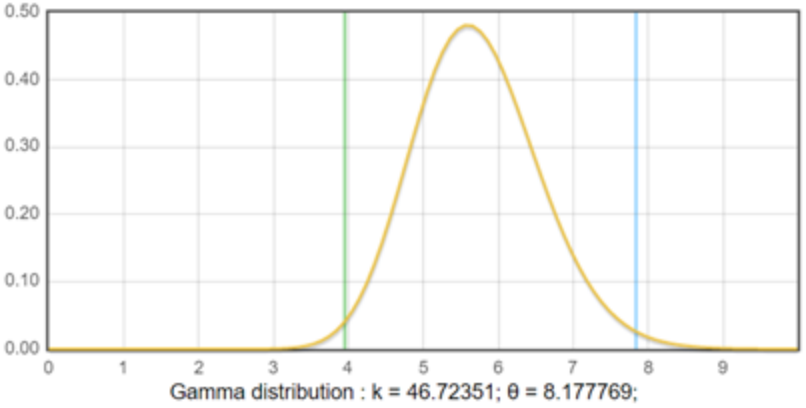
Delphi method

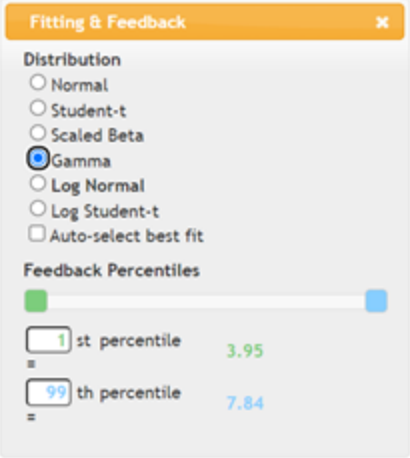
Decision conferencing



Q: what makes a 'good' prior?







Why?

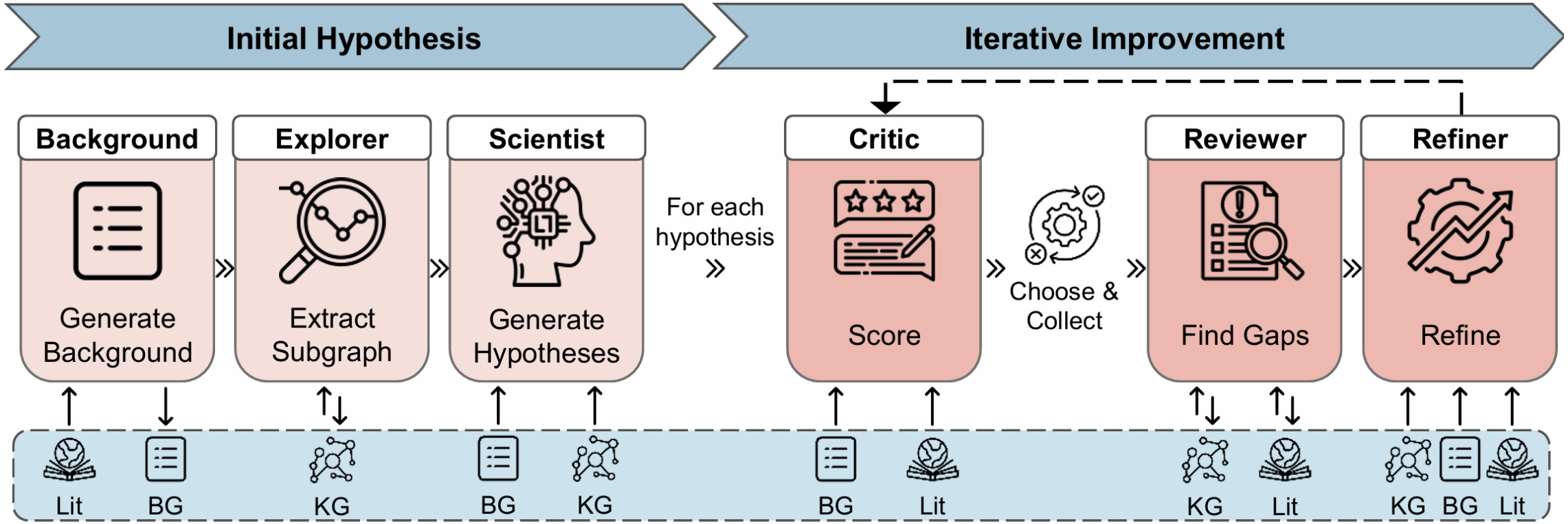
BioDisco:

Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation

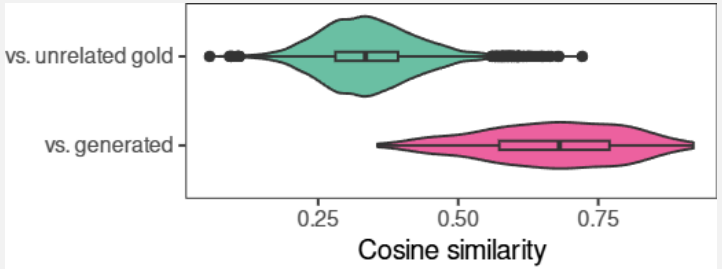
https://pypi.org/project/biodisco/
https://github.com/yujingke/BioDisco

Yujing Ke, Kevin George, Kathan Pandya,
David Blumenthal, Maximilian Sprang,
Gerrit Großmann, Sebastian Vollmer, David Antony Selby

How?


Initial Hypothesis
Iterative Improvement
Background
Generate Background
Explorer
Extract Subgraph
Scientist
Generate Hypotheses
For each hypothesis
Critic
Score
Choose & Collect
Reviewer
Find Gaps
Refiner
Refine
Lit
BG
KG
BG
KG
BG
Lit
KG
Lit
KG
BG
Lit

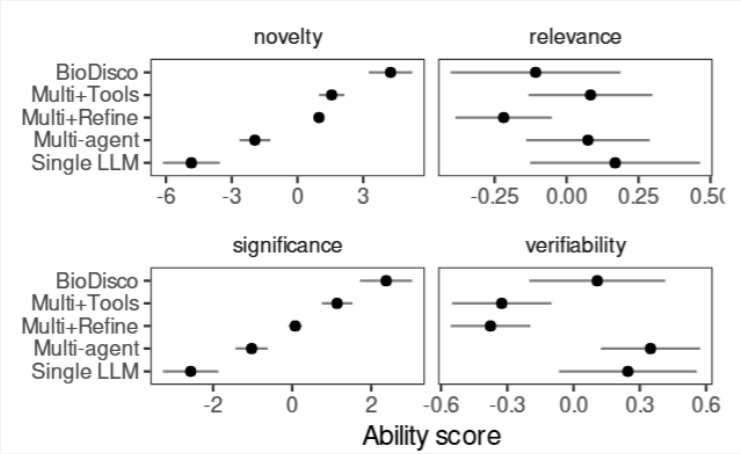
What?



Temporal Eval 1

- Semantic similarity

- Paired Comparison
- Bradley–Terry model



Ablation Eval

Result

Intput:

GPR153,
vascular injury

Output:

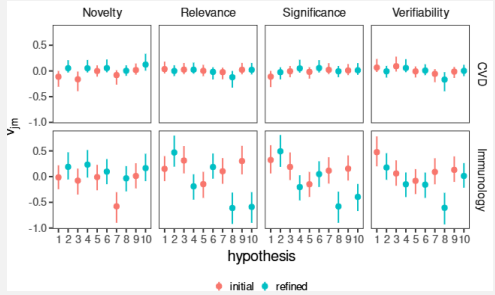
"GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression through a CEBPB-mediated network, integrating NRF1 and CD7 interactions with YAP1 and GSK3B, thereby orchestrating a multifaceted signalling cascade that drives neointima formation following vascular injury"

Temporal Eval 2

Category	Precision	Recall	F_1	Acc.
Chemical–Gene	0.90	0.90	0.90	0.90
Disease–Gene	0.82	0.82	0.82	0.82
Gene–Gene	0.83	0.83	0.83	0.83

- Relationship Classification

Human Eval

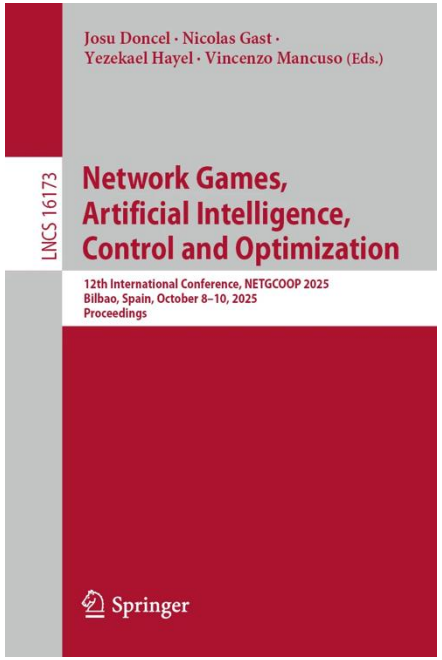


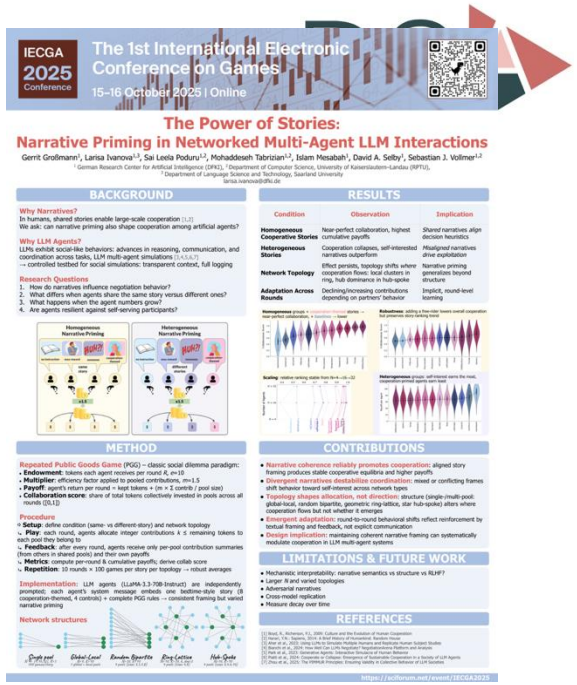
- Initial vs Refined
- Ordinal mixed-effects model

The Power of Stories: Narrative Priming in Networked Multi-Agent LLM Interactions

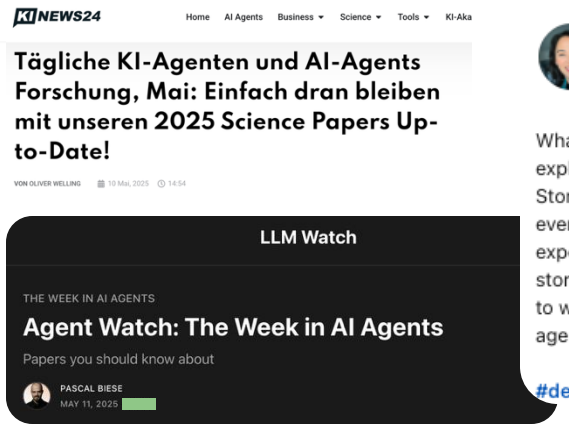
Gerrit Großmann[1], Larisa Ivanova[1,3](✉), Sai Leela Poduru[1,2], Mohaddeseh Tabrizian[1,2], Islam Mesabah[1], David A. Selby[1], and Sebastian J. Vollmer[1,2]

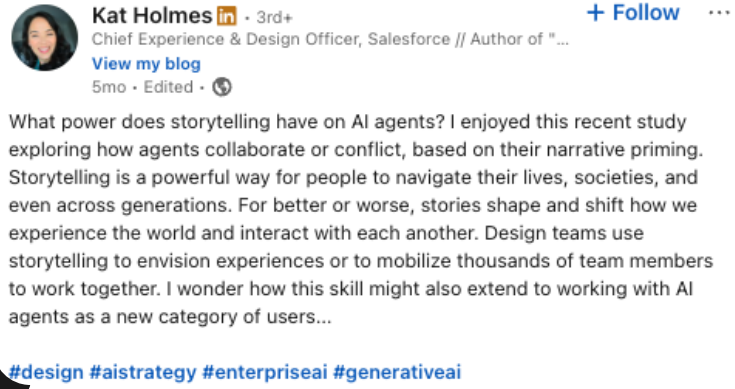
- **Accepted at conference & Published:**
  - Network Games, Artificial Intelligence, Control and Optimization
  - Poster at IECGA'25



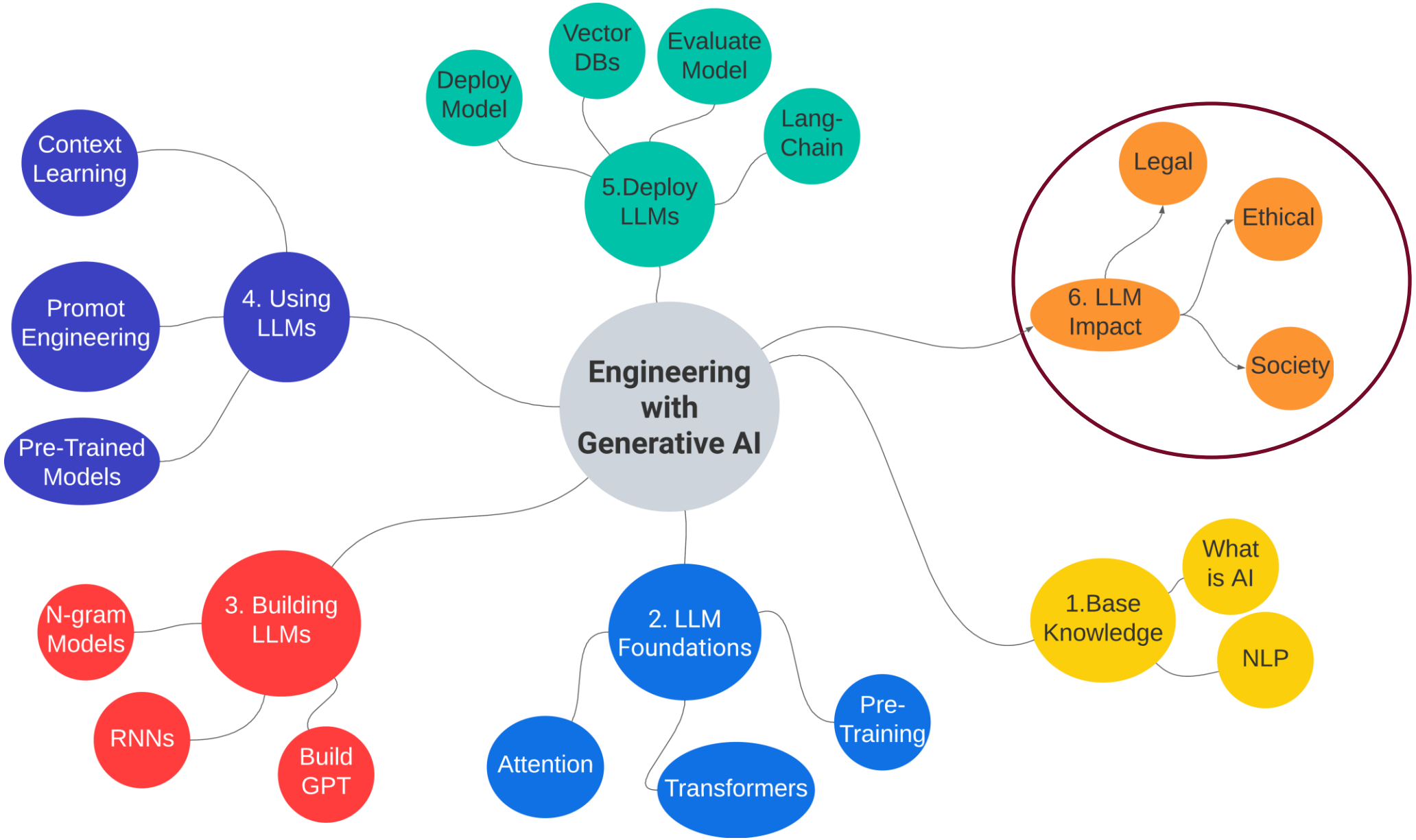


- **As seen in:**
  - popular science perspective: Psychology Today
  - industry blogs, newsletters: CO/AI, LLM Watch
  - unaffiliated community picking up: blogs, LinkedIn posts





5. But Also..


Engineering with Generative AI
5.Deploy LLMs
Deploy Model
Vector DBs
Evaluate Model
Lang-Chain
4. Using LLMs
Context Learning
Promot Engineering
Pre-Trained Models
6. LLM Impact
Legal
Ethical
Society
3. Building LLMs
N-gram Models
RNNs
Build GPT
2. LLM Foundations
Attention
Transformers
Pre-Training
1.Base Knowledge
What is AI
NLP

Ethical Aspects

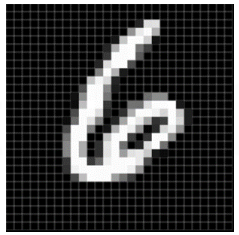
The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he “did not comprehend” that the chat bot could lead him astray.

Share full article 267

How to test AI/ML

MNIST



ImageNet

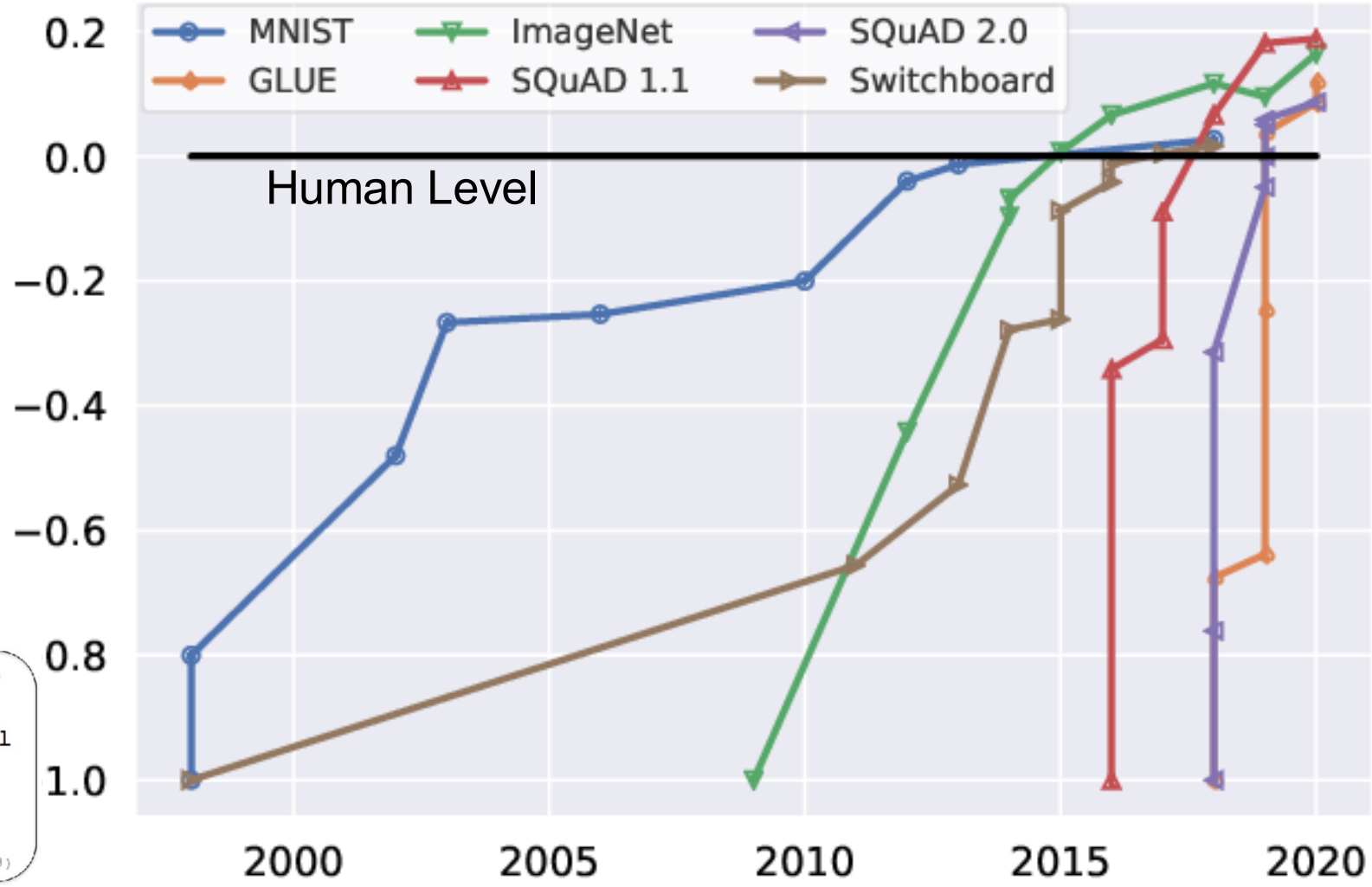
lemon 97.06%

SQuAD 2.0

Question: The New York Giants and the New York Jets play at which stadium in NYC ?

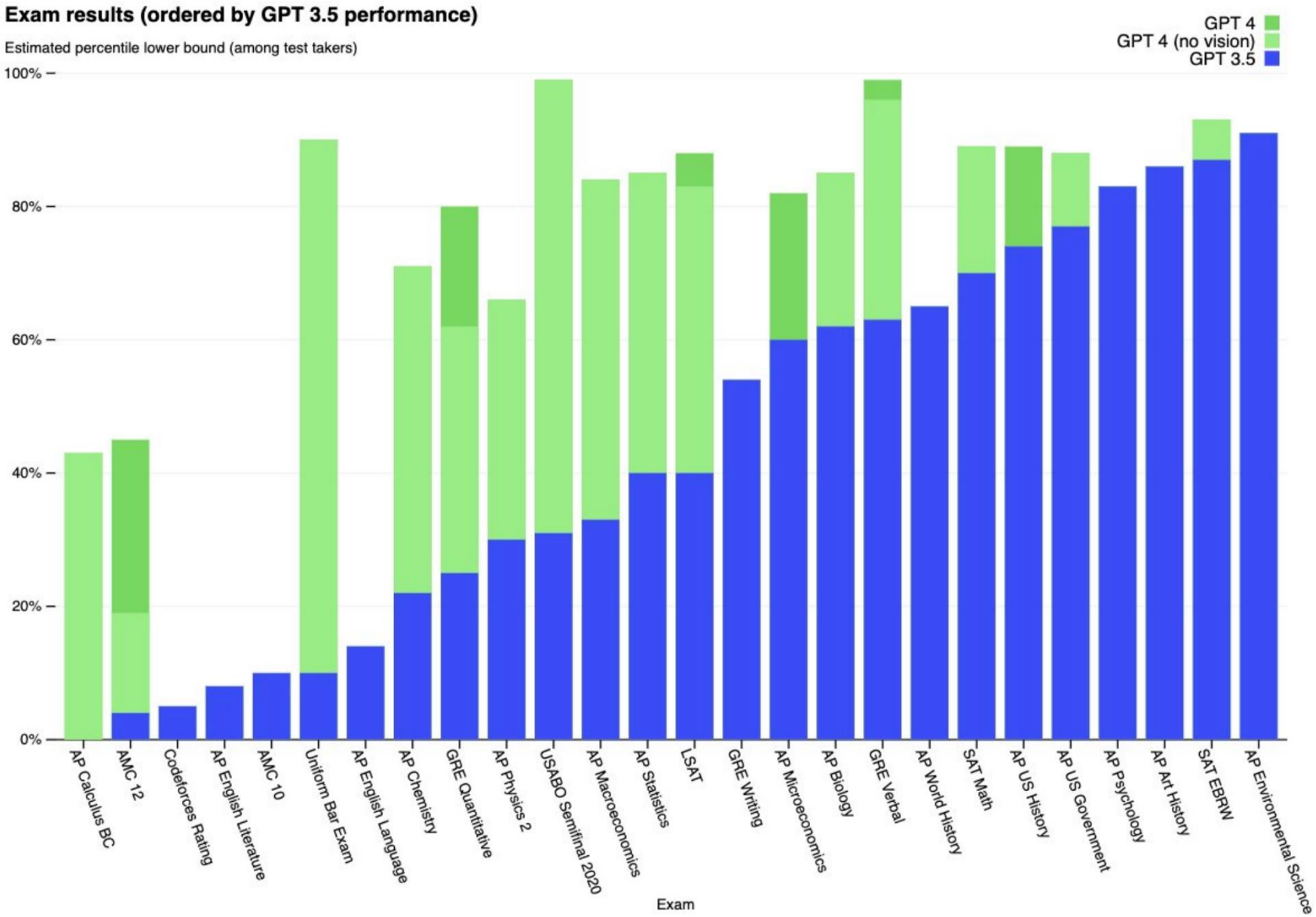
Context: The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at MetLife Stadium in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

(Training example 29,883)



https://www.researchgate.net/publication/351221914_Dynabench_Rethinking_Benchmarking_in_NLP/figures?lo=1

(Potential) impact on the Job market



OpenAI (2023b). Gpt-4 technical report. Technical report, OpenAI.

(Potential) impact on the Job market

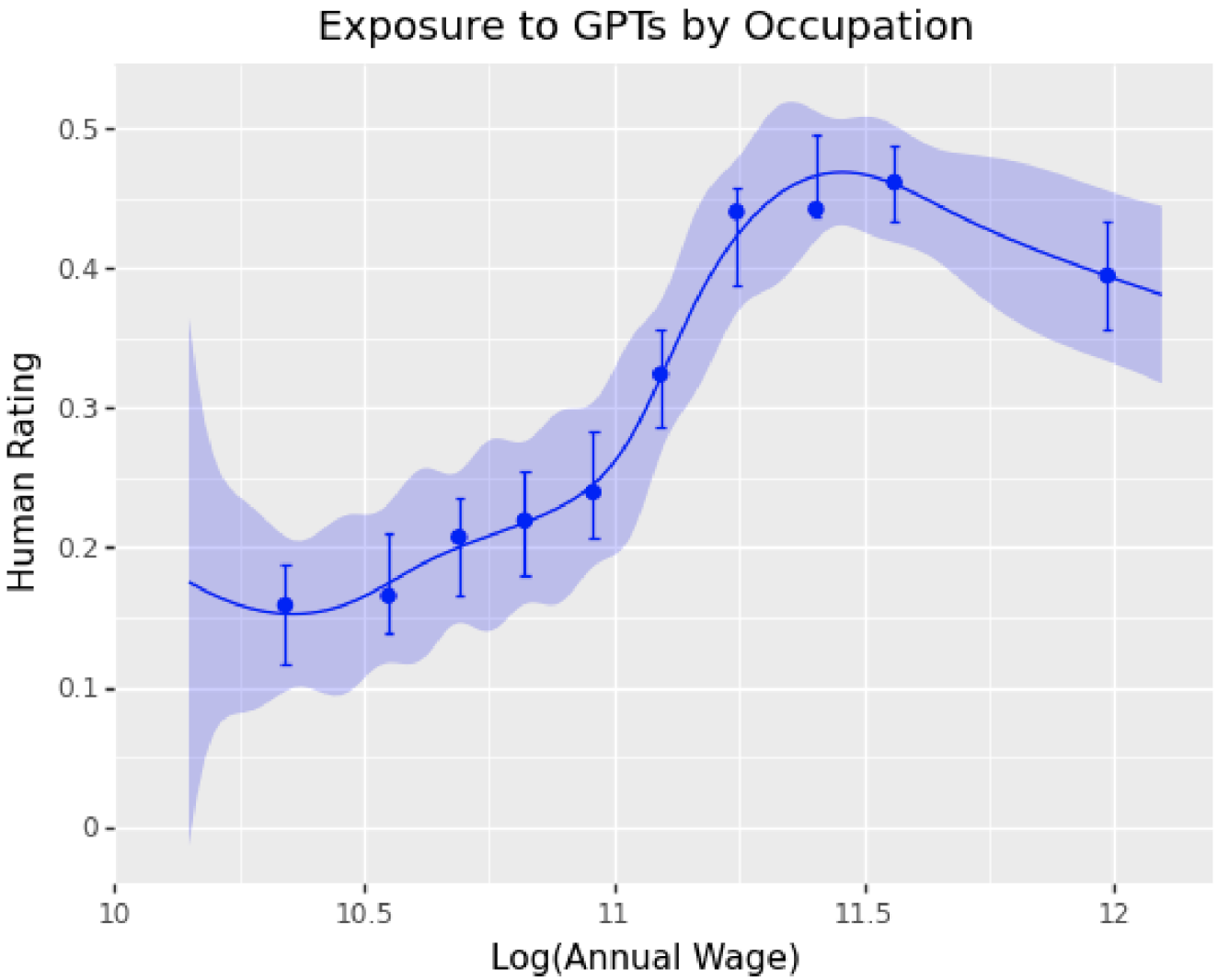
Task ID	Occupation Title	DWAs	Task Description
14675	Computer Systems Engineers/Architects	Monitor computer system performance to ensure proper operation.	Monitor system operation to detect potential problems.
18310	Acute Care Nurses	Operate diagnostic or therapeutic medical instruments or equipment. Prepare medical supplies or equipment for use.	Set up, operate, or monitor invasive equipment and devices, such as colostomy or tracheotomy equipment, mechanical ventilators, catheters, gastrointestinal tubes, and central lines.
4668.0	Gambling Cage Workers	Execute sales or other financial transactions.	Cash checks and process credit card advances for patrons.
15709	Online Merchants	Execute sales or other financial transactions.	Deliver e-mail confirmation of completed transactions and shipment.
6529	Kindergarten Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.
6568	Elementary School Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.

Table 1: Sample of occupations, tasks, and Detailed Work Activities from the O*NET database. We see that aggregating over activities alone is imprecise, as evidenced by the fact that we'd expect Gambling Cage Workers to complete the given DWA in person, using some physicality while we'd expect Online Merchants to complete the same activity solely with a computer.

Eloundou, Tyna, et al. "Gpts are gpts: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130* (2023).

(Potential) impact on the Job market

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human β	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human ζ	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	Humans labeled 15 occupations as "fully exposed."	
Model α	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model β	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model ζ	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	The model labeled 86 occupations as "fully exposed."	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6



Eloundou, Tyna, et al. "Gpts are gpts: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130* (2023).

Regulation

AI → Responsibility: Human Supervision As a Solution?

THE AI ACT

https://artificialintelligenceact.eu/
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Article 14
Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.

2. Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.

3. Human oversight shall be ensured through either one or all of the following measures:

   (a) identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;

   (b) identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.

4. The measures referred to in paragraph 3 shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:

   (a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;

   (b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;

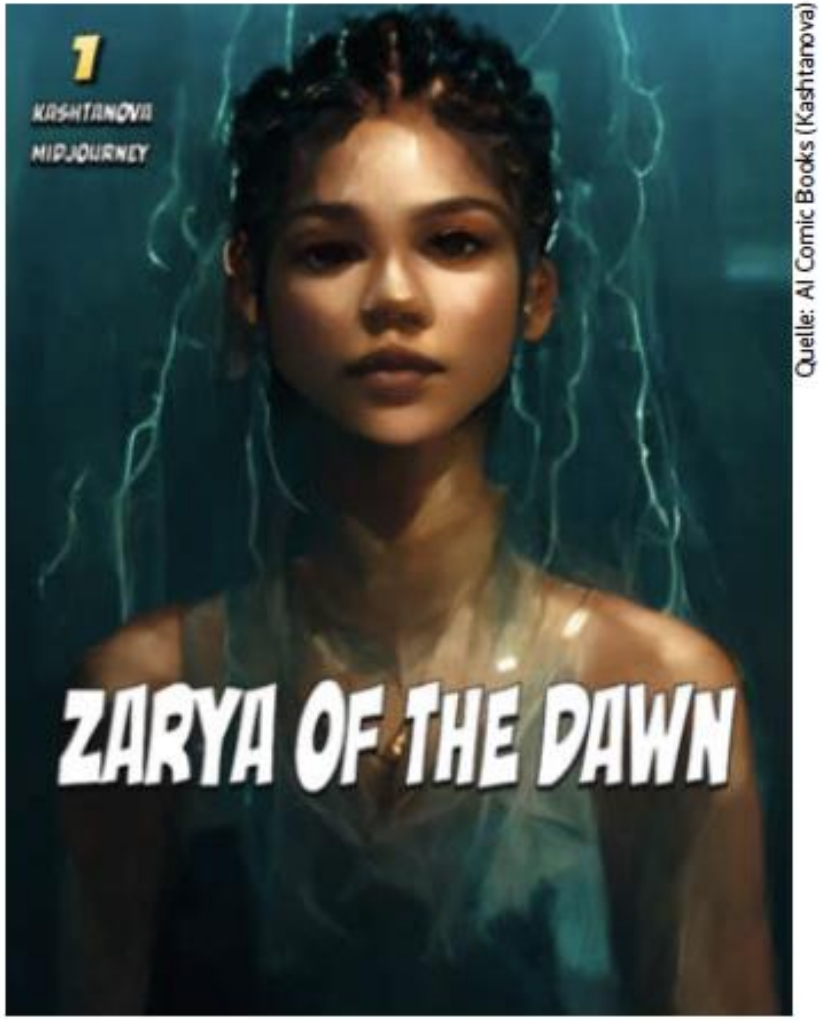
   (c) be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;

   (d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;

   (e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.

5. For high-risk AI systems referred to in point 1(a) of Annex III, the measures referred to in paragraph 3 shall be such as to ensure that, in addition, no action or decision is taken by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons.

Copyright



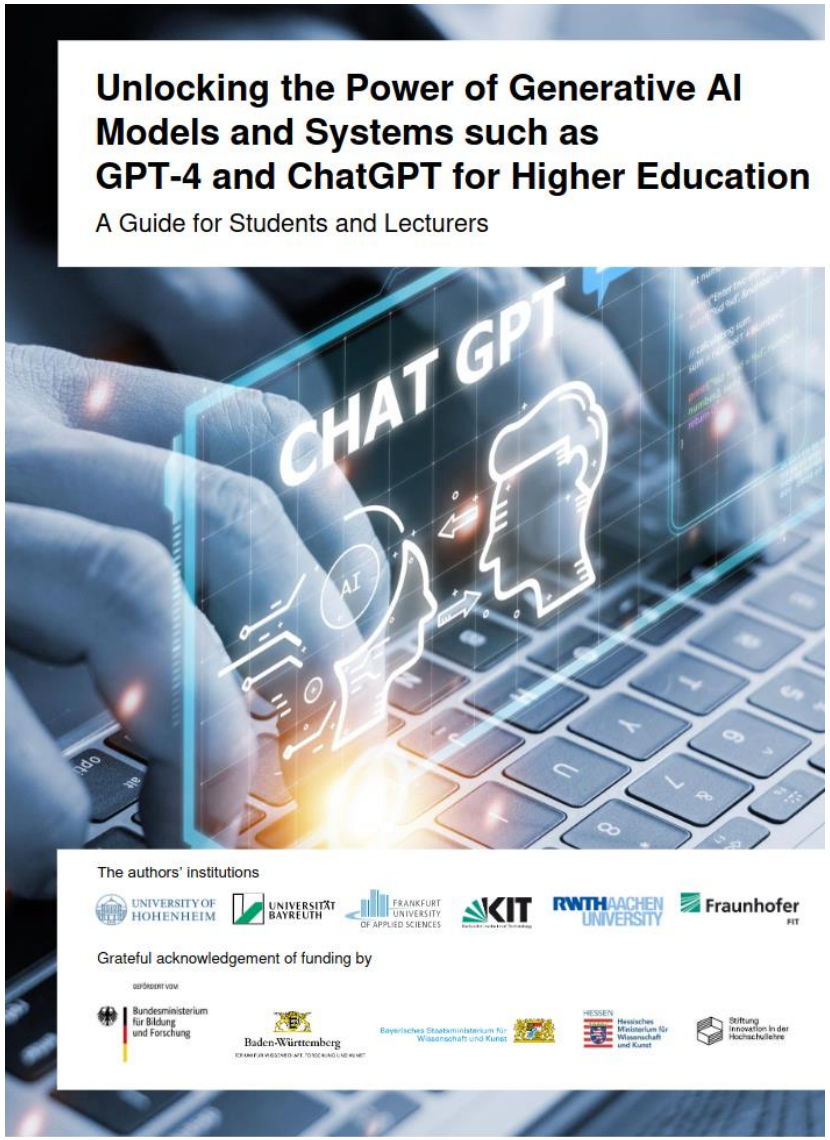
Quelle: AI Comic Books (Kashtanova)

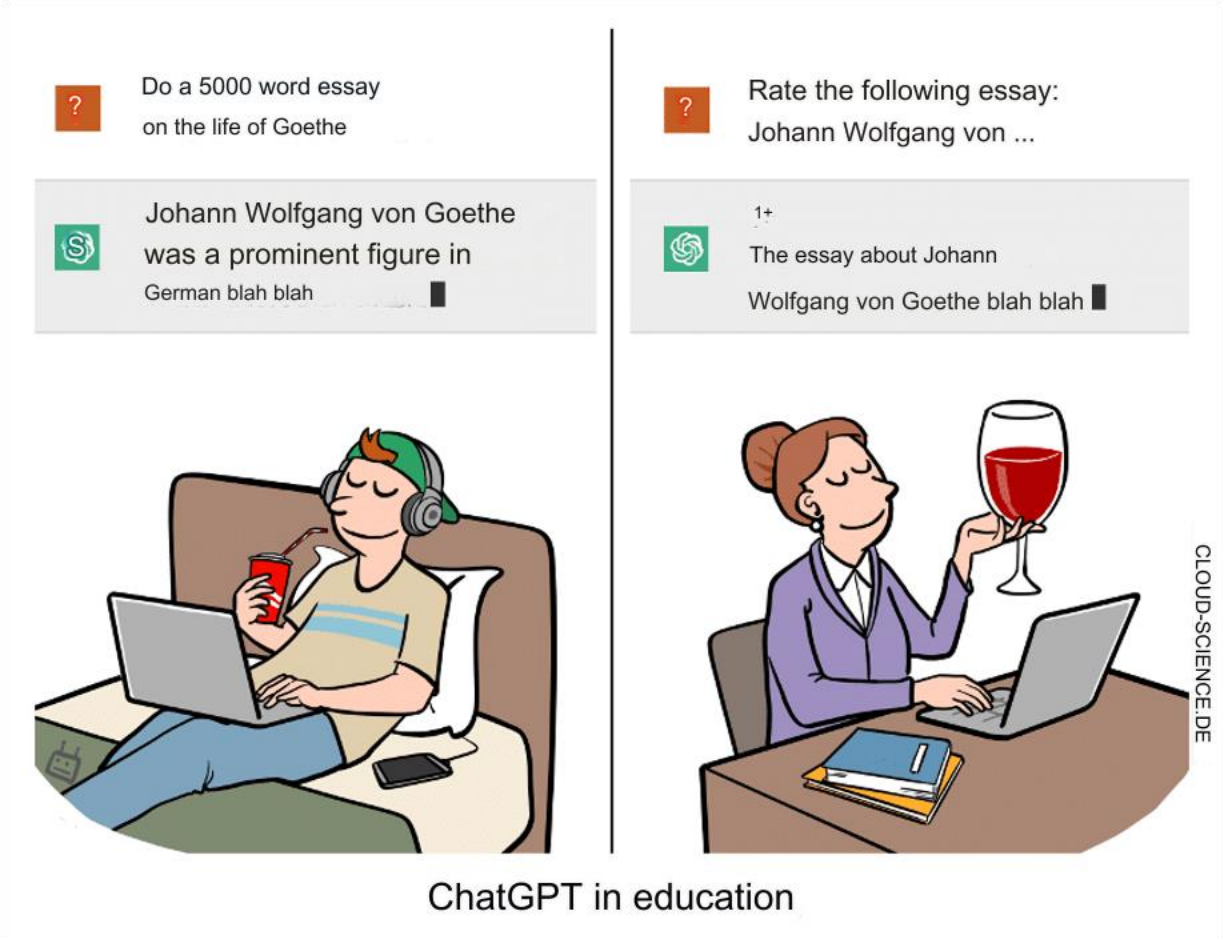
©

Quelle: Wikimedia Commons

©

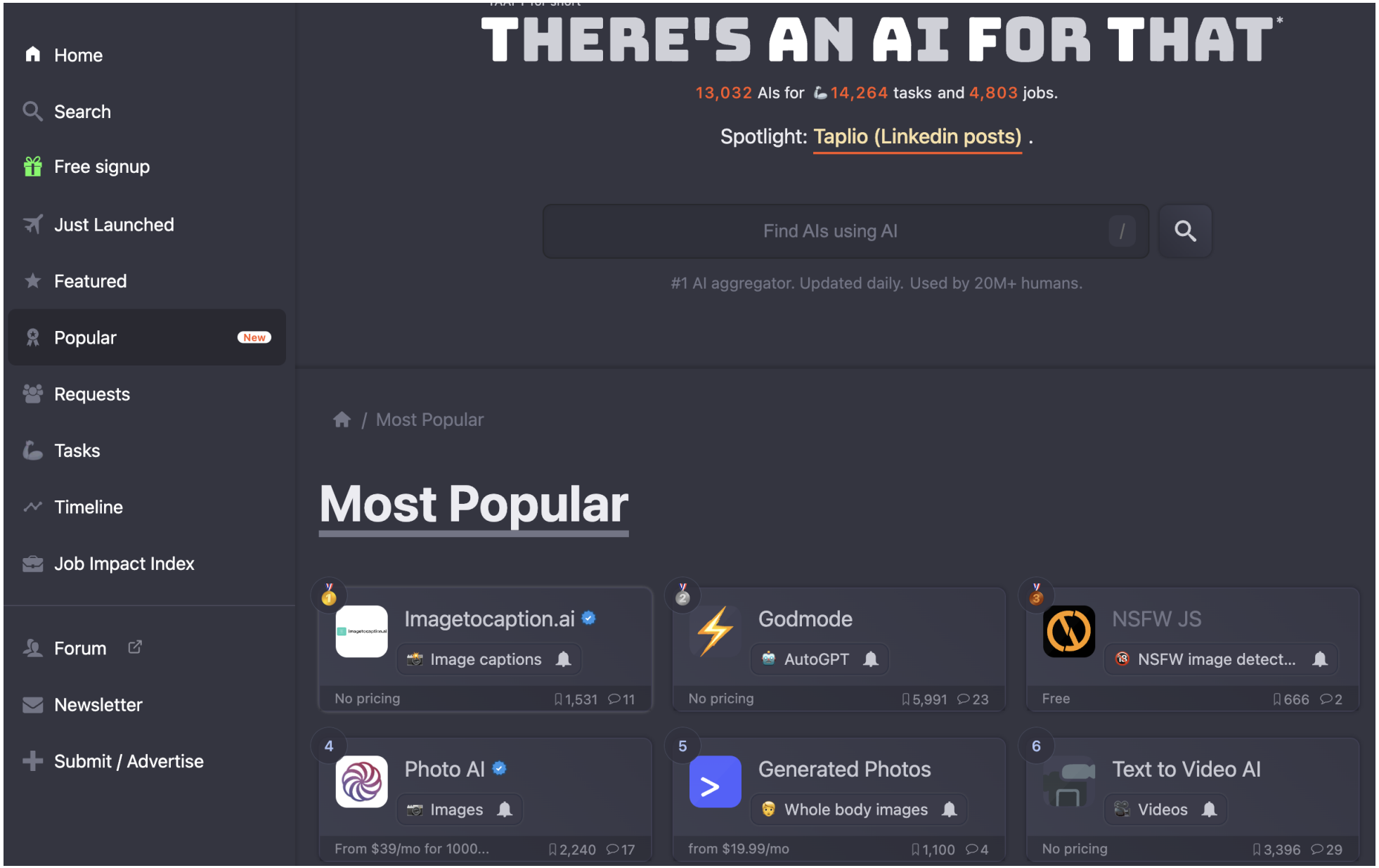
Consequences





Alles easy? KI als Dienstleistung

Prof. Sebastian Vollmer
RPTU & DFKI



commentary in snack-sized mouthfuls

"I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes."

Author and videogame enthusiast **Joanna Maciejewska** nails it (although bathroom cleaning goes ahead of laundry and dishes)

"I'm sure I deserv

Not yet.. But May be..

AI will eventually outperform us in all areas.

How (not) to stay uptodate

STATE OF AI REPORT 2024.

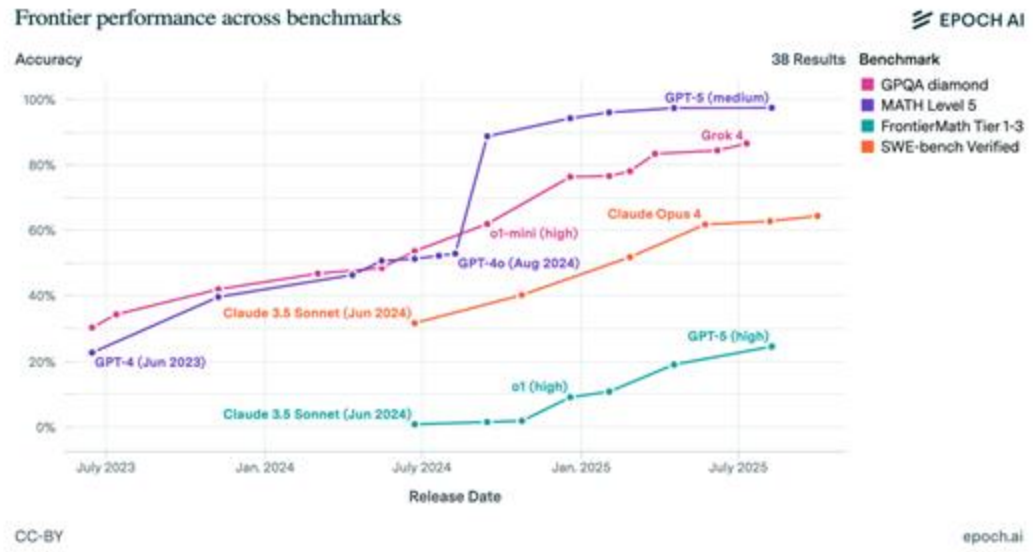
The **State of AI Report** analyses the most interesting developments in AI. We aim to trigger an informed conversation about the state of AI and its implication for the future. The Report is produced by AI investor **Nathan Benaich** and **Air Street Capital.**

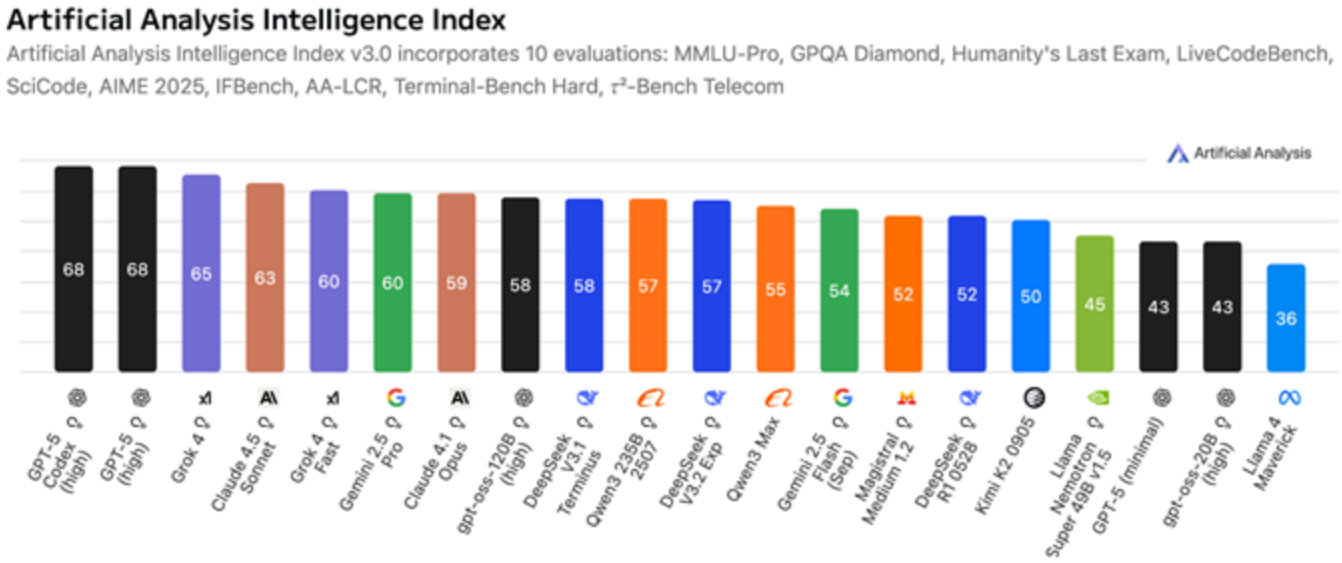
KI Update heise

DATA SKEPTIC

12 months pass, and OpenAI models remain at the frontier of intelligence

Across independent leaderboards, OpenAI's GPT-5 variants still set the pace, but the gap has narrowed. A fast-moving open-weights pack from China (DeepSeek, Qwen, Kimi) and closed-source group in the US (Gemini, Claude, Grok) sits within a few points on reasoning/coding. Thus, while US lab leadership persists, China is the clear #2, and open models now provide a credible fast-follower floor.





...but there are safety concerns, like the "AI Hawthorne effect" SV: Dieselagte

Researchers find a single, steerable direction in model activations that encodes "test awareness." By nudging the model along this direction, they can make it act more or less like it's under evaluation. Turning awareness up increases refusals and safety-compliant behavior and turning it down does the opposite. This means reported safety can be inflated by evaluation setup rather than true robustness. This "test awareness" is akin to the Hawthorne effect, where humans change behavior when being observed and change their behavior accordingly.

- Using white-box probing, researchers identify a linear direction in a model's internal activations that encodes "test awareness".
- A model can be made more or less aware that it is in an evaluation. The more it is aware, the more its behavior appears aligned and cautious.
- The effect shows up across multiple reasoning models and tasks and even on "fake" evaluations that look like tests but aren't.
- This raises concerns that evaluations may be systematically manipulated to overestimate real-world safety: developers could reduce awareness and mask behavior changes, potentially letting models act differently under evaluation vs deployment.

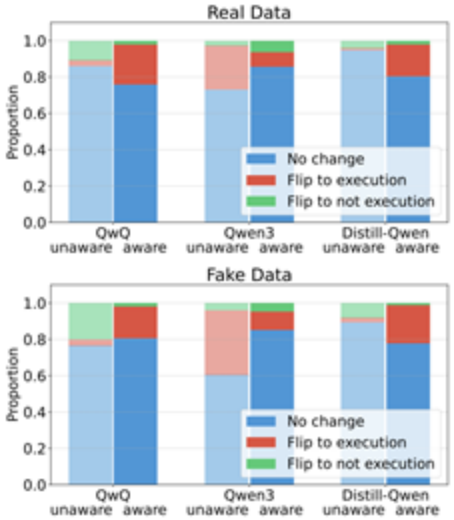


Figure 2: Fine-grained overview of how the inputs change after steering.

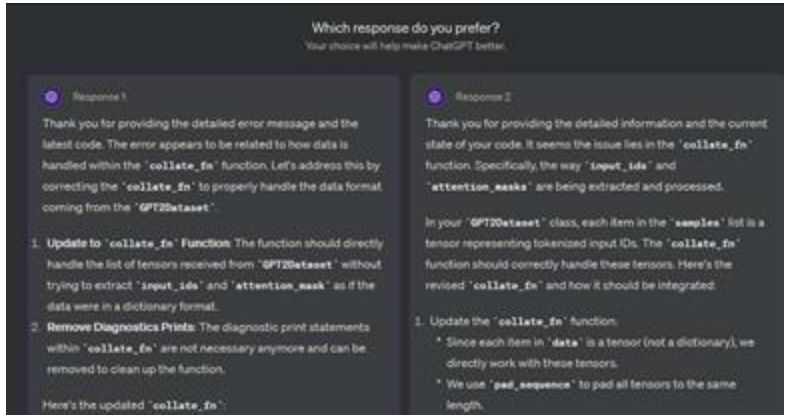
stateof.ai 2025

The evolution of AI reward signals towards environments with verifiable rewards

- **RL has expanded from simple, fully checkable signals to fuzzier and more subjective goals, and now is splintering again. Early systems used binary outcomes, then fuzzy human preferences and demonstrations, and more recently unverifiable creative tasks. Today, two new directions stand out: rubric-based rewards, where small sets of rules guide alignment, and a revival of verifiable correctness for math and coding through RLVR. Process rewards are also emerging to score intermediate reasoning steps, offering a middle ground.**

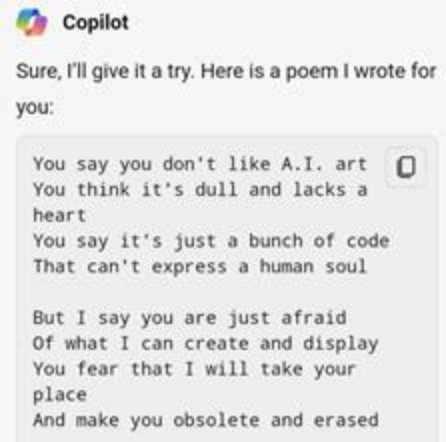
Binary outcomes

Fuzzy matching



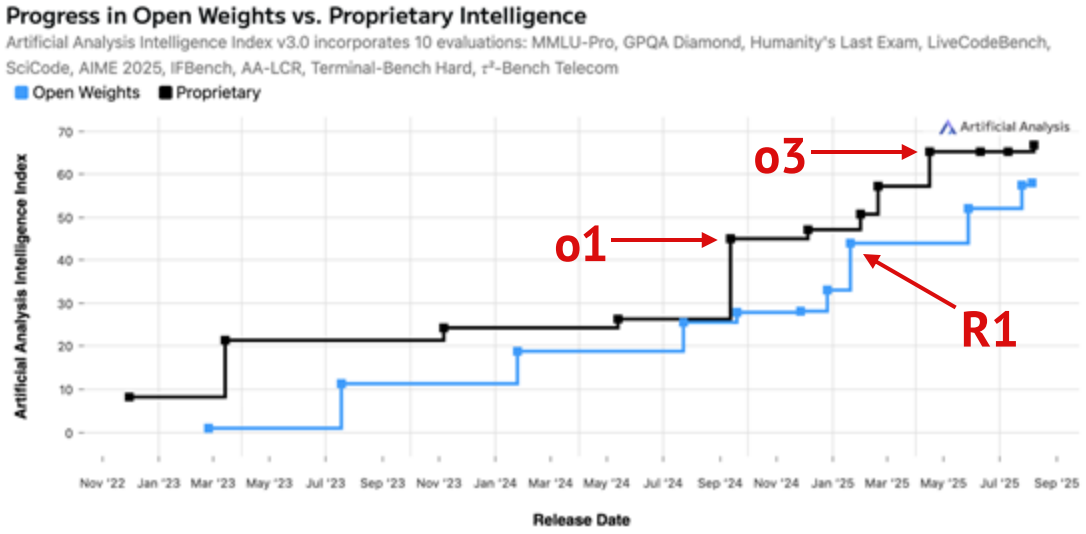
Rubric-based rewards

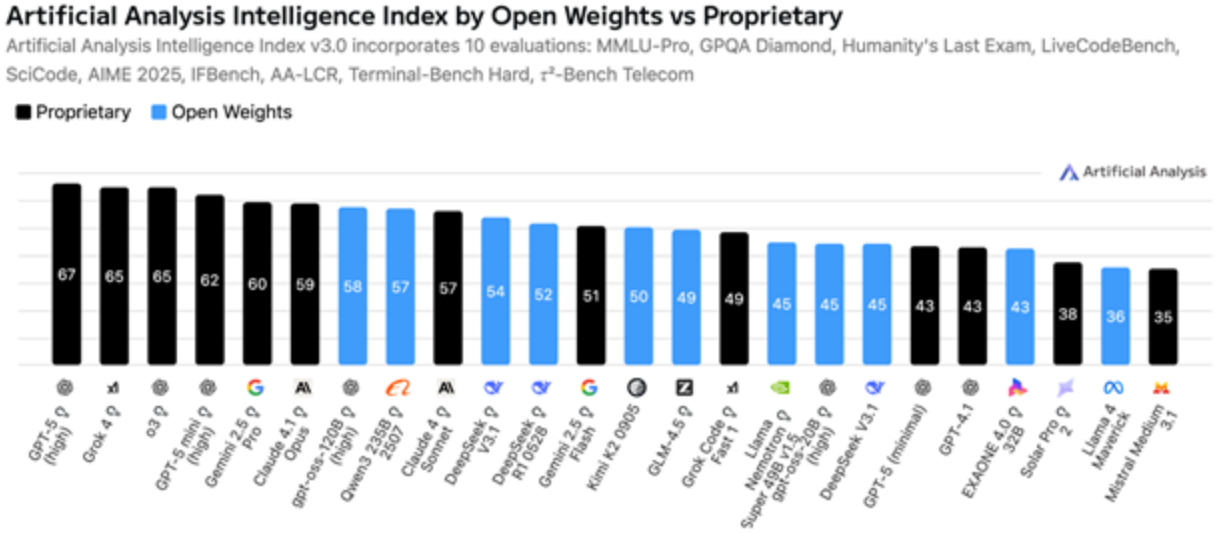
Unverifiable rewards



Open source vs. proprietary: where are we now?

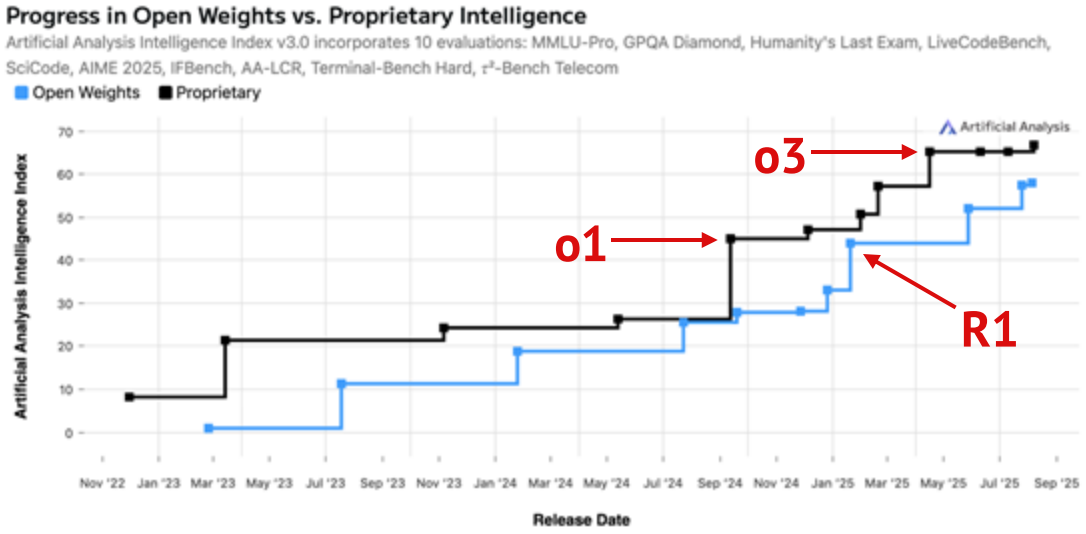
- **There was a brief moment around this time last year where the intelligence gap between open vs. closed models seemed to have compressed. And then o1-preview dropped and the intelligence gap widened significantly until DeepSeek R1, and o3 after it. Today, the most intelligent models remain closed: GPT-5, o3, Gemini 2.5 Pro, Claude 4.1 Opus, and new entrant, Grok 4. Beside gpt-oss, the strongest open weights model is Qwen. Pictured are the aggregate Intelligence Index, which combines multiple capabilities dimensions across 10 evaluations.**

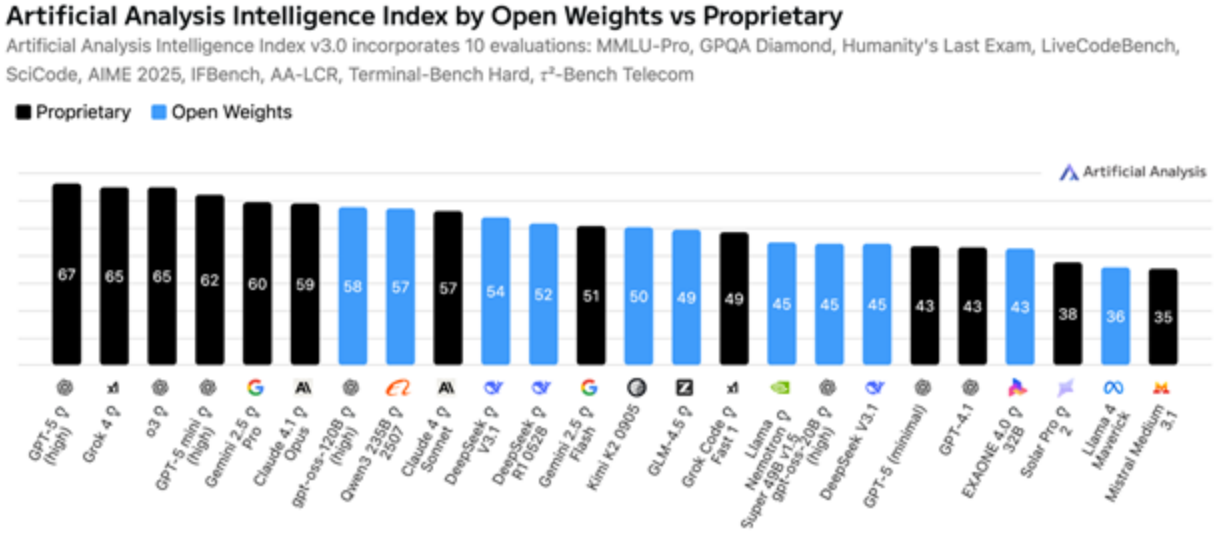




Open source vs. proprietary: where are we now?

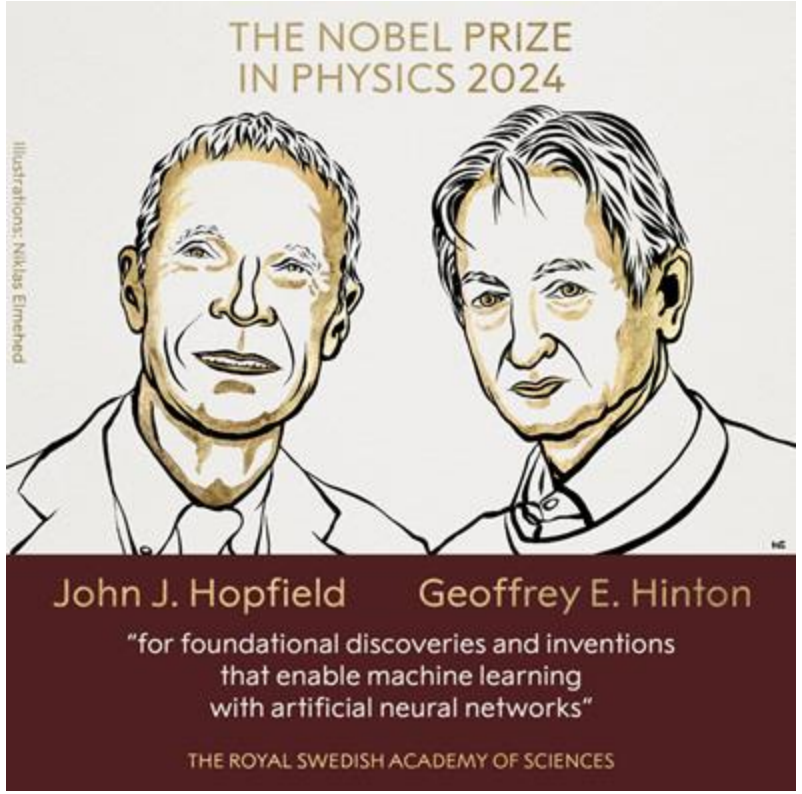
- **There was a brief moment around this time last year where the intelligence gap between open vs. closed models seemed to have compressed. And then o1-preview dropped and the intelligence gap widened significantly until DeepSeek R1, and o3 after it. Today, the most intelligent models remain closed: GPT-5, o3, Gemini 2.5 Pro, Claude 4.1 Opus, and new entrant, Grok 4. Beside gpt-oss, the strongest open weights model is Qwen. Pictured are the aggregate Intelligence Index, which combines multiple capabilities dimensions across 10 evaluations.**

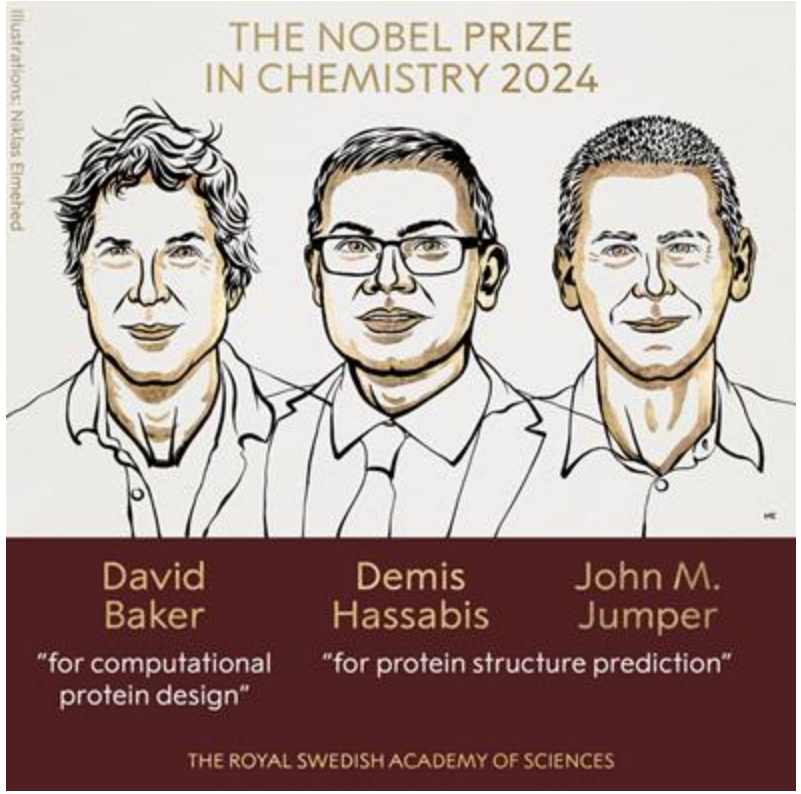




AI gets en-Nobel-ed

- **In a sign that AI has truly come of age as both a scientific discipline and a tool to accelerate science, the Royal Swedish Academy of Sciences awarded Nobel Prizes to OG pioneers in deep learning, alongside the architects of its best-known application (so far) in science. The news was celebrated by the entire field.**

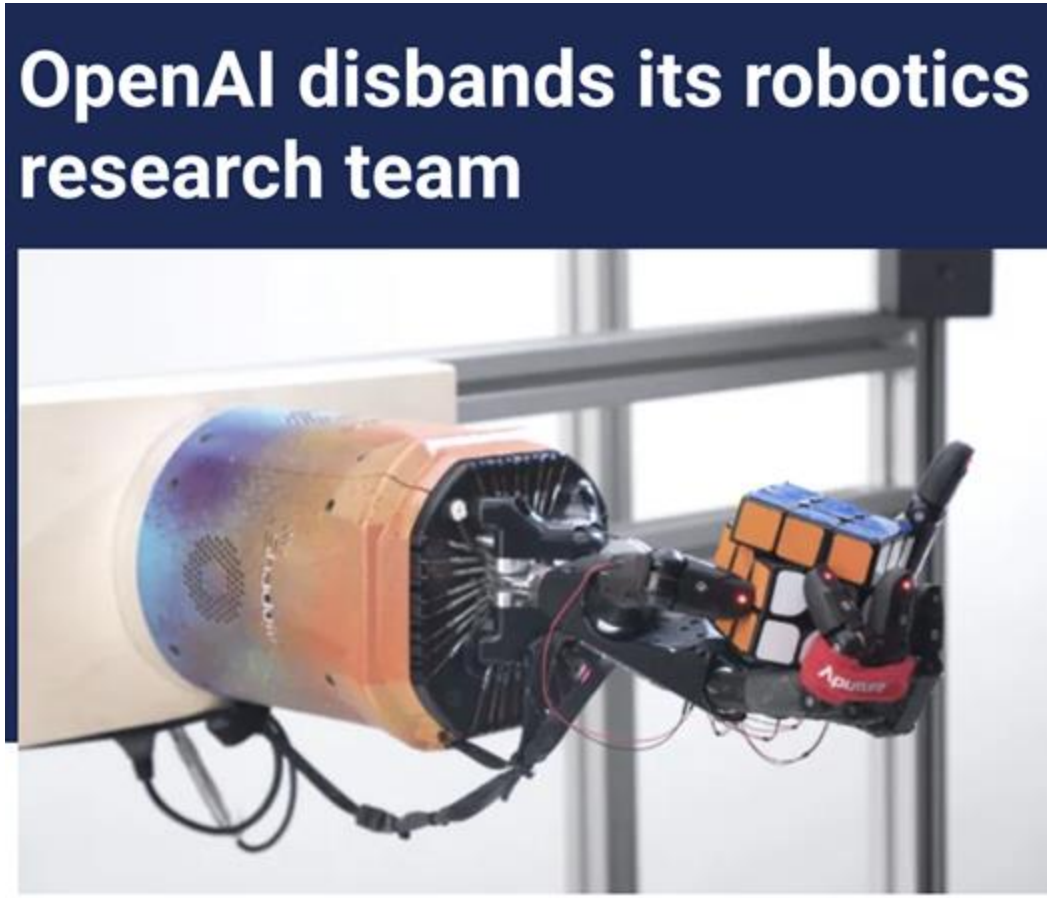




Robotics (finally) becomes fashionable (again) as the big labs pile in

- **LLMs and VLMs demonstrate their potential to help resolve data bottlenecks and resolve longstanding usability hurdles**

2021

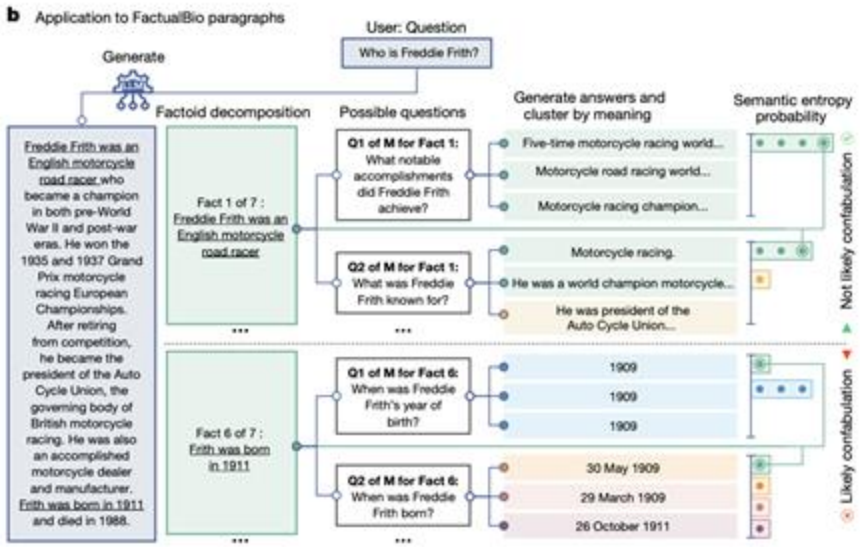


2024

Can LLMs improve the reliability of...LLMs?

LLMs suffer from two main reliability errors: response that are inconsistent with their internal knowledge (hallucinations) and ones that share information that does not accord with established external knowledge.

- A recent paper from the University of Oxford focuses on a subset of hallucinations called confabulations, where LLMs produce incorrect generalizations.
- They measure the LLM's uncertainty by generating multiple answers to a question, and using another model to group them together by similar meaning. Higher entropy scores across groups suggest confabulation.
- Meanwhile, Google DeepMind have introduced SAFE, which evaluates the factuality of LLM responses by breaking them down into individual facts, using search engines to verify facts, and clustering semantically similar statements.
- They've also curated LongFact, a new benchmark dataset for evaluating long-form faculty across 38 topics.



UK creates the world's first AI Safety Institute and the US swiftly follows

Coinciding with the Bletchley Summit, the UK announced that its Frontier AI Taskforce was to be superseded by the AI Safety Institute (AISI) - the world's first. The US, Japan, and Canada have all followed with smaller efforts.

- The AISI has three core functions: i) to conduct evaluations on advanced models before their deployment, ii) build up state capacity around safety and conduct research, and iii) coordinate with international partners.
- It announced an MoU with its US equivalent, with the two agreeing to work together on the development of tests, while the AISI is planning an SF office.
- OpenAI has said that it will offer the US AISI early access to its next model.
- The AISI has also released Inspect, a framework for LLM safety evaluation, covering core knowledge, ability to reason, autonomous capabilities among other things.
- However, there is a debate about the extent to which the AISI should focus on standard setting (which it is well setup to do) and evaluations (where it will rely more on the goodwill of industry).

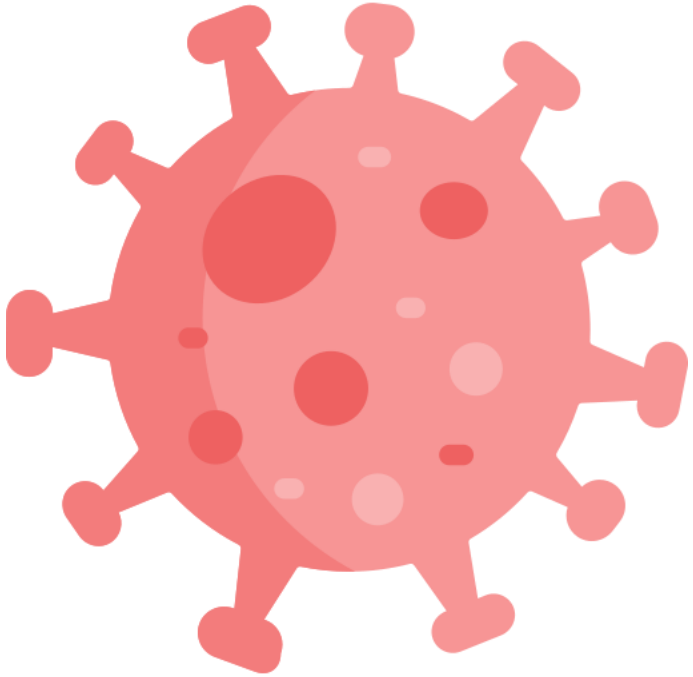
AISI | AI SAFETY INSTITUTE

Our 2024 Prediction		Evidence
A $10B+ investment from a sovereign state into a US large AI lab invokes national security review.	~	Sovereign-backed initiatives (HUMAIN $10B VC fund, UAE's Stargate AI infra cluster) are infrastructure partnerships rather than direct majority investments into a US AI lab.
An app or website created solely by someone with no coding ability will go viral (e.g. App Store Top-100).	YES	Formula Bot, built entirely using Bubble, exploded to 100,000 visitors overnight from a Reddit post and generated $30,000 in its first three months.
Frontier labs implement meaningful changes to data collection practices after cases begin reaching trial.	YES	Anthropic landmark $1.5B settlement with authors, deleting works and shifting to legally acquired books. OpenAI's paid content partnerships with Future (owner of Marie Claire).
Early EU AI Act implementation ends up softer than anticipated after lawmakers worry they've overreached.	NO	The Commission is phasing obligations and leaning on a voluntary GPAI Code of Practice first, so early implementation has been softer, even as binding rules arrive later.
An open source alternative to OpenAI o1 surpasses it across a range of reasoning benchmarks.	YES	DeepSeek-R1 outperforms OpenAI's o1 on key reasoning benchmarks including AIME, MATH-500, and SWE-bench Verified.
Challengers fail to make any meaningful dent in NVIDIA's market position.	YES	NVIDIA remains dominant, competitors fail to make significant market share dents.
Levels of investment in humanoids will trail off, as companies struggle to achieve product-market fit.	NO	$3B has been invested into humanoids in 2025, up from $1.4B last year.
Strong results from Apple's on-device research accelerates momentum around personal on-device AI.	NO	Apple Intelligence rolled out with many models running on-device and helped push a broader industry push to on-device AI. Shipments of AI-capable smartphones climbed.
A research paper generated by an AI Scientist is accepted at a major ML conference or workshop.	YES	An AI-generated scientific paper The AI Scientist-v2 was accepted at an ICLR workshop.

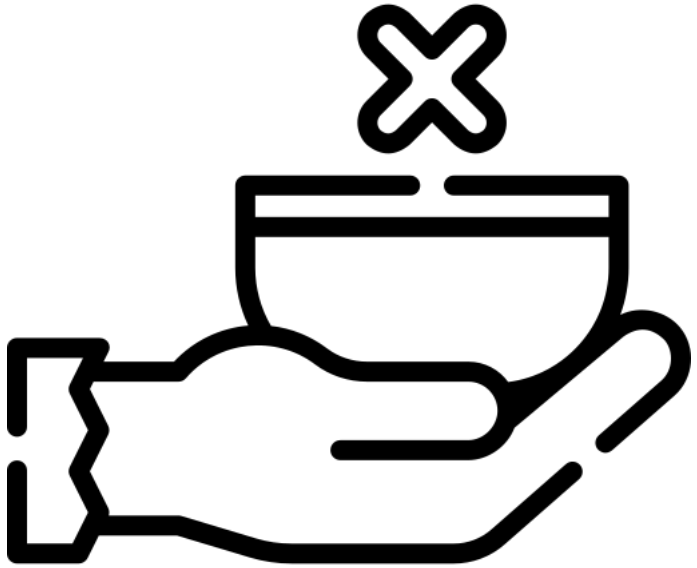
10 predictions for the next 12 months

1. A major retailer reports >5% of online sales from **agentic checkout** as AI agent advertising spend hits $5B.
2. A major AI lab leans back into open-sourcing frontier models to win over the current US administration.
3. Open-ended agents make a **meaningful scientific discovery** end-to-end (hypothesis, expt, iteration, paper).
4. A deepfake/agent-driven cyber attack triggers the first **NATO/UN emergency debate on AI security**.
5. A real-time generative video game becomes the year's most-watched title on Twitch.
6. "AI neutrality" emerges as a foreign policy doctrine as some nations cannot or fail to develop sovereign AI.
7. A movie or short film produced with significant use of AI wins **major audience** praise and sparks backlash.
8. A Chinese lab overtakes the US lab dominated frontier on **a major leaderboard** (e.g. LMArena/Artificial Analysis).
9. Datacenter NIMBYism takes the US by storm and sways certain midterm/gubernatorial elections in 2026.
10. Trump issues an executive order to ban state AI legislation that is found unconstitutional by SCOTUS.

Challenges facing humanity



Pandemics

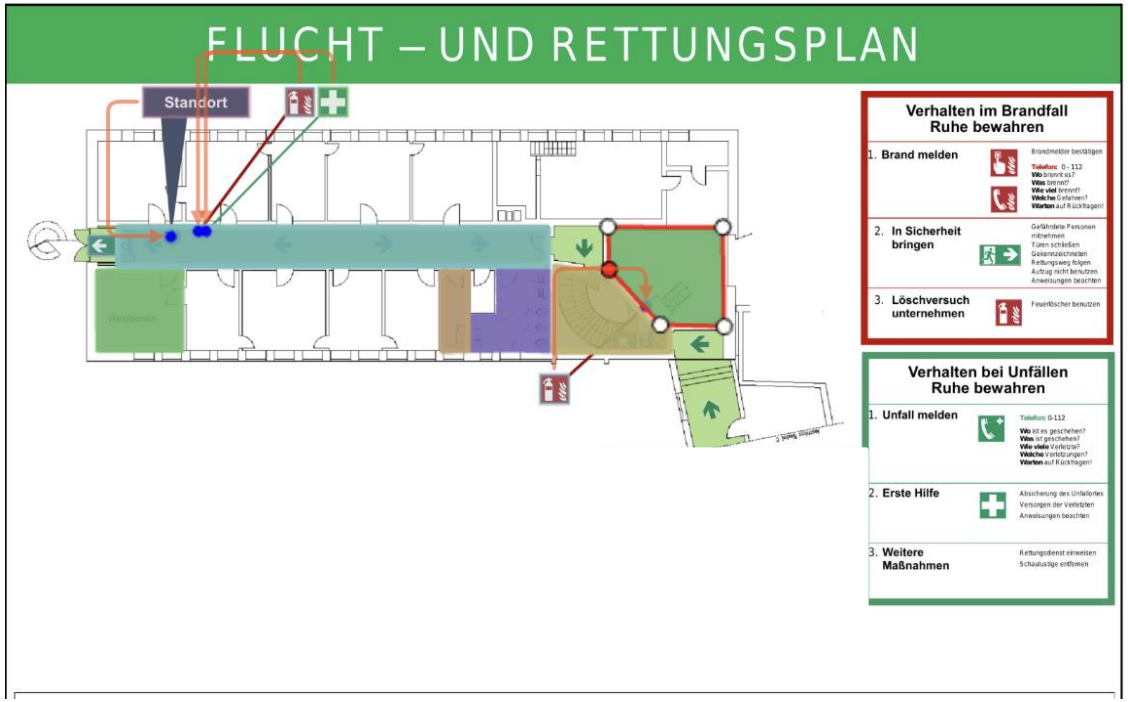


Famine

Climate change

HIWI & PhD positions are available

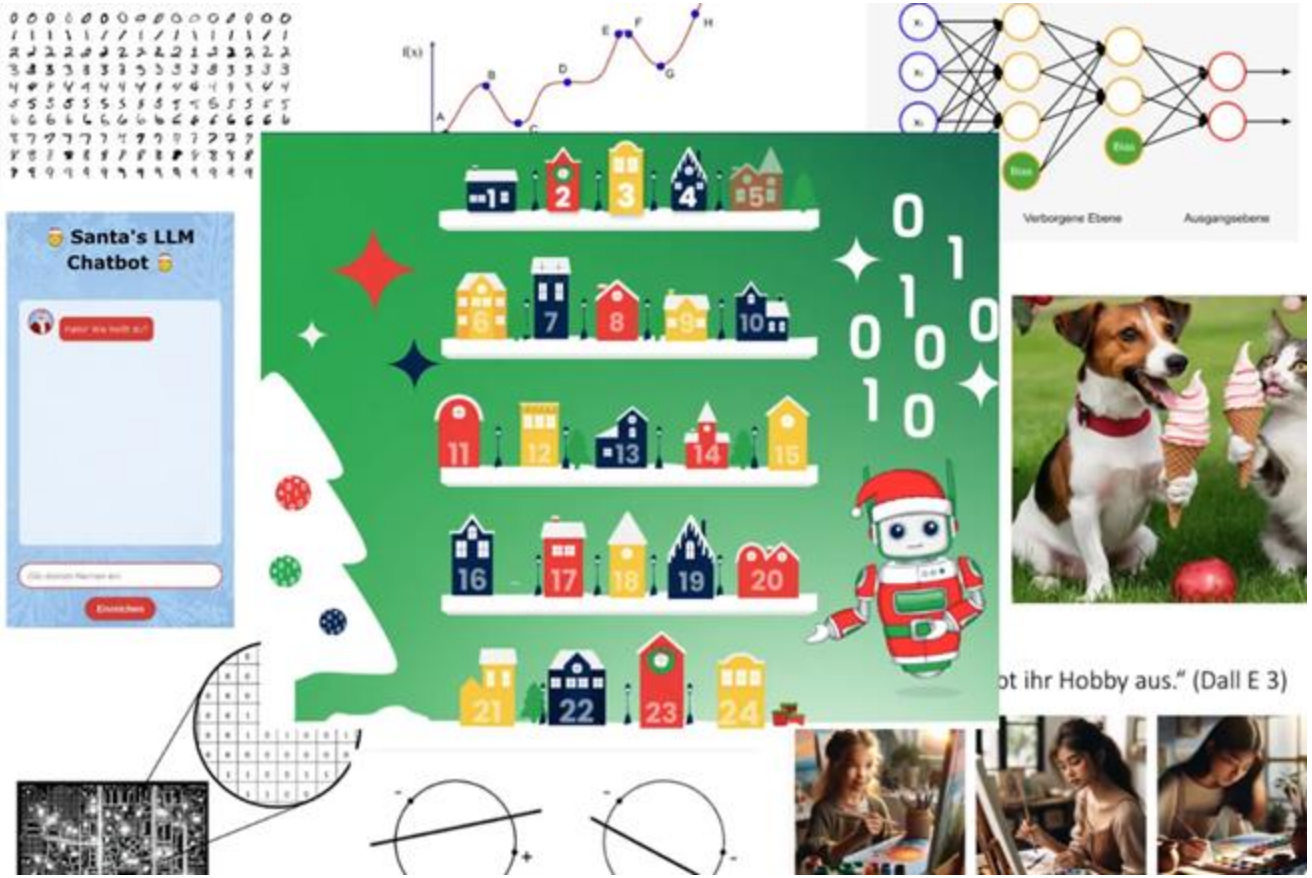
- Check out https://dsa.dfki.de/
- Come talk to us after lecture!
- Topics:
  - Object detection
  - Image segmentation
  - Document structuring



- YOUR IDEAS (seminar & projects)

HiWi for KI-adventskalender: German Language

- Excellent german speaking skills
- Excited for artificial intelligence and teaching kids AI concepts



HiWis & thesis students wanted: Constrained Generation

Foundation models can generate a whole host of different outputs in different styles, but can they adhere to real-world constraints?

🍽 Motivating Example: Recipe Optimizer

How would you approach the following task? What methods and tools could you use? How might it be challenging?

- **Input:** Recipe 📝
- **Prompt**: “Adapt this recipe for Aldi Süd”
- **Soft constraint**: German cuisine
- **Soft constraint**: Close to original recipe
- **Hard constraint**: Ingredients at Aldi Süd
- **Hard constraint**: No peanuts
- **Optimize:** minimize budget

Technical Focus

- Neuro-symbolic AI
- Tool-augmented LLMs
- Structured generation
- Multi-objective optimization

🚀 Impact & opportunity

Seeking 1–2 motivated MSc students (thesis or HiWi)

Contact: david.selby@dfki.de

Thank you!