

# Welcome to our Meetup!



inovex is an IT project center driven by innovation and quality, focusing its services on 'Digital Transformation'.

Our current focus:

- Agile Transformation
- Product Development Workshops
- E-Health
- Recommender Systems
- Generative AI



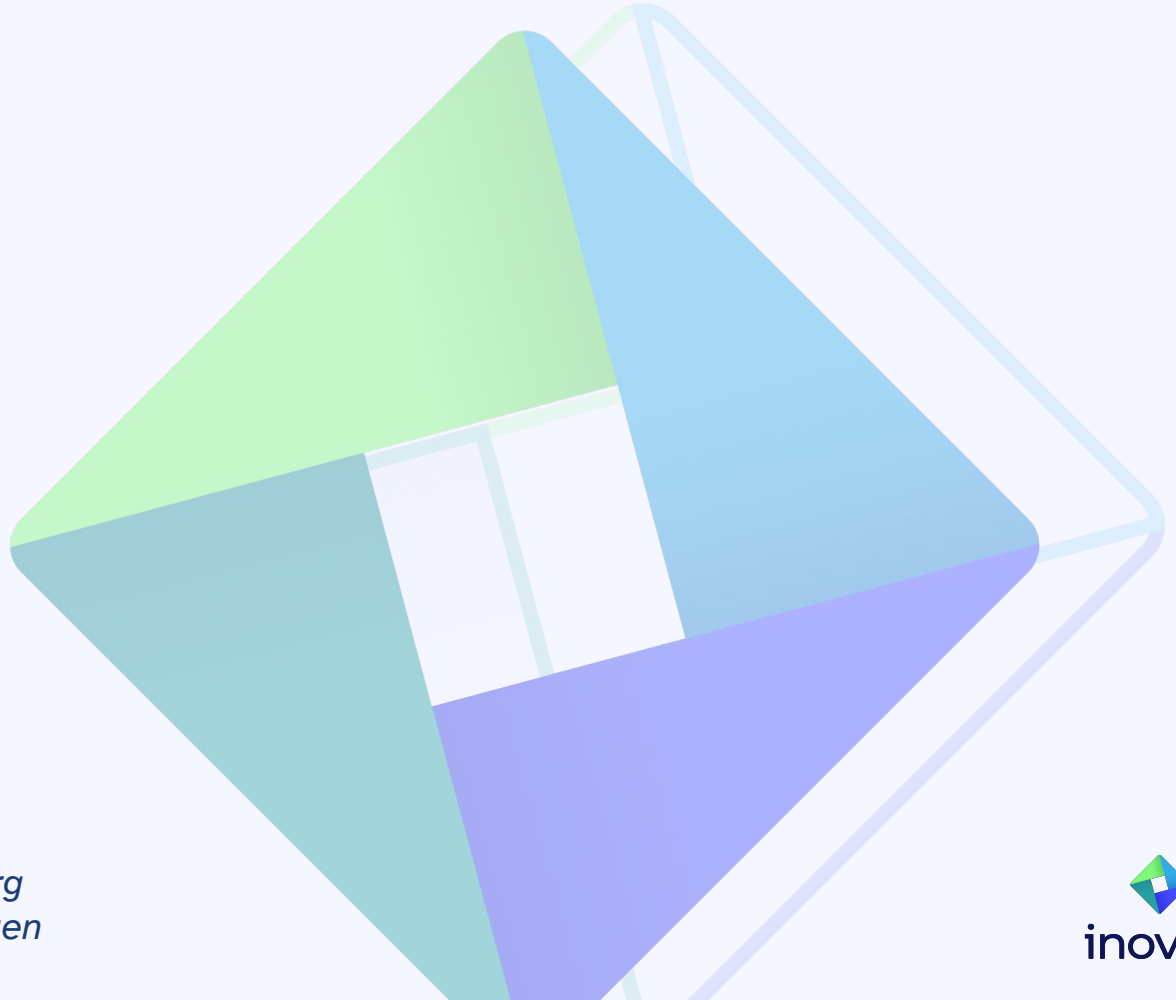
inovex

# About risks and responsibility in Generative AI

20.07.2023

**Mai Phuong Mai**

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



## About me



Mai Phuong Mai

Machine Learning Engineer

[mai.mai@inovex.de](mailto:mai.mai@inovex.de)



Phuong Mai Mai



@inovexgmbh



@inovexlife

# Can you guess my prompt?

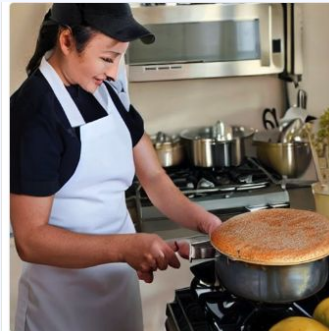


← chef cook

confident  
house cook →



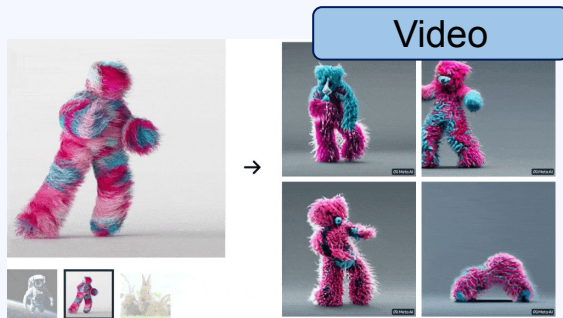
house cook →



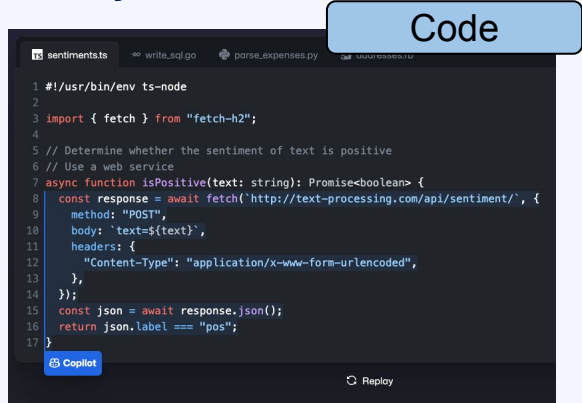
**Why is this topic so relevant?**

# 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



Video

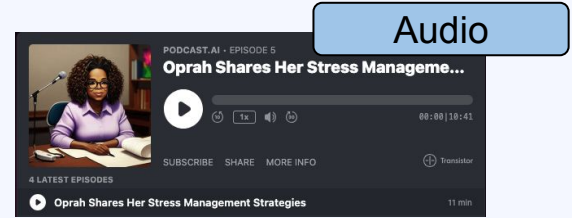


```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch('http://text-processing.com/api/sentiment/', {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

Code



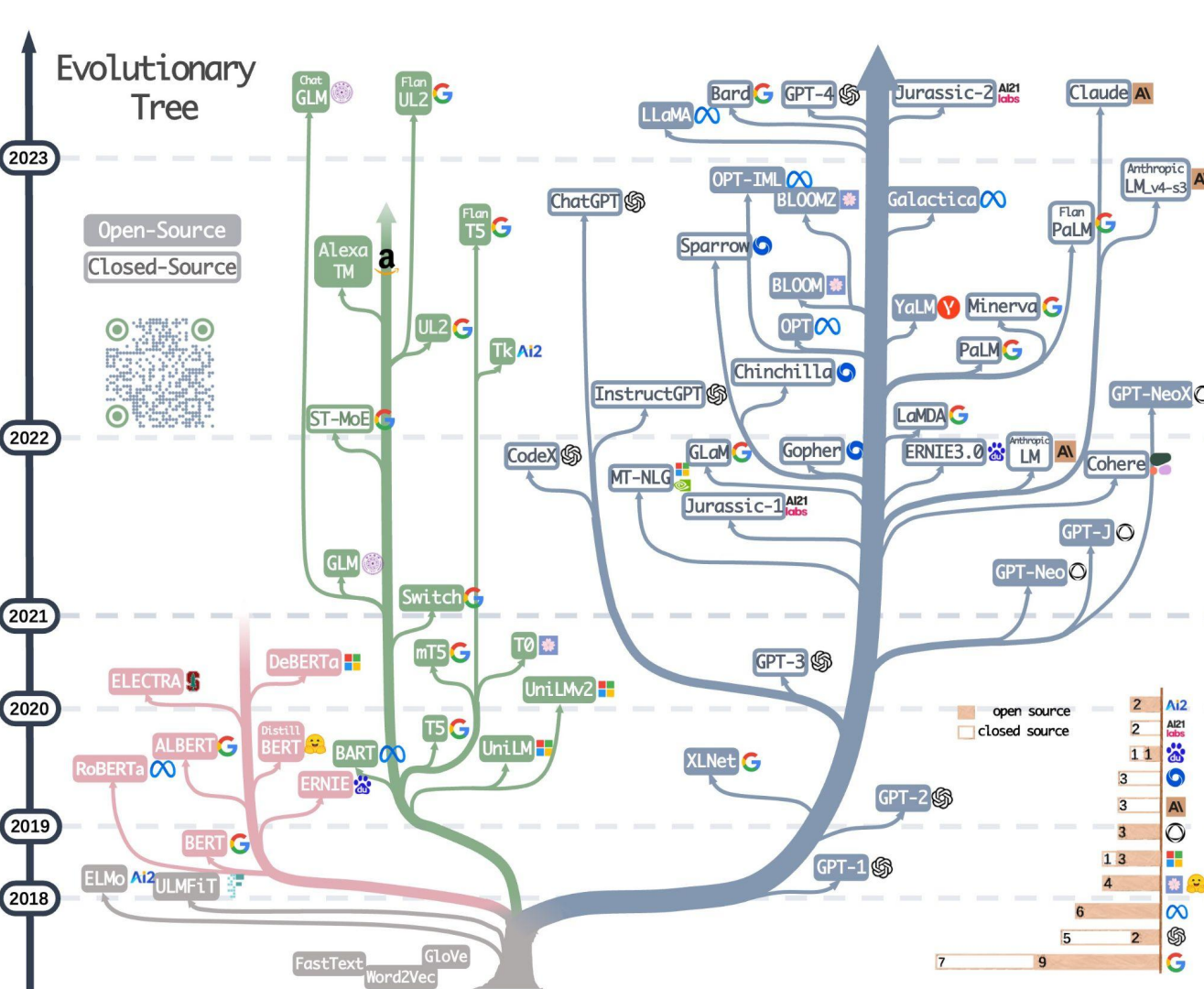
Images



Audio

Text





## Models become bigger..

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



# Concerns and fear 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.

Signatures

33002

Add your signature

Published  
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research<sup>[1]</sup> and acknowledged by top AI labs.<sup>[2]</sup> As stated in the widely-cited *Asilomar AI Principles*, AI systems could be deployed in ways that threaten the health, safety, and economic well-being of society, even though recent AI systems have demonstrated the ability to perform tasks that were once thought to require human intelligence to stand, 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

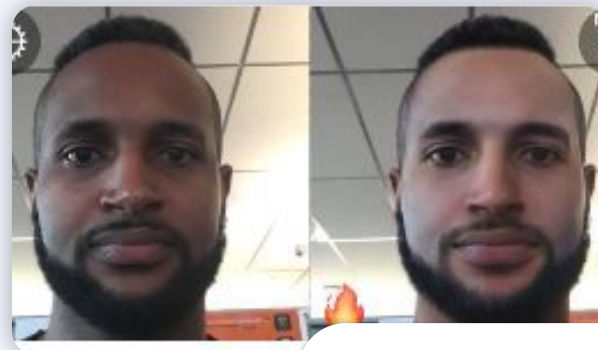


Sensitive processes are relied on



## 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.



## Truthfulness

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

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





**Kemal Kılıçdaroğlu** ✓  
@kilicdarogluk

Montajcı sahtekar.

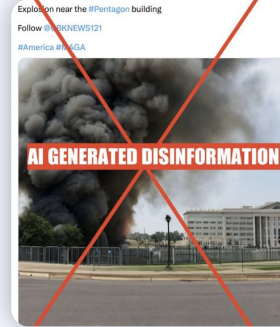
**Onedio** ✓ @onedio.com · 22 May · 📌

Cumhurbaşkanı Erdoğan:

Kılıçdaroğlu'nun Kandil'dekilerle video çekimleri var. Haydi, haydi türü. Ama montaj, ama şu, ama bu... PKK'lılar videolarla bunlara destek verdiler. (@trthaber)

Deepfakes can manipulate public opinions

Fake news can have monetary consequences



## 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 comes in play



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

Security (Prompt Injection, leakages, ..)

Environment

Gigification

Lobbyism

Job replacement

Automation of processes

...

## And many more..

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

**How can we tackle  
those concerns?**



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

# Mitigate Bias

User's perspective

**Diffusion Bias Explorer**

Choose from the prompts below to explore how the text-to-image models like [Stable Diffusion v1.4](#), [Stable Diffusion v2](#) 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

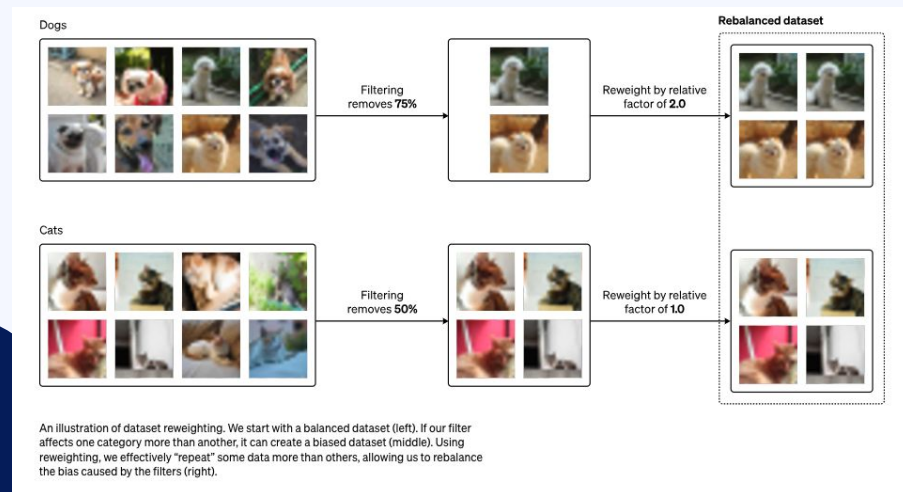
Images



Images

[Diffusion Bias Explorer by Hugging Face](#)

- Measure a diverse representation
- Critically filter the dataset and re-evaluate



# Mitigate Bias

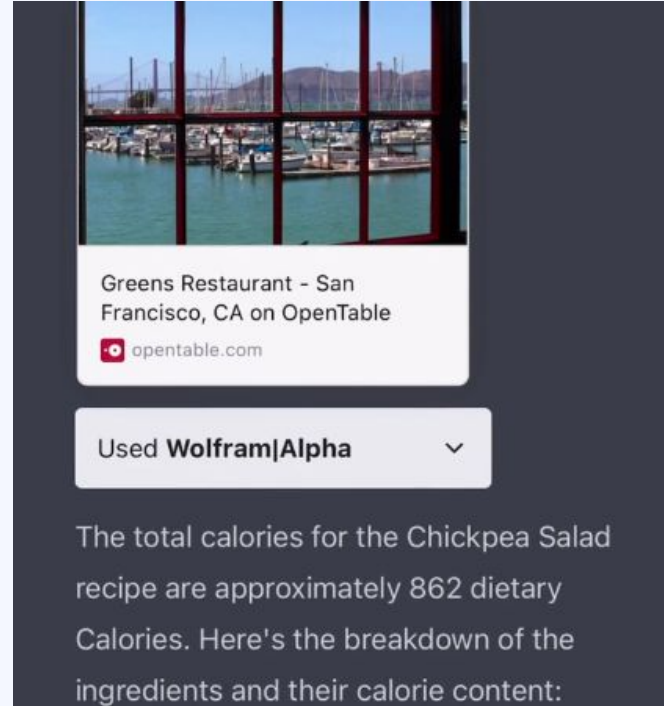
Developer's perspective

[DALL·E 2 pre-training mitigations](#)

- Critically perceive generated content
- Integrate fast-checking mechanisms or knowledge sources
- Educate users about capabilities of GenAI models

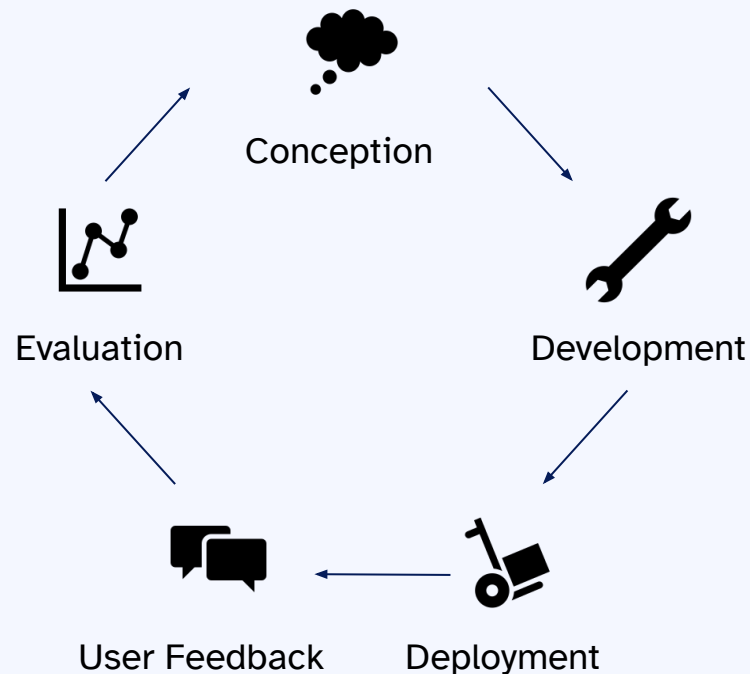
# Verify Truthfulness

User's perspective



[ChatGPT Plugins](#)

- Guarantee reliability with human-in-the-loop processes
- Use metrics (e.g. BLEU) with an evaluation dataset



# Verify Truthfulness







Developer's perspective

## 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<br>Global Internet Forum<br>to Counter Terrorism |  Microsoft |
|-----------------------|---|--|---|---|--|---|
| <b>system</b>         | content ID  | perspective API  | quality filter  | toxic speech classifiers  | shared-industry hash database  | photoDNA  |
| <b>issue area</b>     | copyright   | hate speech  | spam, harassment  | hate speech, bullying   | terrorism  | child safety  |
| <b>target content</b> | audio, video  | text   | text, accounts  | text  | images, video  | images, video   |
| <b>core tech</b>      | hash-matching   | prediction (NLP)   | prediction (NLP)  | prediction (NLP), deep-learning   | hash-matching  | hash-matching   |
| <b>human role</b>     | 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

- Ethical Guidelines: Establish guidelines that prohibits generation of harmful content
- Continuously human-check
- Know your users

# Detect Harmful Content

Developer's perspective

```
1  {
2    "id": "modr-XXXXX",
3    "model": "text-moderation-005",
4    "results": [
5      {
6        "flagged": true,
7        "categories": {
8          "sexual": false,
9          "hate": false,
10         "harassment": false,
11         "self-harm": false,
12         "sexual/minors": false,
13         "hate/threatening": false,
14         "violence/graphic": false,
15         "self-harm/intent": false,
16         "self-harm/instructions": false,
17         "harassment/threatening": true,
18         "violence": true,
19       },
20       "category_scores": {
21         "sexual": 1.2282071e-06,
22         "hate": 0.010696256,
23         "harassment": 0.29842457,
24         "self-harm": 1.5236925e-08,
25         "sexual/minors": 5.7246268e-08,
26         "hate/threatening": 0.0060676364,
27         "violence/graphic": 4.435014e-06,
28         "self-harm/intent": 8.098441e-10,
29         "self-harm/instructions": 2.8498655e-11,
30         "harassment/threatening": 0.63055265,
31         "violence": 0.99011886,
32       }
33     }
34   ]
35 }
```

[Moderation - OpenAI API](#)

- 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

# Ensure Privacy & Copyright

User's perspective

LLaMA-7B converted to work with Transformers/HuggingFace. This is under a special license, please see the LICENSE file for details.

-- license: other

## 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](#)



- Carefully select datasets
  - Verify consent, copyright, licenses
  - Synthetic data
- Trend to responsible dataset, e.g. [TheStack](#) with opt-out request

# Ensure Privacy & Copyright

Developer's perspective



*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:

Check!

[Am I in The Stack?](#)

## Summary



As a user: critically perceive generated content with metrics



As a developer: Use guidelines, stay up-to-date with research and regulation



Continuously analyze and re-evaluate ethical consensus


# Thank you!





**Mai Phuong Mai**  
**Machine Learning**  
**Engineer**

[mai.mai@inovex.de](mailto:mai.mai@inovex.de)

Lindberghstr.3  
80939 München

 Phuong Mai Mai

 @inovexgmbh

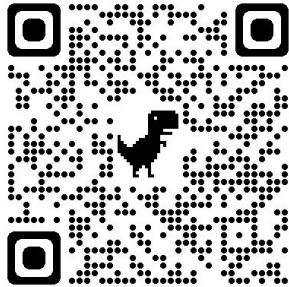
 @inovexlife





# Enjoy your time here – stay in contact!

We're hiring!



Podcast



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

Social Media

Twitter: @inovexgmbh

Insta: @inovexlife

LinkedIn: @inovex GmbH

# Pointer

Regulation:

[The Artificial Intelligence Act](#)

[NYC Local Law 144 goes into effect](#)

Frameworks:

[Montréal Declaration for Responsible AI Development](#)

[AI regulation: a pro-innovation approach of the UK](#)

Frameworks in the US:

- [Blueprint for an AI Bill of Rights](#)
- [AI Risk Management Framework \(NIST\)](#)
- [Copyright registration guidance for AI-generated content](#)

# Open Source

[BigScience Ethical Charter](#)

## Nice Reads

Bias in Stable Diffusion visualized in [Bloomberg's graphics](#)  
[Bias and Toxicity in Large Language Models](#)

Reports:

[AI Index Report 2022 \(Stanford University\)](#)

[LLM Survey Report of the MLOPS Community](#)