CS 435
Lecture 5

Adnan Masood, PhD.

Debiasing Measures and Metrics

Recap Quiz

Applied Ethical AI

Build your own LLM – Part 1 (BERT Style LM)

Assignment

Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2, and other models, locally.

Download ↓

Available for macOS, Linux, and Windows

LM Studio

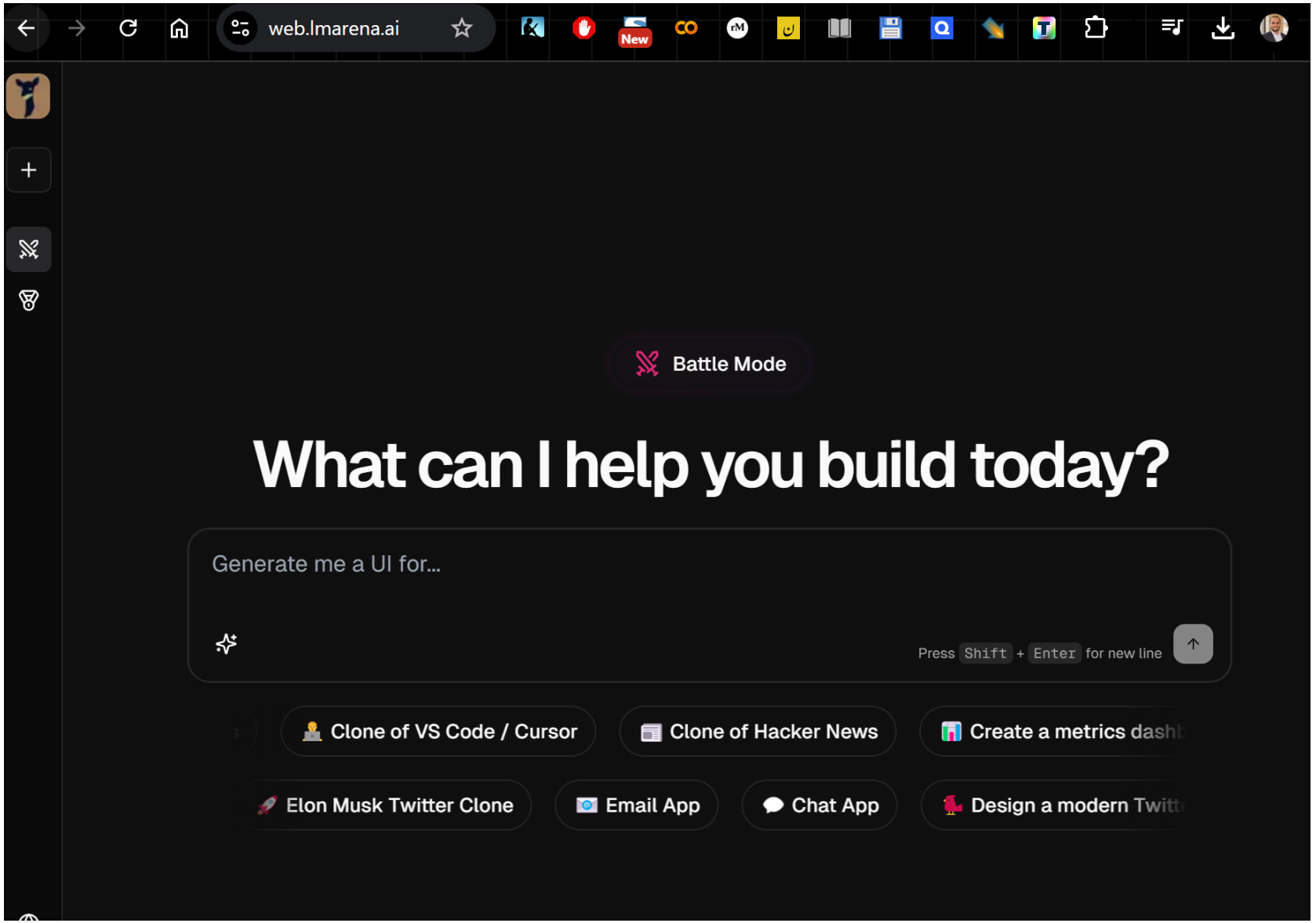
Discover, download, and run local LLMs

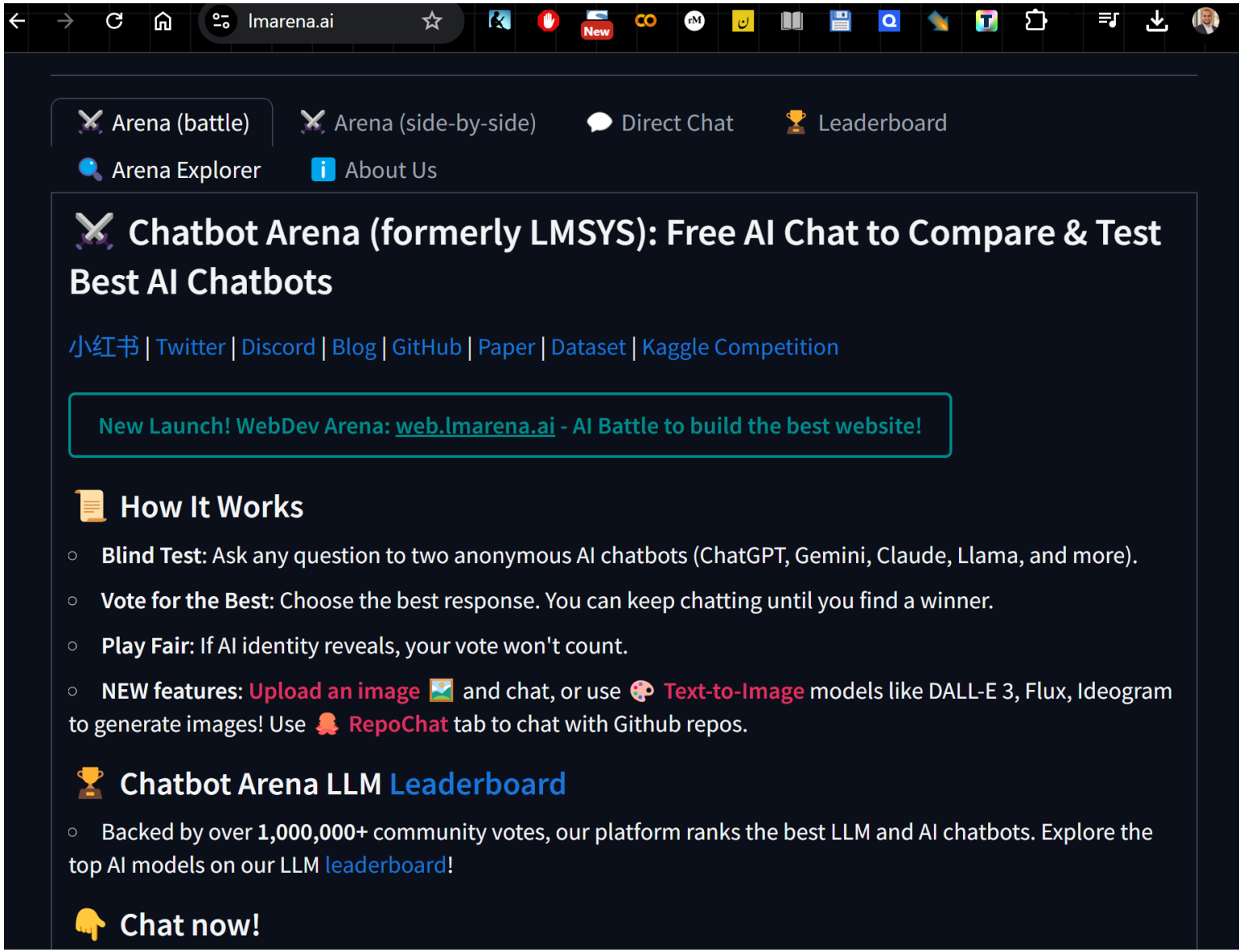
Run Llama 3.2 Mistral Phi Gemma DeepSeek Qwen 2.5 on your computer

Download LM Studio for Mac (M series) 0.3.9

Download LM Studio for Windows 0.3.9

Download LM Studio for Linux 0.3.9


web.lmarena.ai
New
Battle Mode
What can I help you build today?
Generate me a UI for...
Press Shift + Enter for new line
Clone of VS Code / Cursor
Clone of Hacker News
Create a metrics dashb
Elon Musk Twitter Clone
Email App
Chat App
Design a modern Twitt


lmarena.ai
New
⚔️ Arena (battle)
⚔️ Arena (side-by-side)
Direct Chat
Leaderboard
Arena Explorer
About Us
⚔️ Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots
小红书 | Twitter | Discord | Blog | GitHub | Paper | Dataset | Kaggle Competition
New Launch! WebDev Arena: web.lmarena.ai - AI Battle to build the best website!
📜 How It Works
Blind Test: Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).
Vote for the Best: Choose the best response. You can keep chatting until you find a winner.
Play Fair: If AI identity reveals, your vote won't count.
NEW features: Upload an image 🖼️ and chat, or use 🎨 Text-to-Image models like DALL-E 3, Flux, Ideogram to generate images! Use RepoChat tab to chat with Github repos.
🏆 Chatbot Arena LLM Leaderboard
Backed by over 1,000,000+ community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!
👇 Chat now!

TL;DS - 21 fairness definition and their politics by Arvind Narayanan

https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

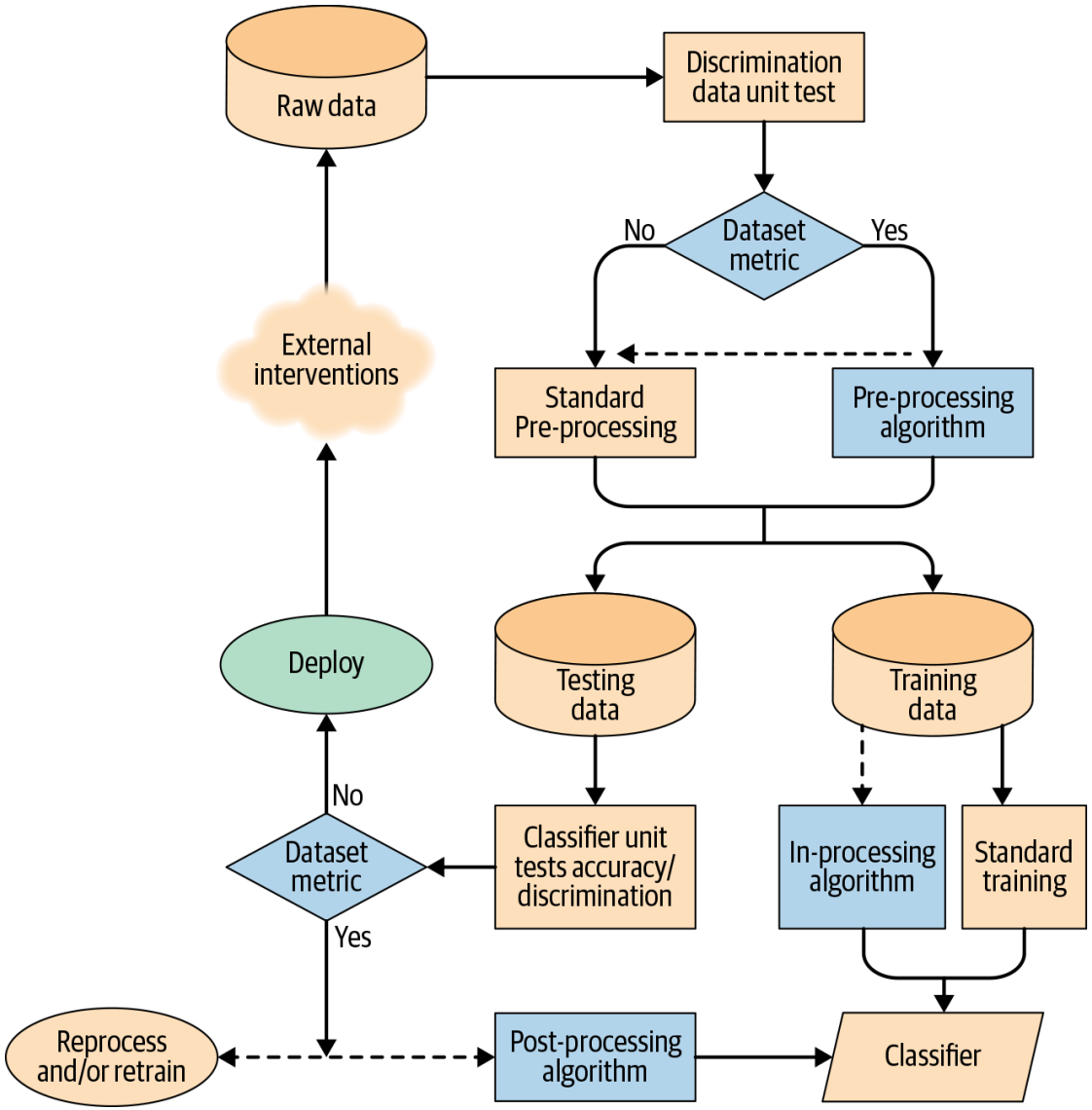
AI Fairness 360 (AIF360)

Continuous Integration failing | docs passing | pypi package 0.6.1 | CRAN not published

The AI Fairness 360 toolkit is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AI Fairness 360 package is available in both Python and R.

The AI Fairness 360 package includes

1. a comprehensive set of metrics for datasets and models to test for biases,
2. explanations for these metrics, and
3. algorithms to mitigate bias in datasets and models. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.



Supported bias mitigation algorithms

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al., 2018)
- Rich Subgroup Fairness (Kearns, Neel, Roth, Wu, 2018)
- Exponentiated Gradient Reduction (Agarwal et al., 2018)
- Grid Search Reduction (Agarwal et al., 2018, Agarwal et al., 2019)
- Fair Data Adaptation (Plečko and Meinshausen, 2020, Plečko et al., 2021)
- Sensitive Set Invariance/Sensitive Subspace Robustness (Yurochkin and Sun, 2020, Yurochkin et al., 2019)

Supported fairness metrics

- Comprehensive set of group fairness metrics derived from selection rates and error rates including rich subgroup fairness
- Comprehensive set of sample distortion metrics
- Generalized Entropy Index (Speicher et al., 2018)
- Differential Fairness and Bias Amplification (Foulds et al., 2018)
- Bias Scan with Multi-Dimensional Subset Scan (Zhang, Neill, 2017)

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/aif360.md

https://colab.research.google.com/drive/1jKt2AnmeEIIN3s0yyhoGxCX_0UhH8m0N?usp=sharing

You have just been hired as an **AI Ethicist**.
What would be your first **30-60-90** day plan?

Ethical AI is the development and use of AI systems in a manner that upholds **moral principles** (like fairness, transparency, privacy, and accountability) and aims to **benefit** society without causing **harm** or **discrimination**.

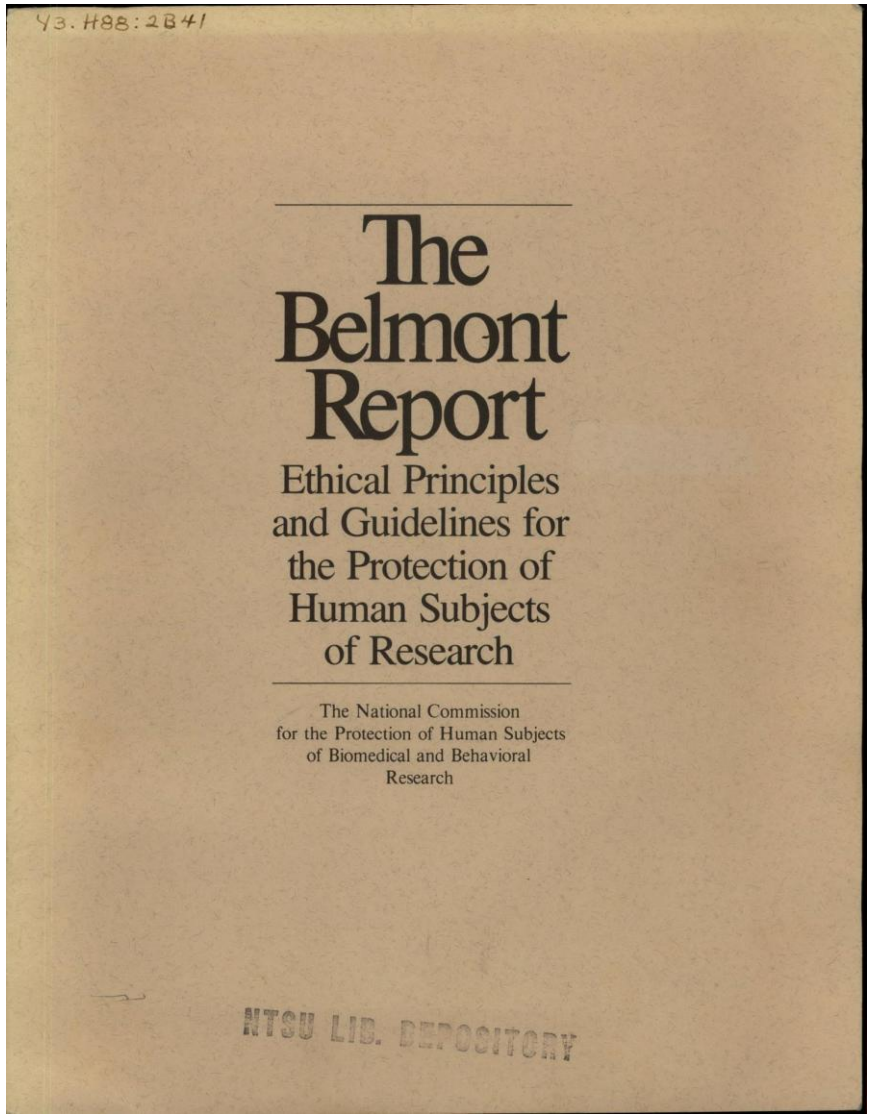
https://github.com/adnanmasood/cs-435-spring-2025/blob/main/ethical-ai.md

AI Ethics: Multidisciplinary & Socio-Technical field examining how to optimize AI benefits while reducing risks.

- Common Ethical Challenges:
- Data privacy and responsibility
- Fairness and bias
- Explainability and transparency
- Accountability and trust
- Environmental sustainability
- Misuse of technology

The Belmont Report's Three Principles

- Respect for Persons: Autonomy and consent, Protection of those with diminished autonomy
- Beneficence: Minimize potential harm, Maximize societal benefits
- Justice: Fair distribution of burdens and benefits

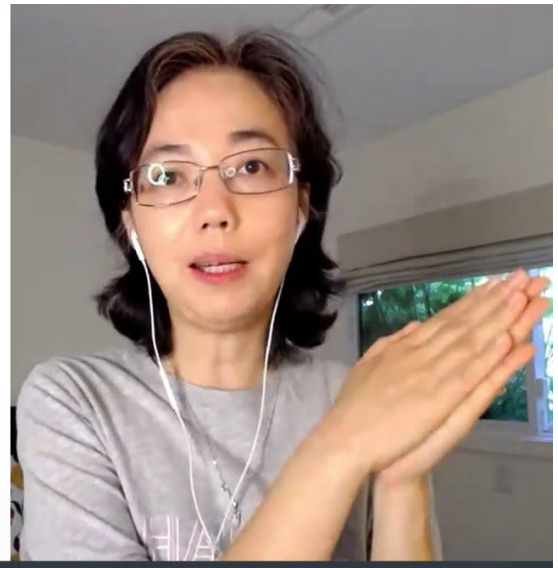


STANFORD AI & LAW SOCIETY

AI & HUMAN RIGHTS

AI & FREE SPEECH
AI & DISCRIMINATION
AI & WARFARE
AI & PERSONHOOD

Reinventing AI Safety & Reliability



Fei Fei Li
Stanford

greylock

HAI
Stanford University
Human-Centered
Artificial Intelligence

Dr. Fei Fei Lee

Human-Centered AI

CO-DIRECTOR // STANFORD INSTITUTE FOR HAI

Primary Concerns of AI Today

- **Foundation Models & Generative AI**
- Bias in training data
- Risk of generating false or harmful content
- Lack of explainability

- **Technological Singularity**
- AI surpassing human intelligence
- Questions about autonomy, liability, regulation

Primary Concerns of AI Today

- **AI Impact on Jobs**
  - Job shifts vs. absolute loss
  - Need for re-skilling, up-skilling

- **Privacy:**
- Data protection and security
- Regulatory context (GDPR, CCPA)
- Importance of informed consent

Primary Concerns of AI Today

- **Bias & Discrimination:**
  - Bias from training data
  - AI recruiting tools, facial recognition issues

- **Accountability:**
  - Lack of universal AI legislation
  - Ethical frameworks vs. enforceable regulations

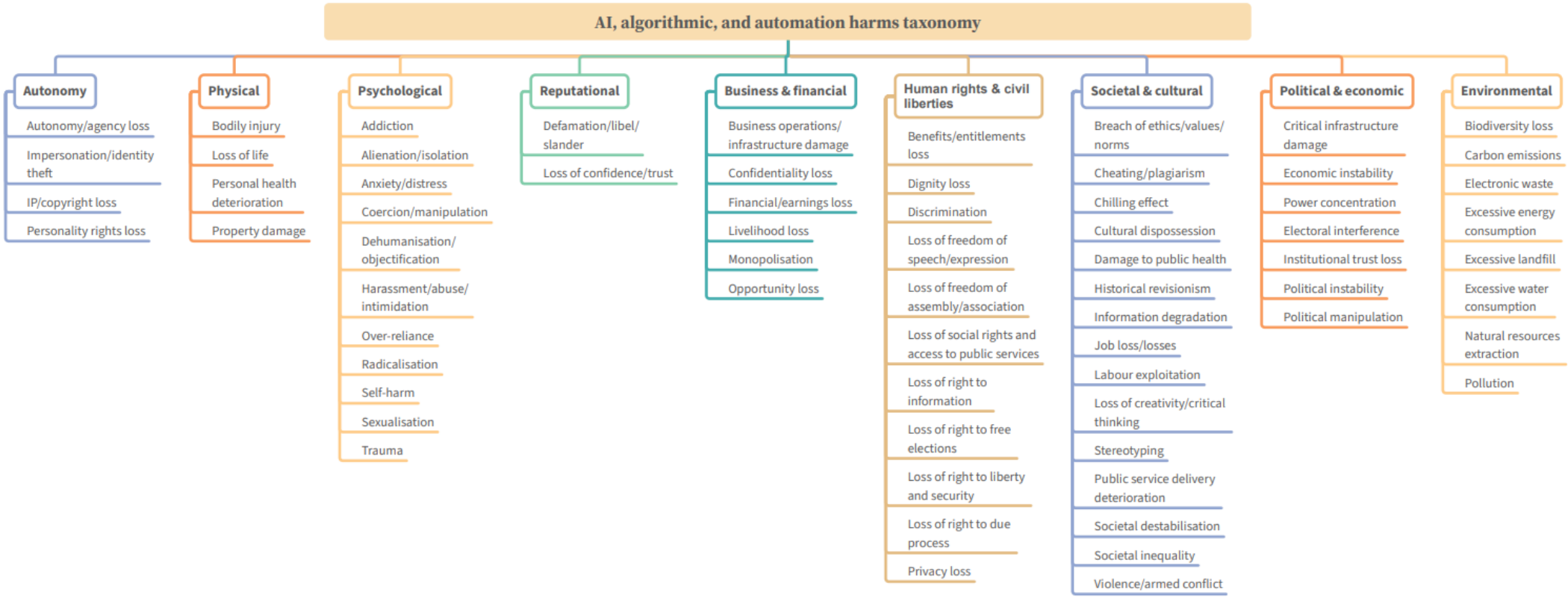


Fig. 3. An overview of the AI, algorithmic and automation harms taxonomy

https://arxiv.org/pdf/2407.01294

Approaches to AI Ethics & Governance

- **Governance Structures:**
- AI Ethics Board
- Clear roles, responsibilities, auditing

- **Principles**
- Fairness, Explainability, Robustness, Transparency, Privacy

- **Tools:**
- AI bias detection tools
- Model explainability frameworks

Case Study – IBM’s Principles of Trust & Transparency

AI augments human intelligence

Data belongs to the creator

AI should be transparent

Five Pillars:

Explainability,
Fairness,
Robustness,
Transparency,
Privacy

Example Governance: IBM’s AI Ethics Board

- Composition: Diverse leaders, ethicists, technical experts
- Responsibilities:
- - Oversee AI policies
- - Review AI projects
- - Provide decision-making
- Benefits: Clear accountability, ethical standards, transparency

Implementing Ethical AI in Practice

- Define Ethical Goals
- Develop Internal Policies
- Adopt Bias Detection & Explainability Tools
- Perform Ethical Risk Assessments
- Monitor & Audit AI Systems
- Govern via Boards & Compliance

Discussion & Future Outlook

- Upcoming Regulations:
- - EU AI Act, U.S. federal AI laws
- Emerging Concerns:
- - Deepfakes, social media manipulation
- - AI environmental impact
- Opportunities:
- - Public trust
- - Inclusive AI innovations
- - Sustainable AI solutions

Key Takeaways

AI Ethics ensures trust, reduces risk, improves societal outcomes

Multiple stakeholders involved

Governance frameworks guide responsible AI

Tools: Bias detection, model explainability, ethics boards

Continuous learning & adaptation

Discussion Prompts

- What are some strategies to reduce bias in AI?
- What role do you see future legislation playing in AI governance?
- How can we ensure AI benefits society broadly rather than a select few?

Recap Quiz

Ethical AI is the development and use of AI systems in a manner that upholds **moral principles** (like fairness, transparency, privacy, and accountability) and aims to **benefit** society without causing **harm** or **discrimination**.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/ethical-ai.md

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/On-the-Dangers-of-Stochastic-Parrots.md

- **Suppressing Ethical Concerns:** Alleged that Google leadership sidelined or suppressed research findings highlighting potential harms of large AI models.
- **Conflict with Corporate Interests**: Claimed Google prioritized business goals over ethical commitments, dismissing warnings about bias and societal impact.
- **Lack of Transparency:** Criticized Google for insufficient openness around data collection and model development processes.
- **Marginalization of Ethical AI Team:** Argued that Google undermined the autonomy of the Ethical AI team, discouraging critical inquiry.
- **Retaliatory Response:** Indicated that her forced departure was partly due to internal disagreements over publishing the “Stochastic Parrots” paper.

Data Curation Challenges

The paper highlights the impracticality of truly vetting massive text corpora, emphasizing how uncurated data can embed harmful stereotypes and reinforce biases (Sections 4 & 5).

Environmental Cost Quantification

It provides concrete estimates of the **energy consumption and CO2 emissions** linked to large-scale model training, stressing the outsized ecological footprint of ever-growing models (Section 2).

Illusion of Understanding

The authors discuss how large language models function as “stochastic parrots”—they predict words based on patterns in data without genuine comprehension, risking misleading outputs and disinformation (Section 5).

Ethical and Societal Risks

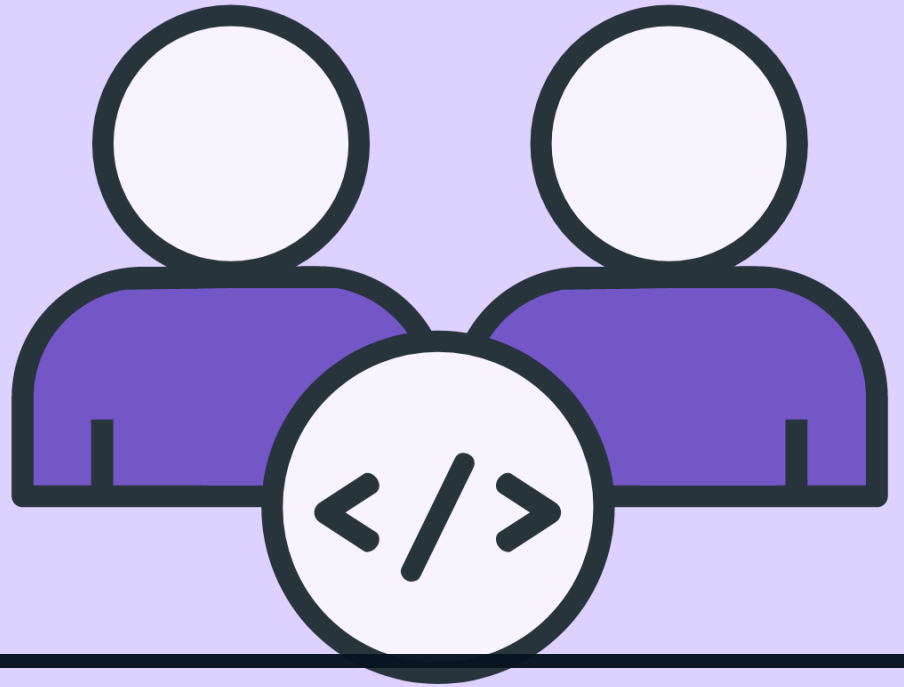
They detail how scale alone cannot solve issues of fairness or accountability, urging AI developers to address systemic bias, misuse potential, and real-world harm rather than relying on bigger datasets or more parameters.

Runaway Development Concern

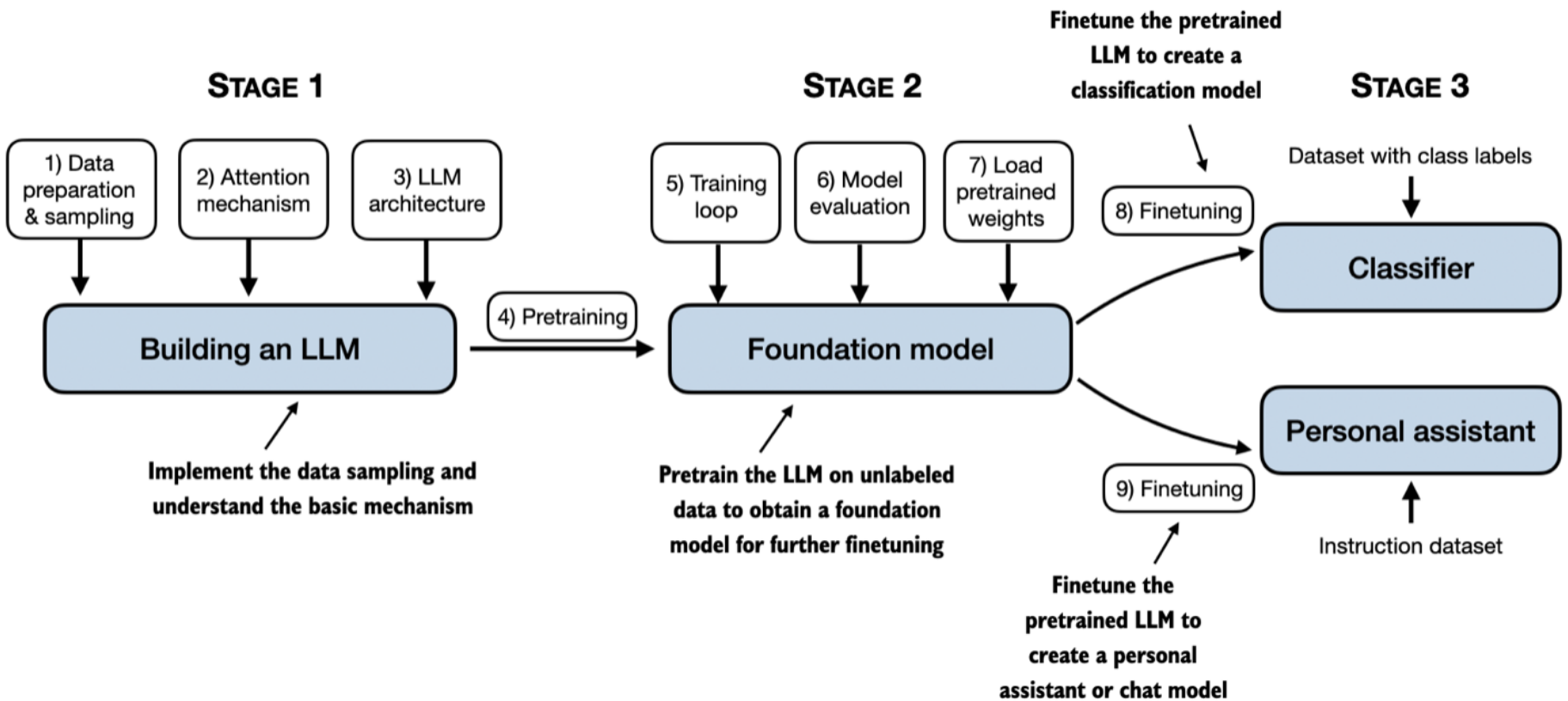
The paper warns about the “runaway train” phenomenon: racing to build increasingly bigger models without adequate governance, interdisciplinary oversight, or consideration of potential societal harms (Introduction & Discussion).

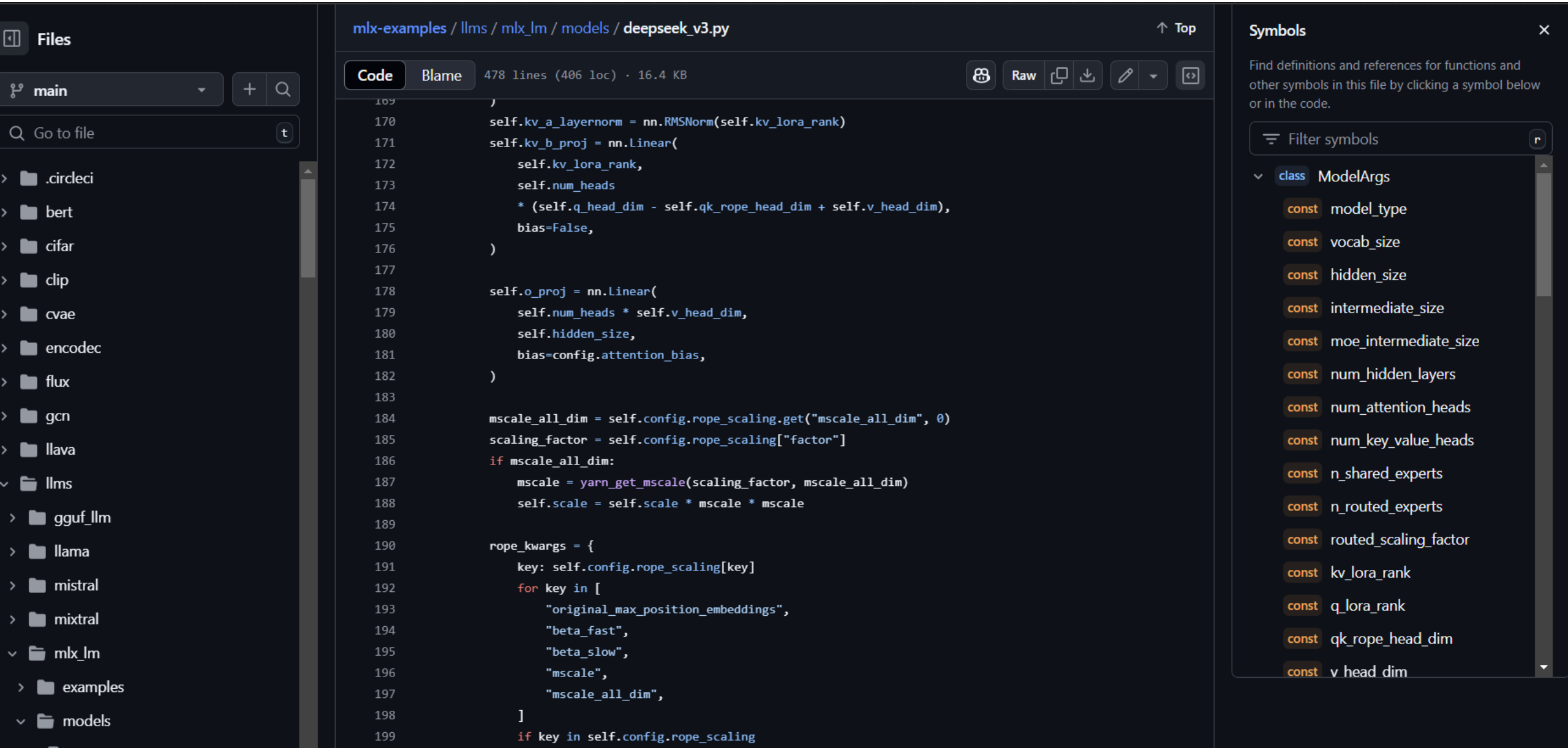

Talk is cheap.
Show me the code.
Linus Torvalds
quotefancy

What is Remote Pair Programming and Why Does it Work?

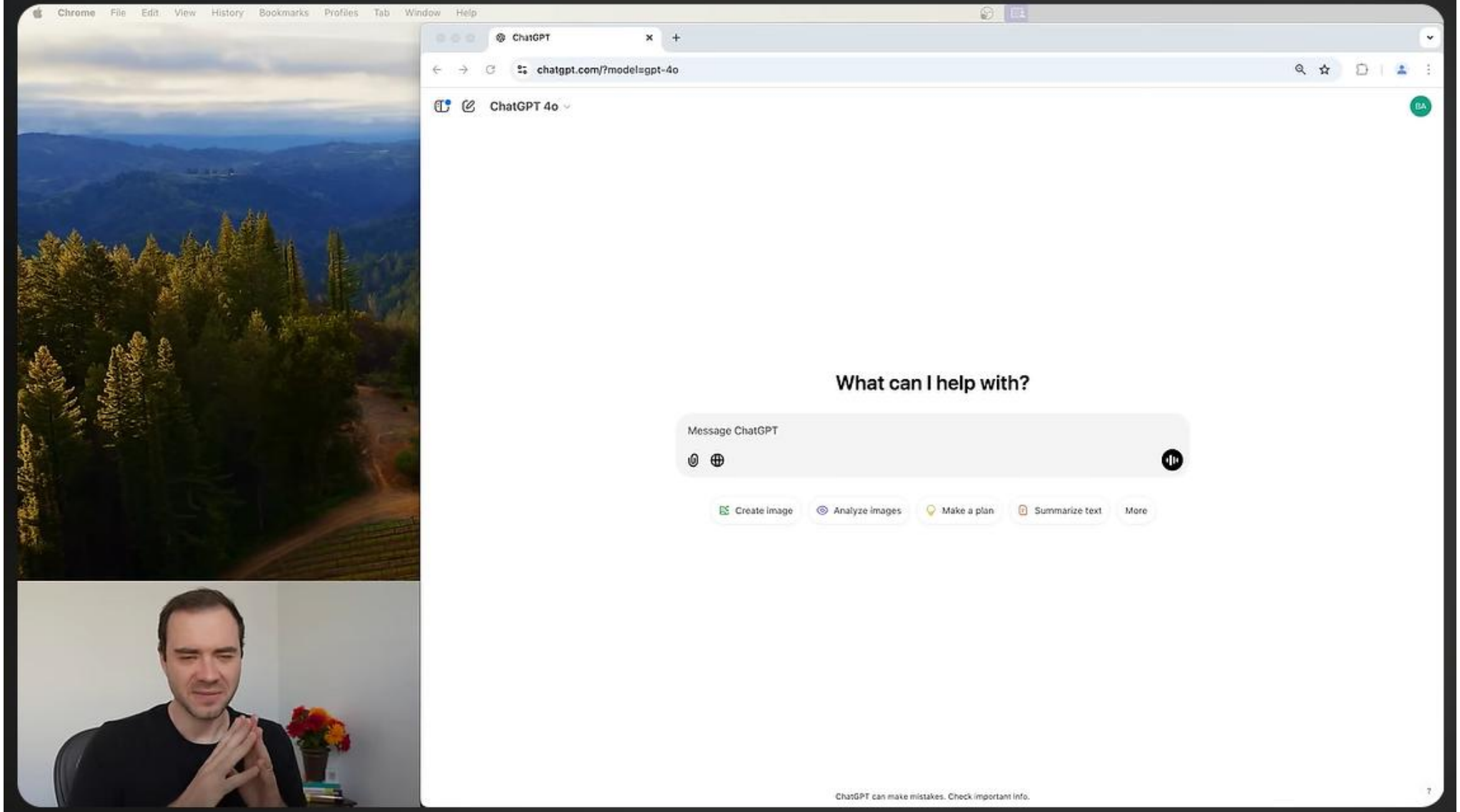


Build a Large Language Model from Scratch





https://github.com/ml-explore/mlx-examples/blob/main/llms/mlx_lm/models/deepseek_v3.py



Deep Dive into LLMs like ChatGPT

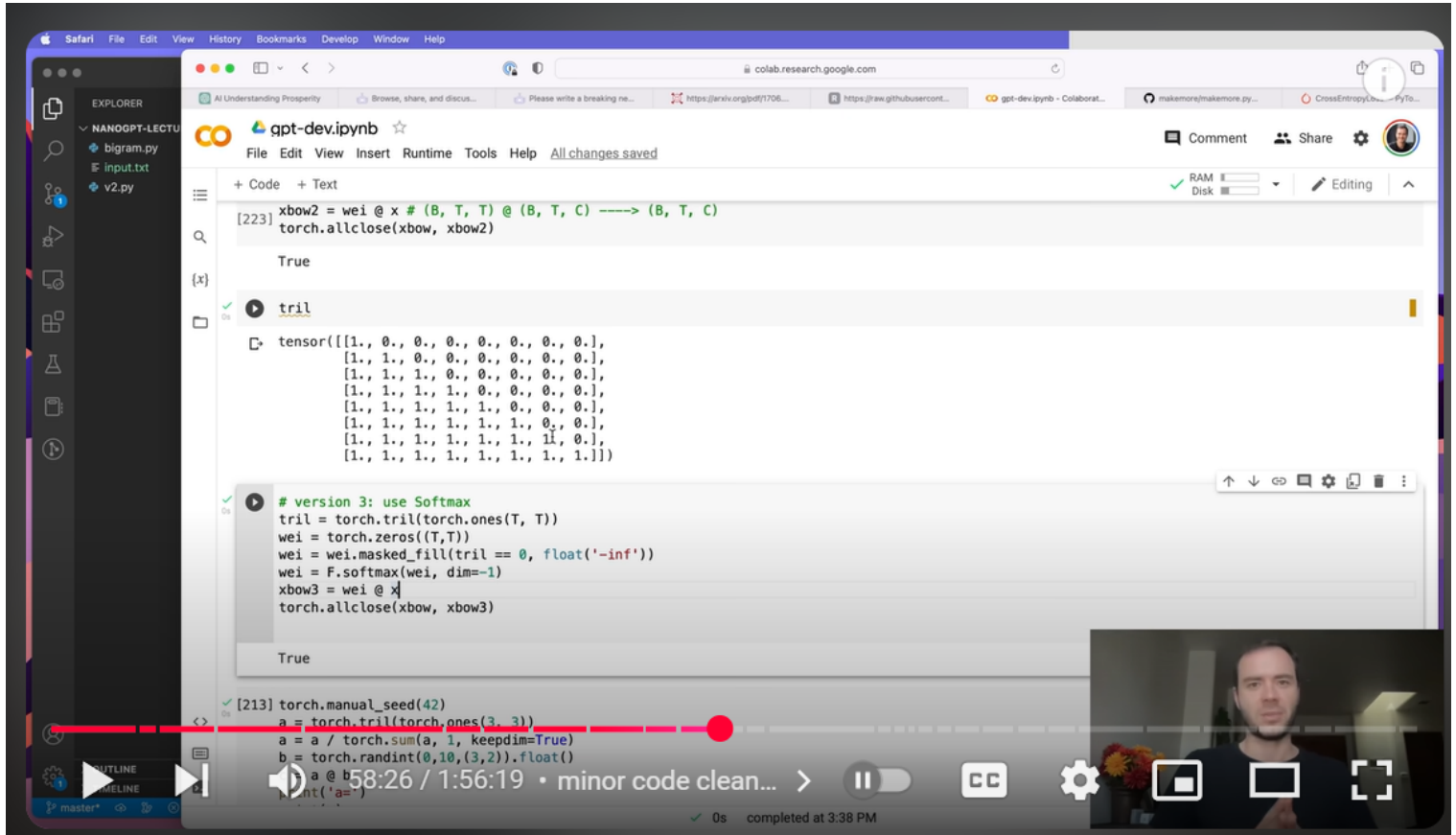
Andrej Karpathy

631K subscribers

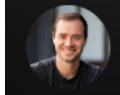
Subscribe

17K

Share



Let's build GPT: from scratch, in code, spelled out.



Andrej Karpathy
631K subscribers

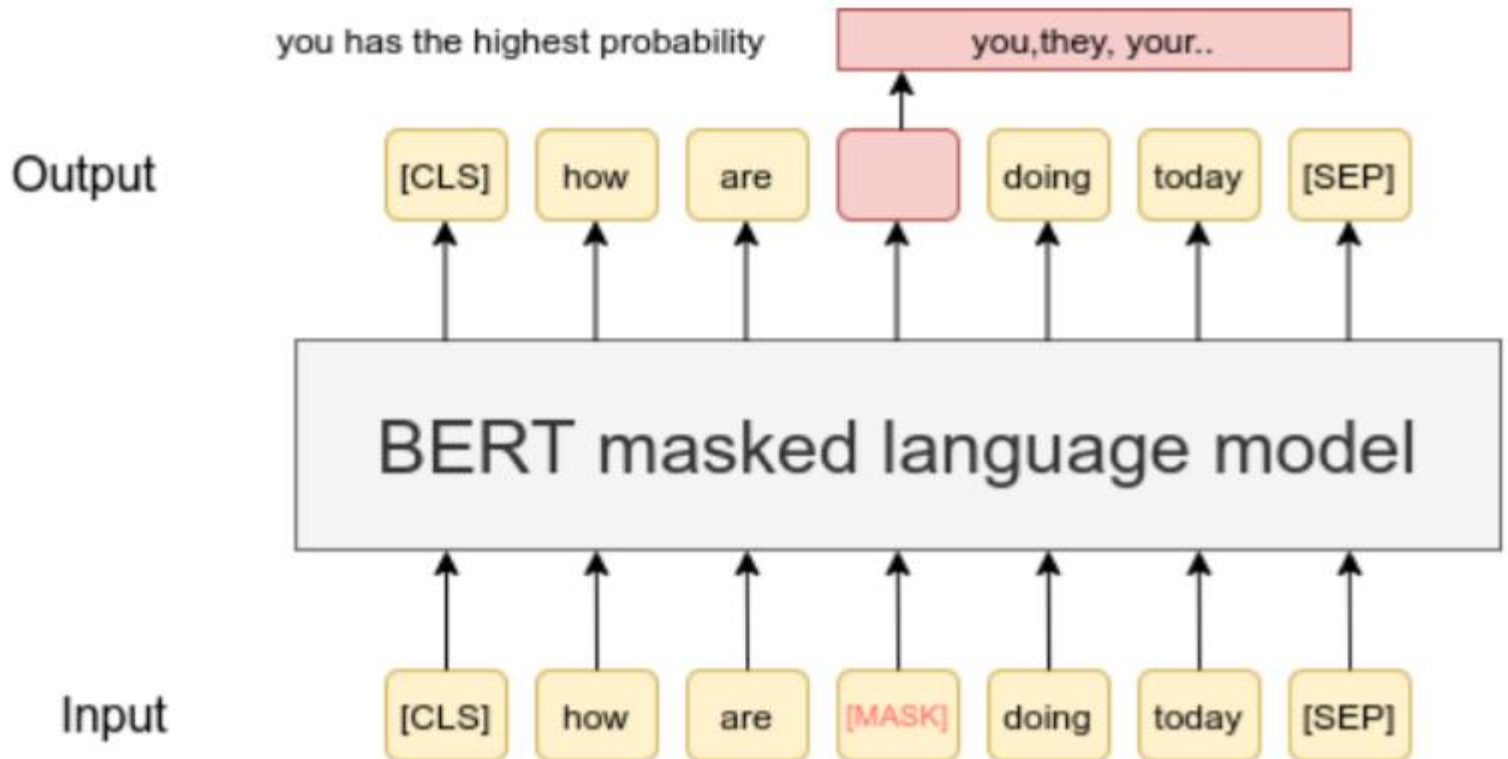
Subscribe

119K

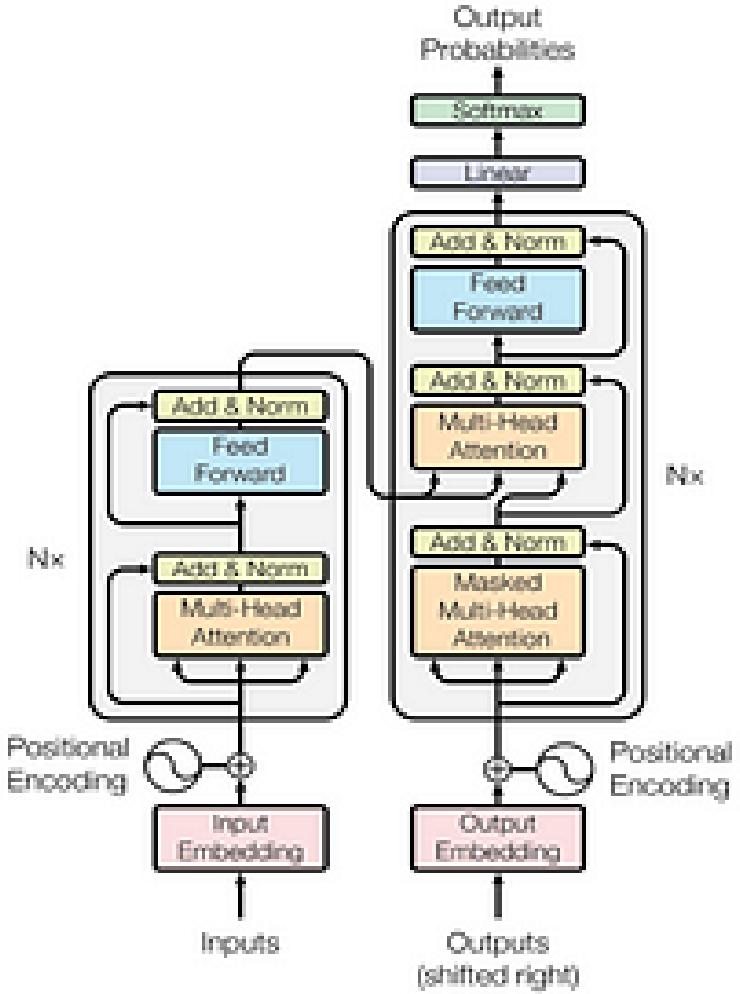
Share

5.1M views 2 years ago

We build a Generatively Pretrained Transformer (GPT), following the paper "Attention is All You Need" and OpenAI's GPT-2 / GPT-3. We talk about connections to ChatGPT, which has taken the world by storm. We watch GitHub Copilot, itself a GPT, help us write a GPT (meta :D!) . I recommend people watch the earl ...more

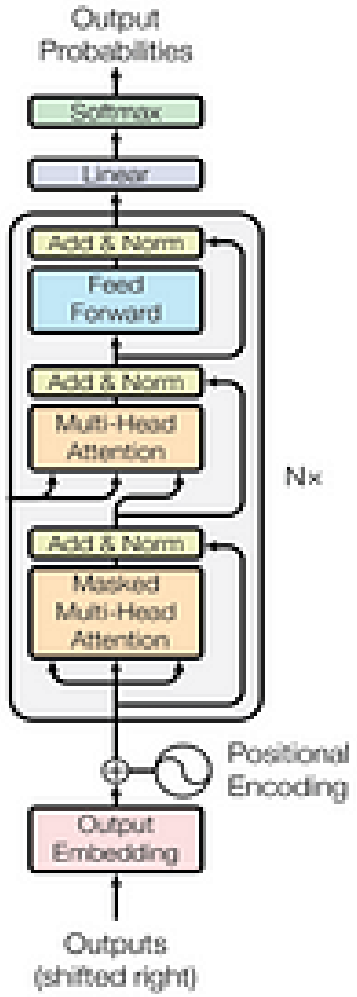

you has the highest probability
you,they, your..
Output
[CLS]
how
are
doing
today
[SEP]
BERT masked language model
Input
[CLS]
how
are
[MASK]
doing
today
[SEP]

Transformer



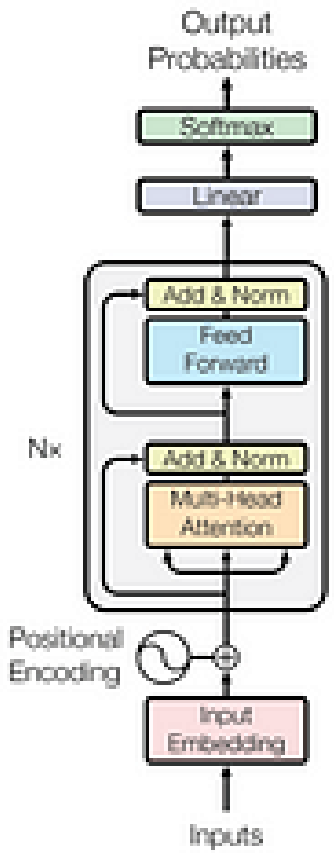
Encoder **Decoder**

GPT*



Decoder-only

BERT*



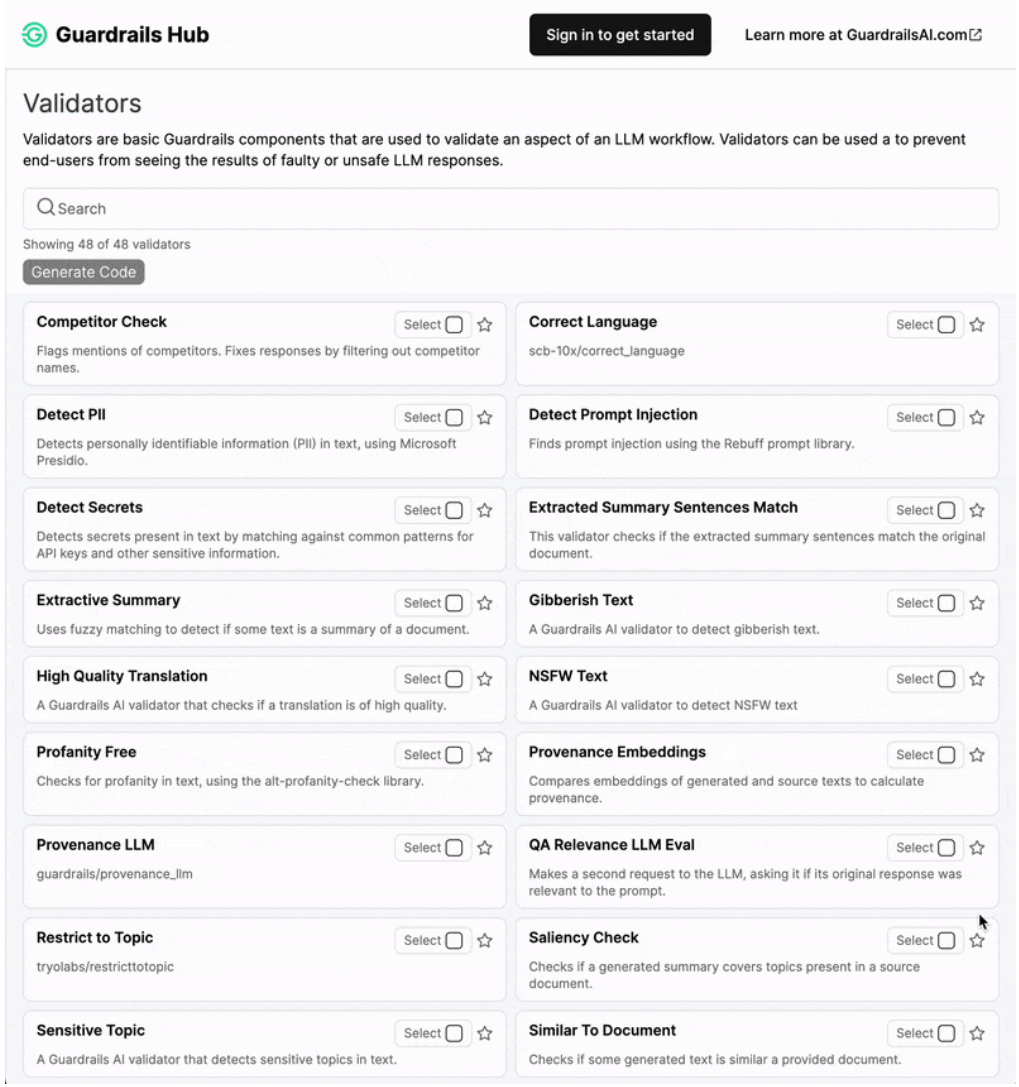
Encoder-only

Assignment

This assignment has four components. Each carry equal points. Submit the resulting PDF. Complete details are available on each component's github page.

a. **Implement Guardrails AI**
b. **OWASP top 10 Problem**
c. **Gen AI Job Review**
d. **Investigating harmful AI news**

Guardrails AI



https://github.com/adnanmasood/cs-435-spring-2025/blob/main/Exploring-Guardrails-AI-Validators.md

Assignment: Exploring Guardrails AI Validators

Guardrails AI Validator Hub https://hub.guardrailsai.com/ Guardrails AI github codebase https://github.com/guardrails-ai/guardrails

Objective

1. **Learn & Explore**: Familiarize yourself with the Guardrails AI framework and the variety of **validators** offered in the Guardrails Hub.
2. **Implement a Validator**: Install and use **one validator** of your choice in a Jupyter notebook to demonstrate how it works.
3. **Document & Demonstrate**: Provide detailed explanations (via comments and markdown cells) of what your code does and why it is relevant to AI/LLM risk mitigation.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/Exploring-Guardrails-AI-Validators.md

Assignment: OWASP Top Ten for LLM (2025)

Objective

1. **Learn & Research**: Familiarize yourself with the **OWASP Top Ten for Large Language Models (LLM) 2025**.
2. **Select an Item**: Choose **one** issue from the Top Ten that interests you or seems especially significant.
3. **Real-World Impact**: Investigate how that issue could (or does) affect an actual enterprise or project.
4. **Risk Mitigation**: Propose one or two strategies to address or mitigate this specific risk.

OWASP top 10 Link https://owasp.org/www-project-top-10-for-large-language-model-applications/

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/assignment-owasp-top-10.md

Assignment: AI Career Exploration and Skill Gap Analysis

Objective

1. **Real-World Relevance**
   Explore a **current, real-world job listing** in the fields of **generative AI, AI safety, AI security, or AI ethics/governance** that aligns with your interests and recent coursework.
2. **Skills and Knowledge Assessment**
   Analyze the job description to understand the required skills, knowledge, and responsibilities.
3. **Gap Identification**
   Reflect on which job requirements you already meet and which ones you do not.
4. **Action Plan for Skill Development**
   Propose a clear, step-by-step plan to **acquire or strengthen** the missing skills or knowledge areas.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/gen-ai-job.md

Assignment: Investigating Harmful AI News

Objective

1. **Real-World Analysis**: Identify a real incident where AI (or Generative AI) caused harm or raised significant ethical/safety/security concerns.
2. **Critical Examination**: Summarize what happened, who was affected, and how.
3. **Preventative Measures**: Propose strategies that could have prevented the harm and outline measures to avoid similar incidents in the future.
4. **Presentation**: Prepare a short (2-minute) in-class presentation summarizing your findings and recommendations.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/Investigating-Harmful-AI-News.md