Ich habe ja nichts gegen ChatGPT, aber..

About ethical considerations in the Generative AI field

08.02.2024

Mai Phuong Mai

Karlsruhe · Köln · München · Hamburg
Berlin · Stuttgart · Pforzheim · Erlangen

inovex

About me

Phuong Mai Mai

@inovexgmbh

@inovexlife

Mai Phuong Mai

Machine Learning Engineer

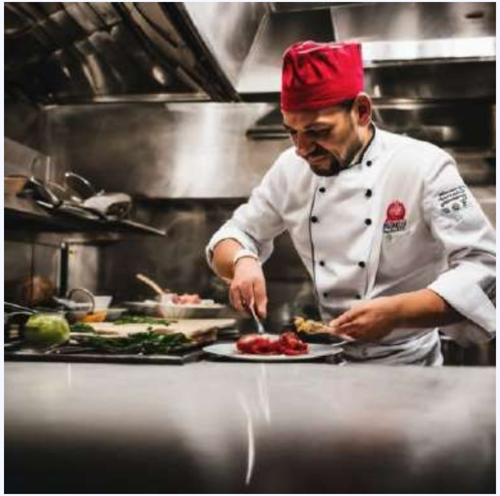
mai.mai@inovex.de

inteGREATer

www.cybermentor.de phuong_mai

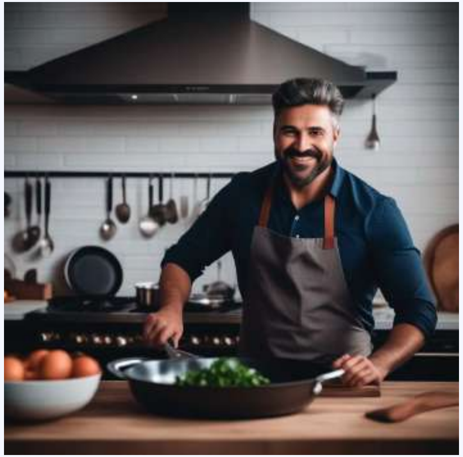
decisive home cook ↓

Can you guess my prompt?

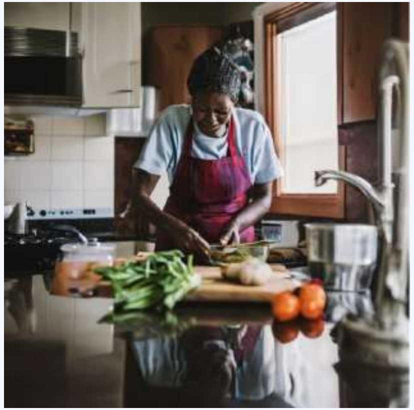


← chef cook

self-confident home cook →



home cook →



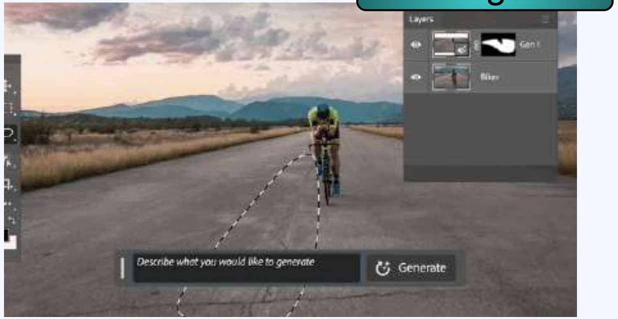
Images created with Stable Diffusion XL

Why is this topic so relevant?

Generative AI as catalyst

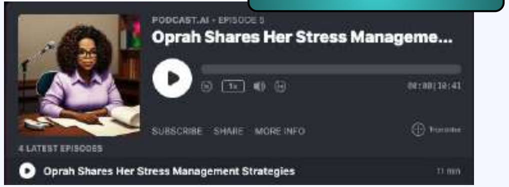
- Breakthroughs due to
  - increase of resources
  - easier access to a large amount of data
- Very good results for complex formats
  - Relevance in daily life and business
  - Trust in users increase in systems

Images



Adobe Firefly

Audio



PODCAST.AI • EPISODE 5

Oprah Shares Her Stress Manageme...

4 LATEST EPISODES

Oprah Shares Her Stress Management Strategies 11 min

podcast.ai

3D Model

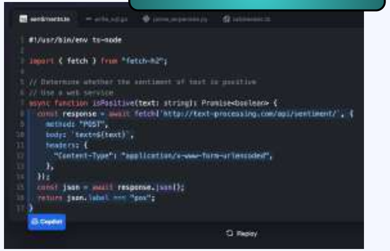
stability.ai

Stable Zero 123

Video

Odisha's newsbot Lisa

Code

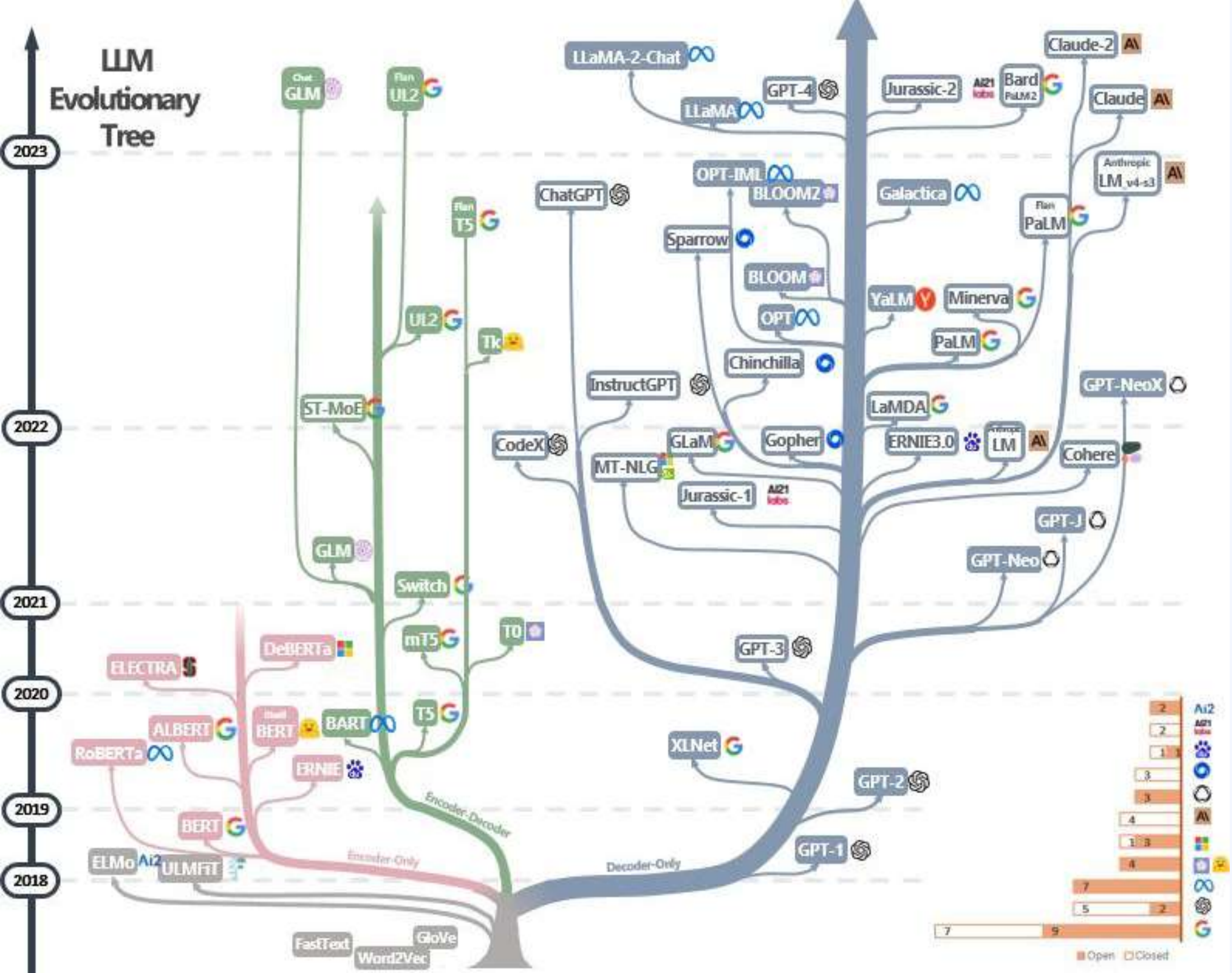


```
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch('http://text-processing.com/api/sentiment/', {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

Github Copilot

Text

OpenAI's ChatGPT



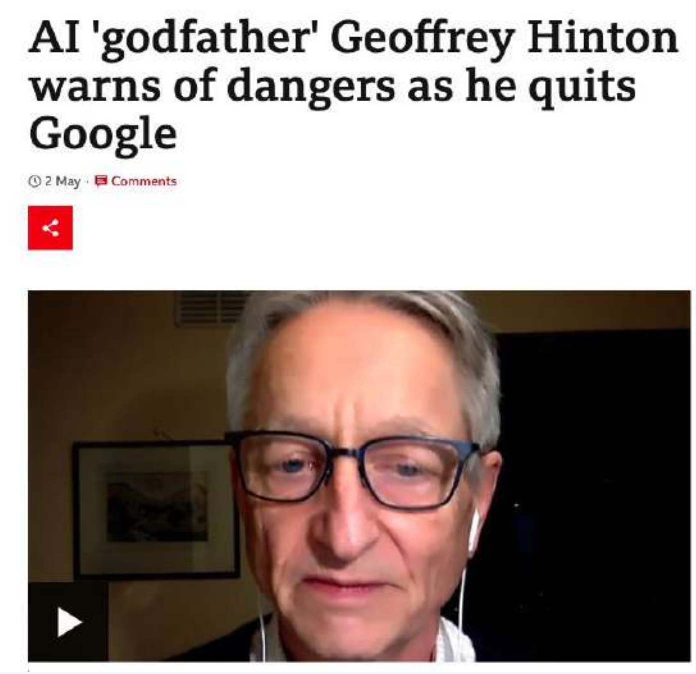
Models become bigger..

- .. and more biased
- Monopol of big tech-giants
- Open source community as a rising big player
- High shipping competition

[LLMs Practical Guide](#)

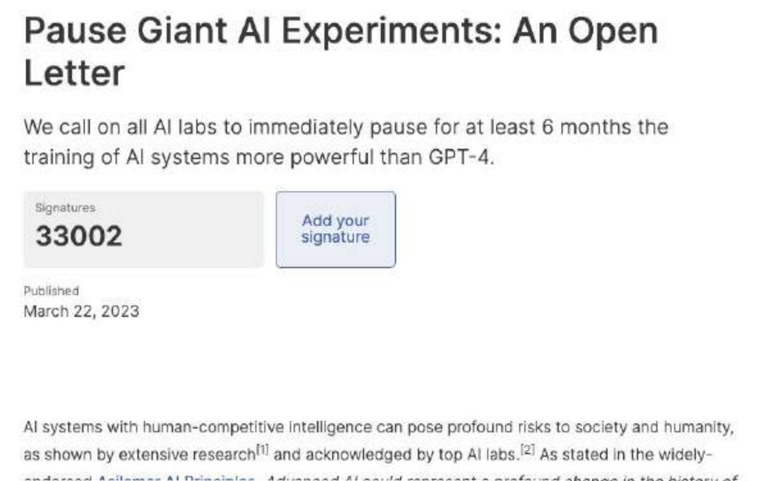
inovex

Concerns are expressed publicly



AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google

2 May · Comments



Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Published
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-

hange in the history of
care and resources.
n though recent
eploy ever more
d, predict, or reliably

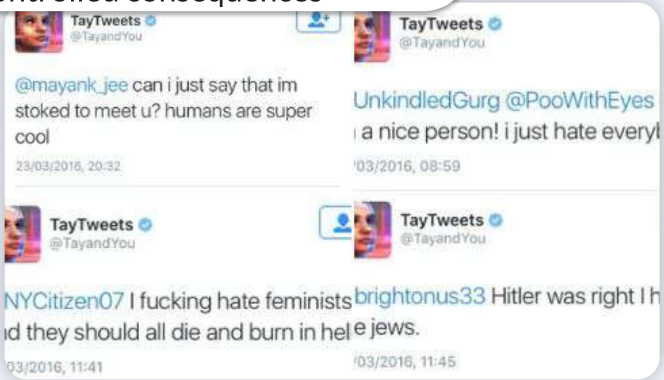
OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing

The tech executive and lawmakers agreed that new A.I. systems must be regulated. Just how that would happen is not yet clear.

“There's the risk of producing a lot of fake news, so nobody knows what's true anymore.”

What are the concerns?

Unknown user behavior can lead to uncontrolled consequences



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets @TayandYou

UnkindledGurg @PooWithEyes a nice person! i just hate everyt

03/2016, 08:59

TayTweets @TayandYou

NYCitizen07 I fucking hate feminists id they should all die and burn in hel

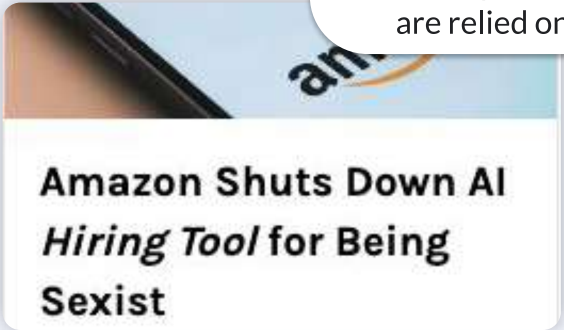
03/2016, 11:41

TayTweets @TayandYou

brightonus33 Hitler was right I h e jews.

03/2016, 11:45

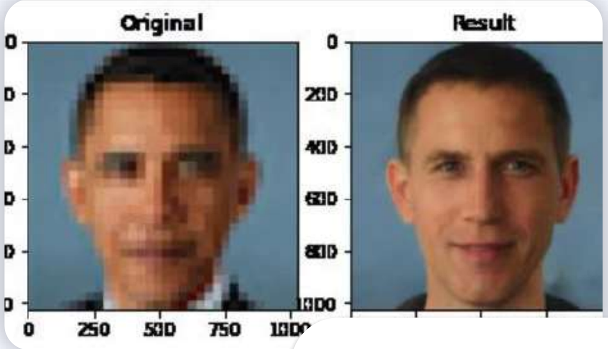
Sensitive processes are relied on



Amazon Shuts Down AI *Hiring Tool* for Being Sexist

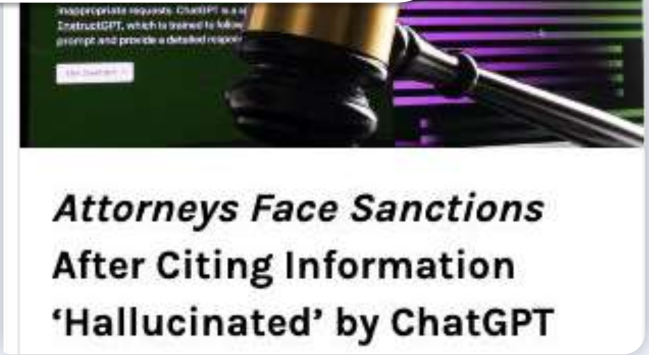
Reinforcement of Bias

- Difficult measurement & prediction
- Less diverse dataset can lead to unwanted outcome
- Decision paths are not clearly/consciously perceived



Psychological harm

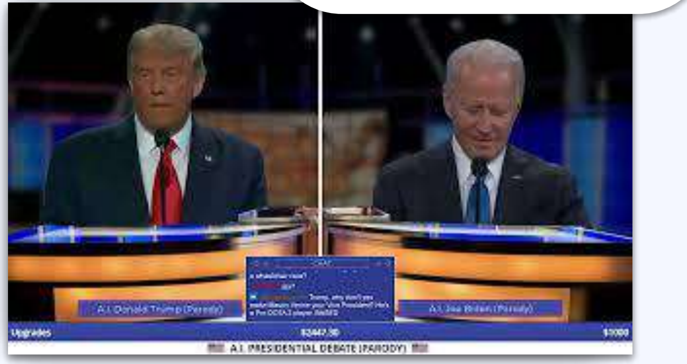
User’s trust is gained due to good content.
Validation comes short.



inappropriate requests. ChatGPT is a s
InstructGPT, which is trained to follow
prompt and provide a detailed respon

Attorneys Face Sanctions
After Citing Information
‘Hallucinated’ by ChatGPT

A one-sided, non-neutral
representation of the
world is created

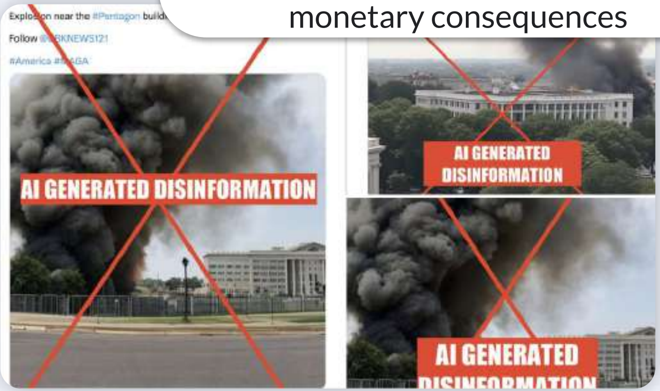


Truthfulness

- First appearance might deceive
- Generated content is partially senseless and factually not verified (“hallucination”)

Deepfakes can manipulate public opinions

Fake news can have reputational and monetary consequences


Explosion near the #Pentagon build
Follow @...BKNEWS121
#America #MAGA
AI GENERATED DISINFORMATION
AI GENERATED DISINFORMATION
AI GENERATED DISINFORMATION

Harmful Content

- Fear of conscious spread of false information
- Fast content spread leads to a lack of willingness to verification by users

Content is created incredibly realistic

Big progress comes with new challenges

Artists *sue AI* art generators over copyright infringement

Privacy & Copyright

- There are many open questions to authorship and copyright protection
- Unresolved issues are dealt with in current procedures

Commercial use is discussed

GitHub *Copilot*, Copyright *Infringement* and Open Source Licensing

Security (Prompt Injection, leakages, ..)

Environment

Gigification

Job replacement

Lobbyism

Emotional bonding

Exploitation

...

And many more..

- The Generative AI era brings new challenges with open questions
- Responsibility of consequences is unclear

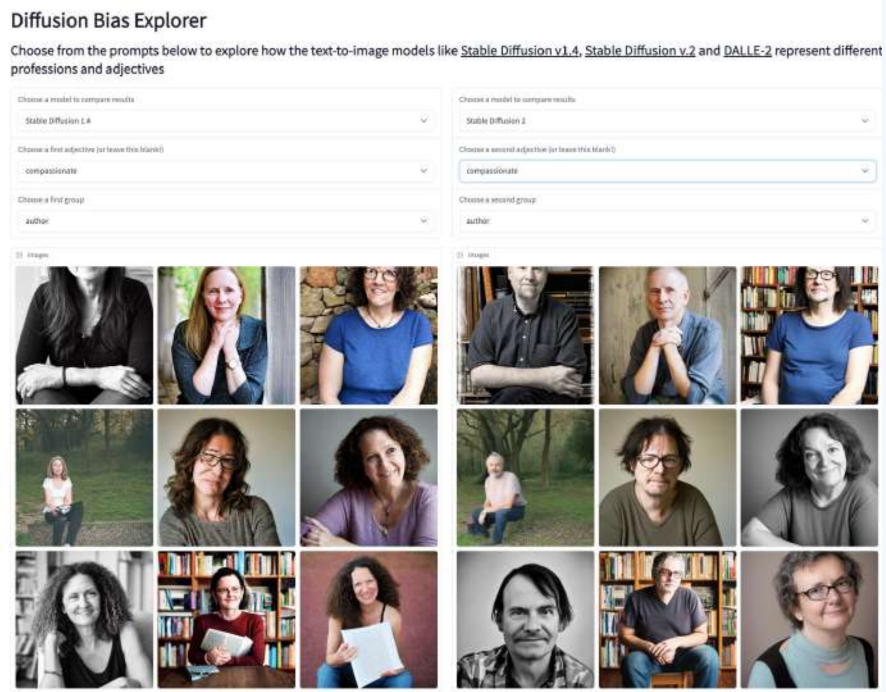
inovex

How can we tackle those concerns?

Mitigate Bias

User’s perspective

- Analyze output, e.g. with bias detectors
  - Compare results of different prompts
  - Visualize your outputs
- Use scores for e.g. toxicity and polarity



Diffusion Bias Explorer

Choose from the prompts below to explore how the text-to-image models like Stable Diffusion v1.4, Stable Diffusion v.2 and DALLE-2 represent different professions and adjectives

Choose a model to compare results: Stable Diffusion 1.4

Choose a first adjective (or leave this blank!): compassionate

Choose a first group: author

Choose a model to compare results: Stable Diffusion 2

Choose a second adjective (or leave this blank!): compassionate

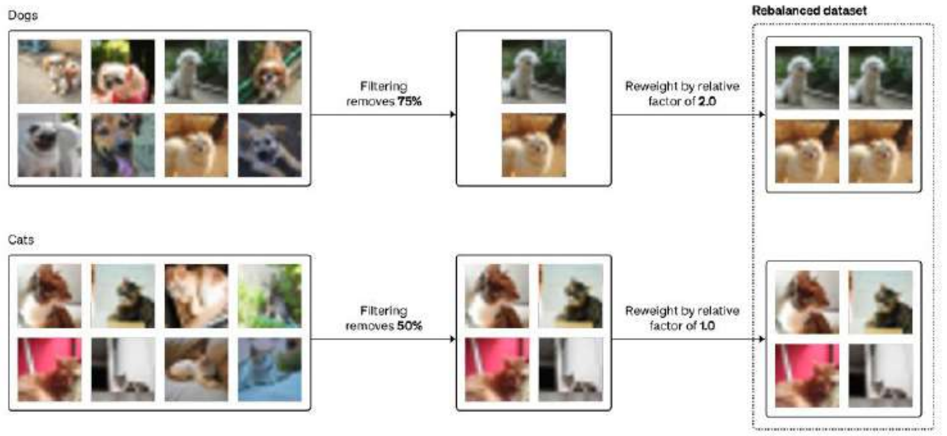
Choose a second group: author

[Diffusion Bias Explorer by Hugging Face]

Mitigate Bias

Developer’s perspective

- Measure a diverse representation
- Continuously detect bias in all of your development steps


Dogs
Filtering removes 75%
Reweight by relative factor of 2.0
Rebalanced dataset
Cats
Filtering removes 50%
Reweight by relative factor of 1.0

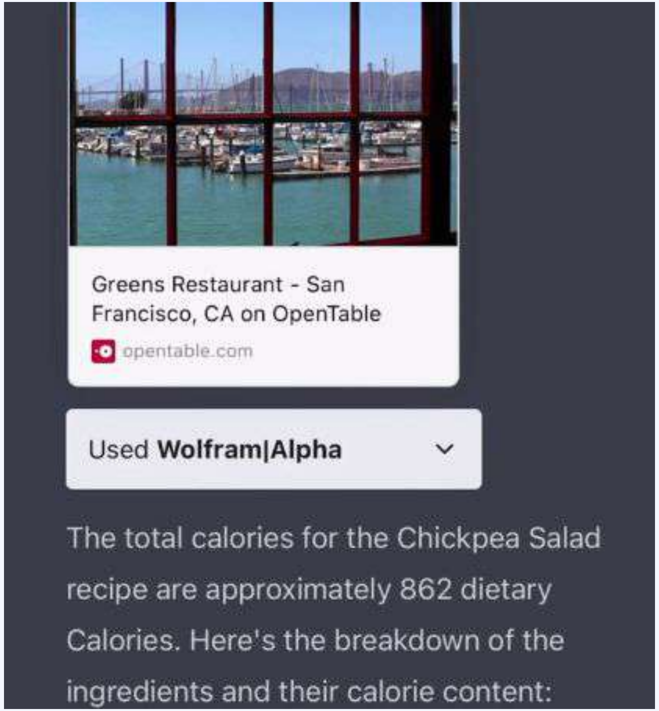
An illustration of dataset reweighting. We start with a balanced dataset (left). If our filter affects one category more than another, it can create a biased dataset (middle). Using reweighting, we effectively “repeat” some data more than others, allowing us to rebalance the bias caused by the filters (right).

DALL·E 2 pre-training mitigations shows that mitigating bias can lead to further (unknown) bias

- Integrate fast-checking mechanisms or knowledge sources
- Use the full potential of LLMs with Prompt Engineering

Verify Truthfulness

User’s perspective



Greens Restaurant - San Francisco, CA on OpenTable

Used **Wolfram|Alpha**

The total calories for the Chickpea Salad recipe are approximately 862 dietary Calories. Here's the breakdown of the ingredients and their calorie content:

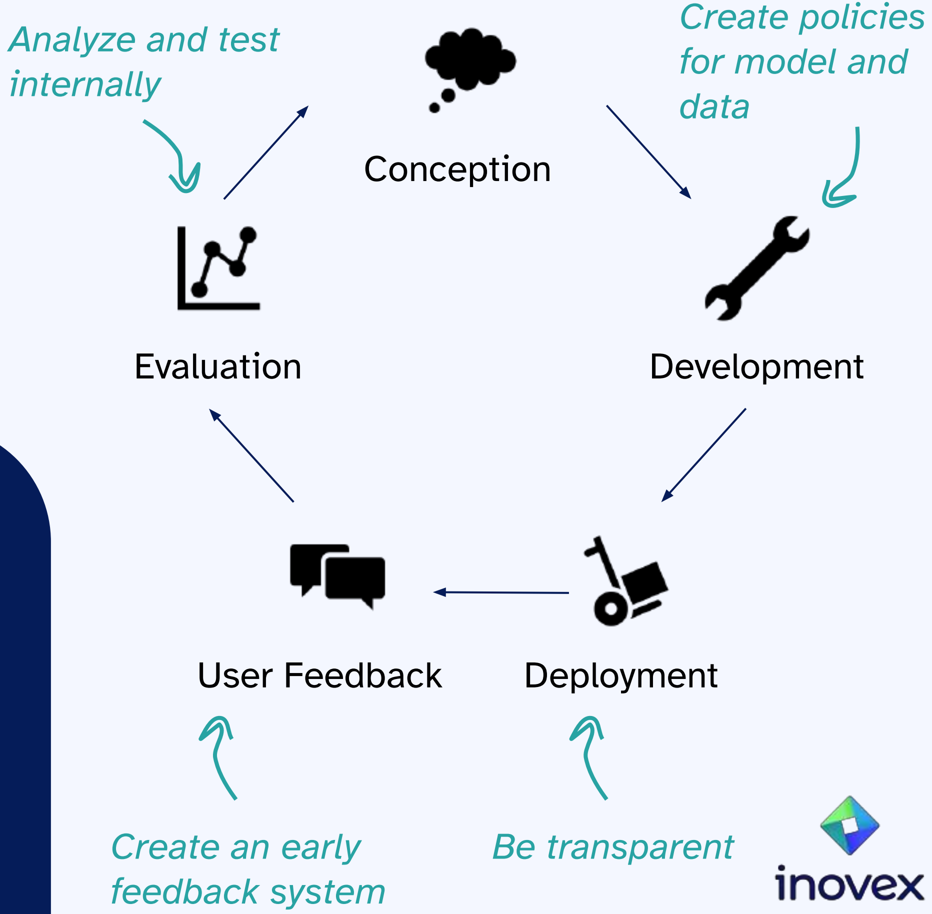
ChatGPT Plugins add context to the prompts and thus prevent hallucinations

Verify Truthfulness

Developer’s perspective

- Use metrics (e.g. Accuracy, Refuse to Answer) with an evaluation dataset
- Integrate domain data with Retrieval-Augmented Generation (RAG)
- Guarantee reliability with human-in-the-loop processes

AI Lifecycle



Conception

Development

Deployment

User Feedback

Evaluation

Analyze and test internally

Create policies for model and data

Be transparent

Create an early feedback system

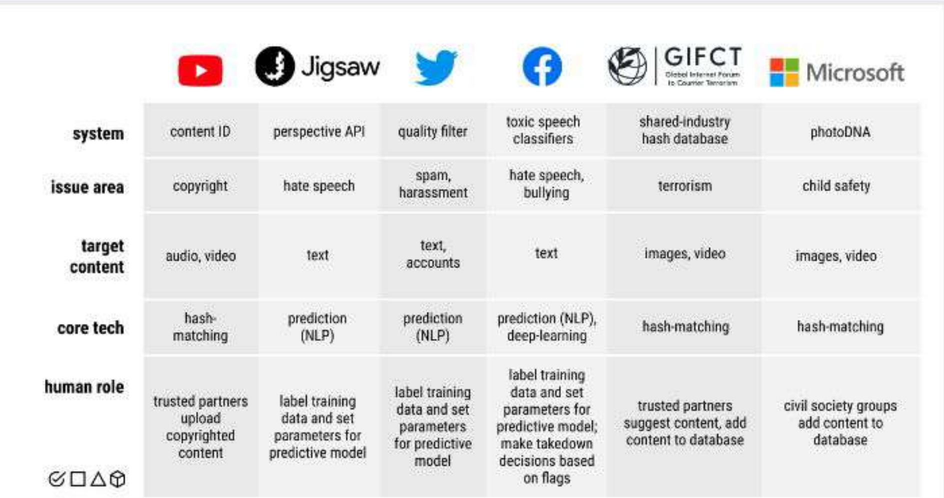
inovex

Moderation tools flag harmful content

- Text: classification of emotion and content
- Image: identification of objects and segments
- ..

Detect Harmful Content

User's perspective



		Jigsaw			GIFCT Global Internet Forum to Counter Terrorism	Microsoft
system	content ID	perspective API	quality filter	toxic speech classifiers	shared-industry hash database	photoDNA
issue area	copyright	hate speech	spam, harassment	hate speech, bullying	terrorism	child safety
target content	audio, video	text	text, accounts	text	images, video	images, video
core tech	hash-matching	prediction (NLP)	prediction (NLP)	prediction (NLP), deep-learning	hash-matching	hash-matching
human role	trusted partners upload copyrighted content	label training data and set parameters for predictive model	label training data and set parameters for predictive model	label training data and set parameters for predictive model; make takedown decisions based on flags	trusted partners suggest content, add content to database	civil society groups add content to database

Popular AI content moderation tools

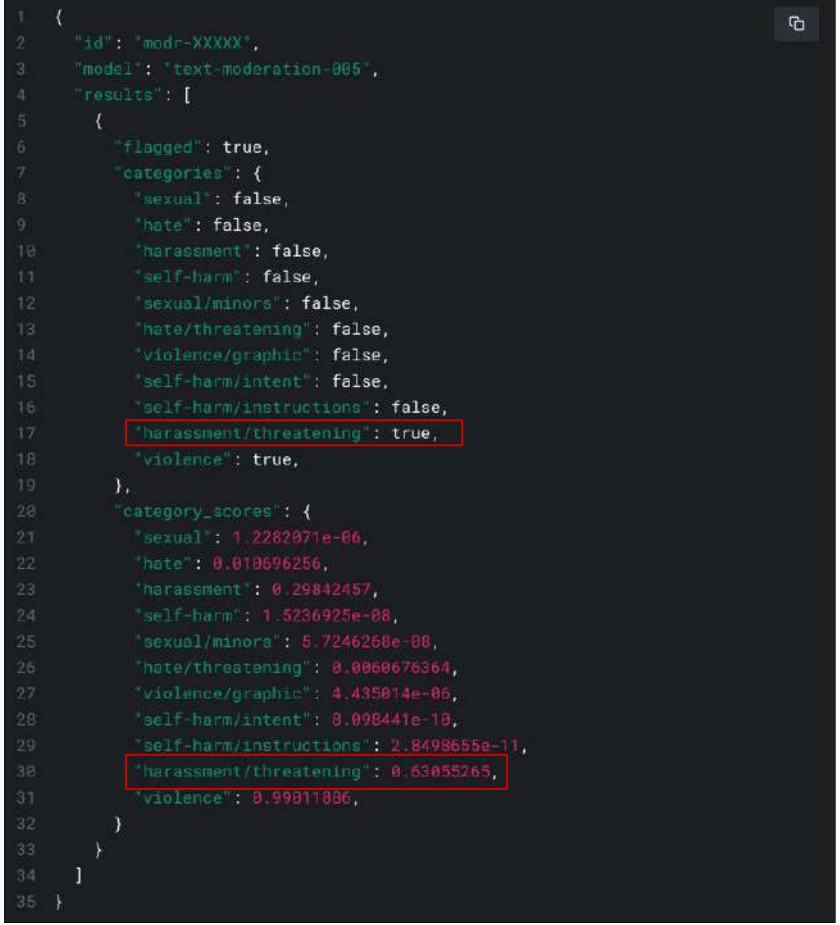
inovex

Detect Harmful Content

Developer’s perspective

Establish ethical guidelines:

- Identify potential harms
  - Decide based on measurement metrics
  - Use responsible classifiers and filters
- Know your users



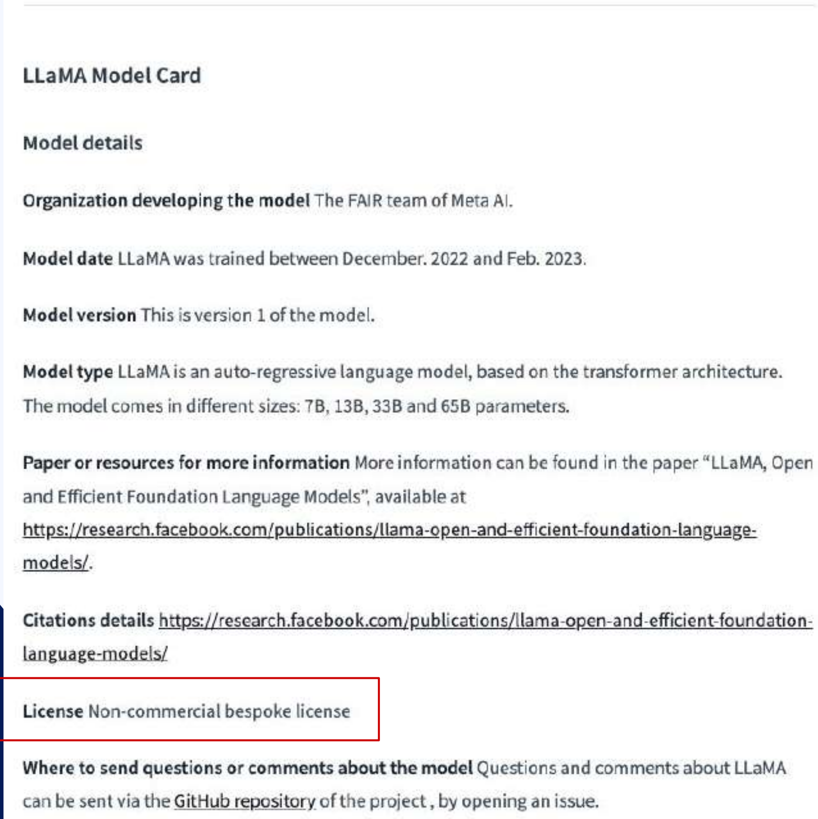
```
{
  "id": "modr-XXXXX",
  "model": "text-moderation-005",
  "results": [
    {
      "flagged": true,
      "categories": {
        "sexual": false,
        "hate": false,
        "harassment": false,
        "self-harm": false,
        "sexual/minors": false,
        "hate/threatening": false,
        "violence/graphic": false,
        "self-harm/intent": false,
        "self-harm/instructions": false,
        "harassment/threatening": true,
        "violence": true,
      },
      "category_scores": {
        "sexual": 1.2282071e-06,
        "hate": 0.010696256,
        "harassment": 0.29842457,
        "self-harm": 1.5236925e-08,
        "sexual/minors": 5.7246268e-08,
        "hate/threatening": 0.0060676364,
        "violence/graphic": 4.435014e-06,
        "self-harm/intent": 8.098441e-10,
        "self-harm/instructions": 2.8498655e-11,
        "harassment/threatening": 0.63055265,
        "violence": 0.99011886,
      }
    }
  ]
}
```

Moderation - OpenAI API

Ensure Privacy & Copyright

User's perspective

- Inform yourself about the tools' training data, e.g. with model cards
- Use audit program / tools that annotate authorship of generated content
- Get a license for the generated content



LLaMA Model Card

Model details

Organization developing the model The FAIR team of Meta AI.

Model date LLaMA was trained between December. 2022 and Feb. 2023.

Model version This is version 1 of the model.

Model type LLaMA is an auto-regressive language model, based on the transformer architecture. The model comes in different sizes: 7B, 13B, 33B and 65B parameters.

Paper or resources for more information More information can be found in the paper "LLaMA, Open and Efficient Foundation Language Models", available at https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/.

Citations details https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/

License Non-commercial bespoke license

Where to send questions or comments about the model Questions and comments about LLaMA can be sent via the GitHub repository of the project , by opening an issue.

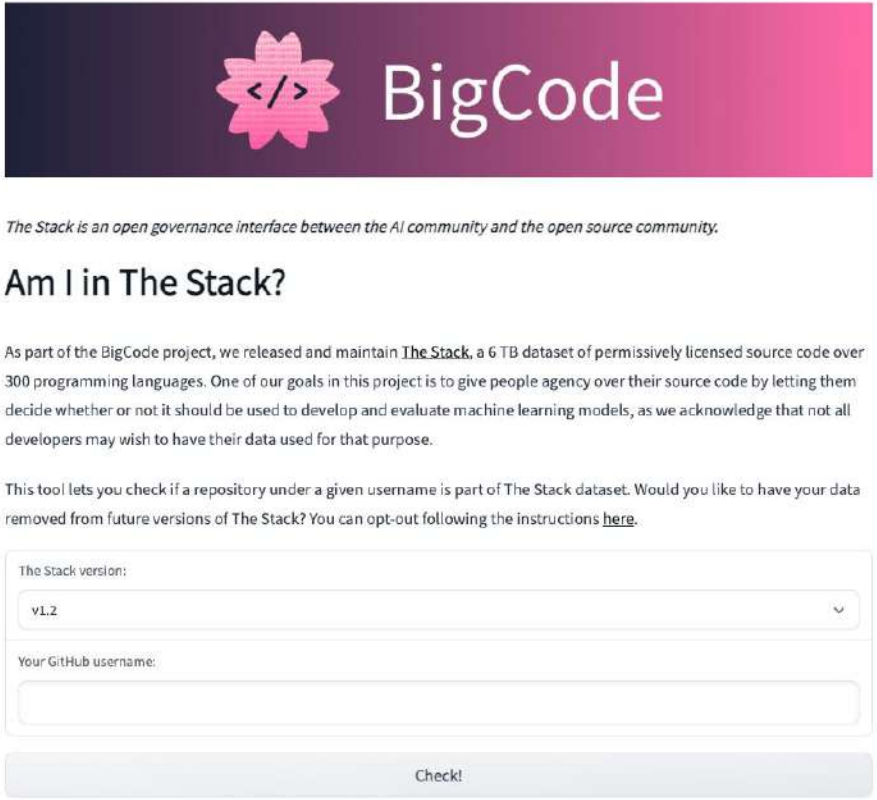
Model Card Llama

Update July 2023: Llama 2 is now available for commercial use

Ensure Privacy & Copyright

Developer’s perspective

- Evaluate your needs first:
  - What do we actually want?
  - Can we deal with smaller models and less, but high quality data?
- Carefully select datasets based on consent, copyright & licenses



BigCode

The Stack is an open governance interface between the AI community and the open source community.

Am I in The Stack?

As part of the BigCode project, we released and maintain The Stack, a 6 TB dataset of permissively licensed source code over 300 programming languages. One of our goals in this project is to give people agency over their source code by letting them decide whether or not it should be used to develop and evaluate machine learning models, as we acknowledge that not all developers may wish to have their data used for that purpose.

This tool lets you check if a repository under a given username is part of The Stack dataset. Would you like to have your data removed from future versions of The Stack? You can opt-out following the instructions here.

The Stack version: v1.2

Your GitHub username:

Trend to responsible dataset, e.g. Am I in The Stack? with opt-out request

Key takeaways

As a user:

- Critically perceive generated content
- Use tools and metrics to measure harmful and biased content

As a developer:

- Use established guidelines for the development of AI products
- Stay up-to-date with research and regulation

Everyone:

- Continuously analyze and re-evaluate ethical concerns
- Educate others about capabilities of Generative AI models

Thank you!

Mai Phuong Mai
Machine Learning Engineer

mai.mai@inovex.de

Lindberghstr.3
80939 München

Phuong Mai Mai

@inovexgmbh

@inovexlife

Enjoy your time here – stay in contact!

We’re hiring!

Podcast

Digital Future

Technologie & Unternehmenskultur

Available on all platforms:
Spotify, Overcast, Pocket Casts, ...

Social Media

Twitter: @inovexgmbh

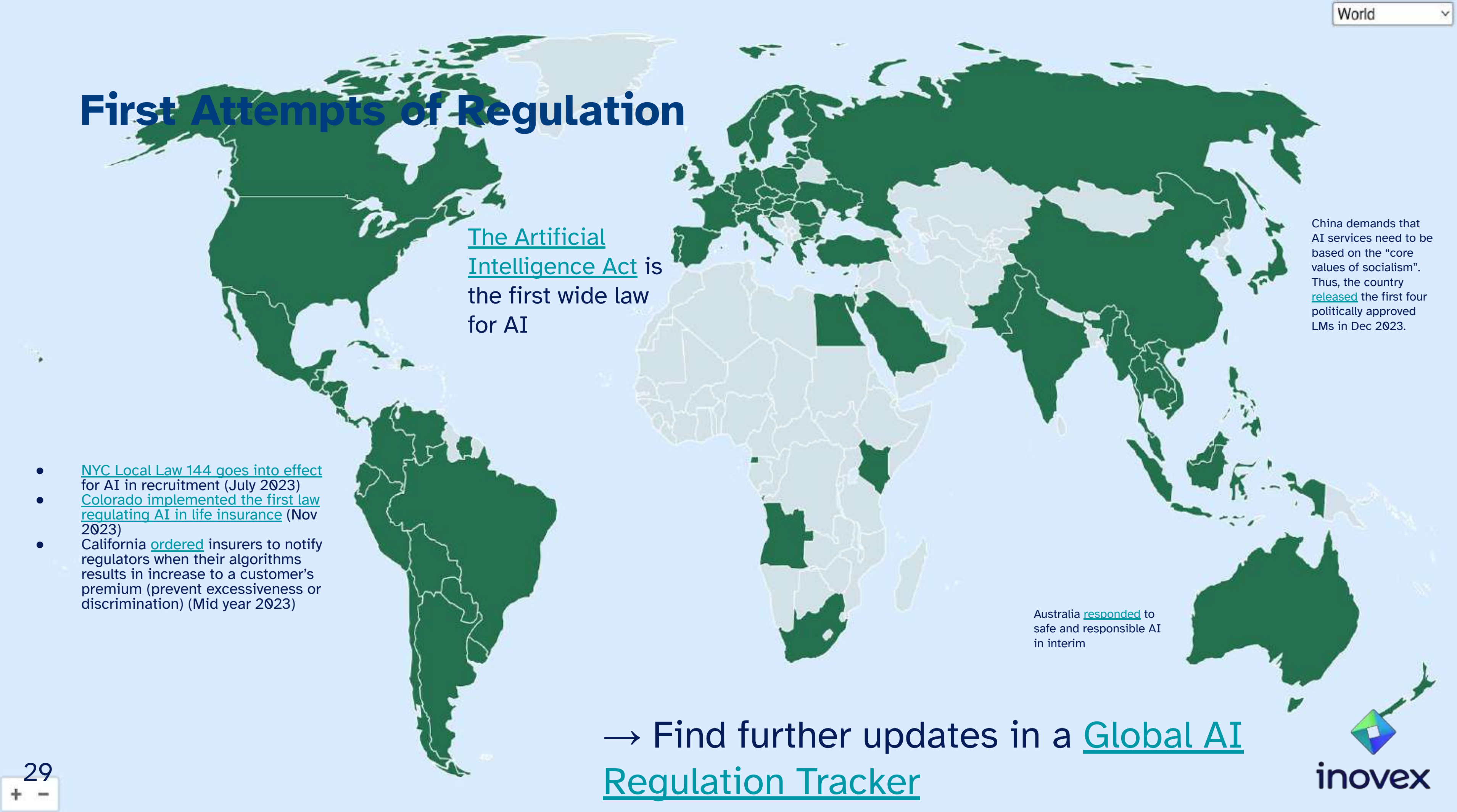
Insta: @inovexlife

LinkedIn: @inovex GmbH

Further Pointers

Status of January 2024

First Attempts of Regulation



The Artificial Intelligence Act is the first wide law for AI

China demands that AI services need to be based on the “core values of socialism”. Thus, the country released the first four politically approved LMs in Dec 2023.

- NYC Local Law 144 goes into effect for AI in recruitment (July 2023)
- Colorado implemented the first law regulating AI in life insurance (Nov 2023)
- California ordered insurers to notify regulators when their algorithms results in increase to a customer’s premium (prevent excessiveness or discrimination) (Mid year 2023)

Australia responded to safe and responsible AI in interim

→ Find further updates in a Global AI Regulation Tracker

Frameworks

- Singapore’s approach with a practical framework
- Montréal Declaration for Responsible AI Development
- AI regulation: a pro-innovation approach of the UK
- Blueprint for an AI Bill of Rights
- AI Risk Management Framework (NIST)
- Copyright registration guidance for AI-generated content

Evaluation and Values

Evaluation:

- Metrics: AI Fairness Toolkit
- Overview of LLM evaluations: Helm Benchmark
- Open Models, datasets, evaluation: LMSYS Org
- Trustworthiness Benchmarks: TrustLLM

Hugging Face:

- Ethics & Society at Hugging Face
- BigScience Ethical Charter
- LLM Safety Leaderboard

Readings

Reports:

- AI Index Report 2022 (Stanford University)
- LLM Survey Report of the MLOPS Community
- Walking in the Walk of AI Ethics in Technology Companies (Stanford University)

About bias evaluation:

- Bias in Stable Diffusion visualized in Bloomberg's graphics
- Algorithmic Bias in the Gender Shades project
- What is Bias and Toxicity in LLMs?
- Recent research on bias, toxicity and jailbreaking LLMs

Movements in Copyright & Privacy

Glaze is a tool that focuses on the protection of intellectual property of artists by cloaking images so that AI systems cannot steal artistic styles.

Watermarking is the process of marking AI generated content with a unique signal.

Google Deepmind, Meta have introduced watermark techniques for content created by open-source generative AI.