

CS 194/294-267 Understanding Large Language Models: Foundations and Safety

https://rdi.berkeley.edu/understanding_llms/s24

Teaching Staff

Instructor: **Prof. Dawn Song**

Co-instructor: **Dr. Dan Hendrycks**

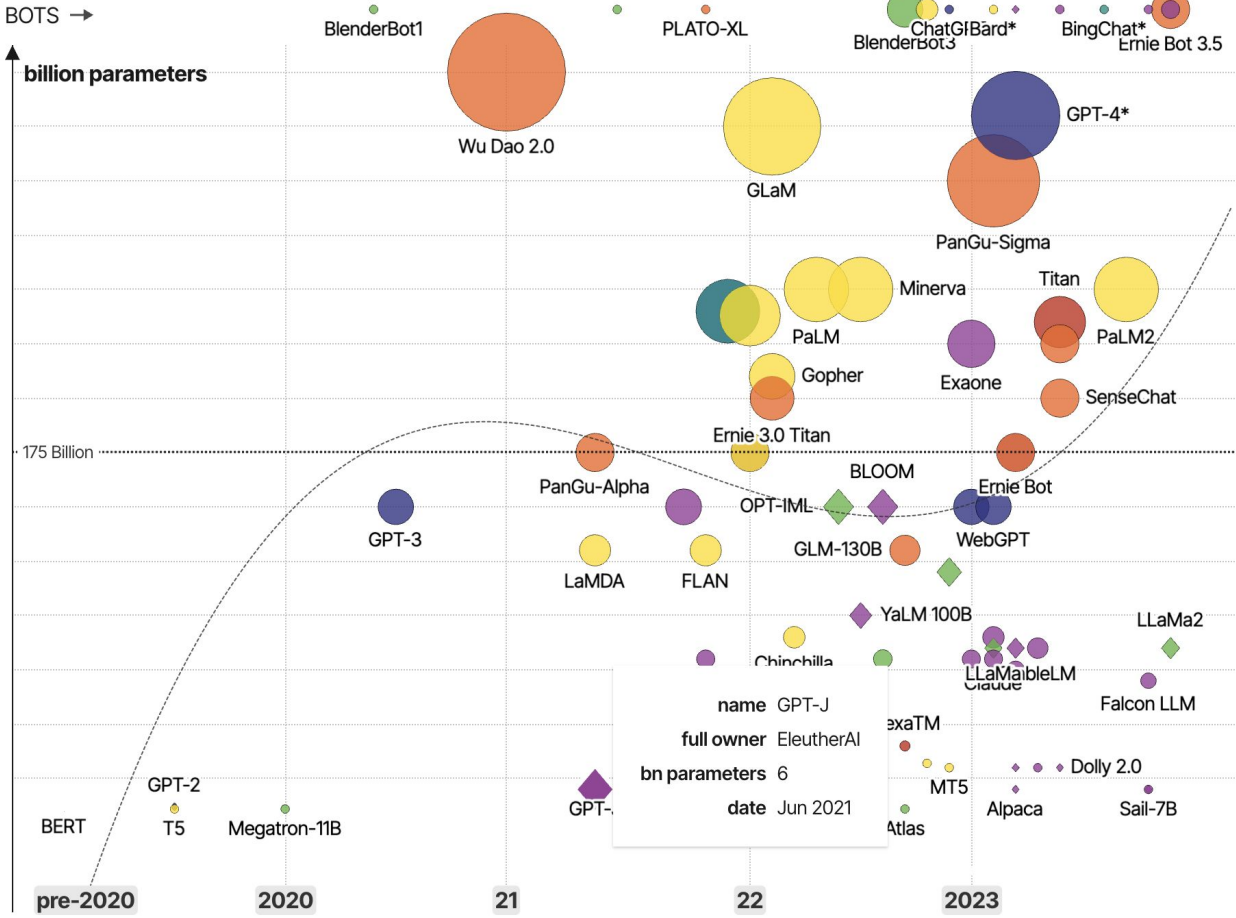
GSI: Yu Gai

Reader: Tara Pande

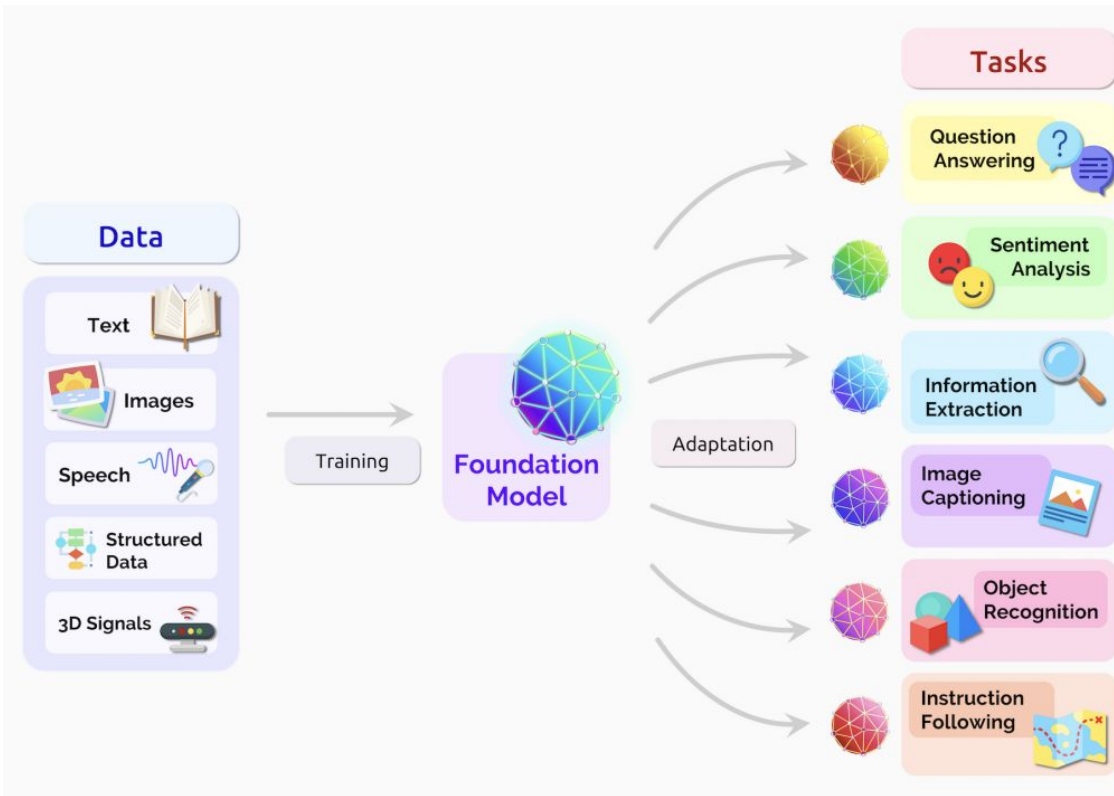
Exponential Growth in LLMs

Large Language Models (LLMs) & their associated bots like ChatGPT

● Amazon-owned ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other



Powering Rich New Capabilities

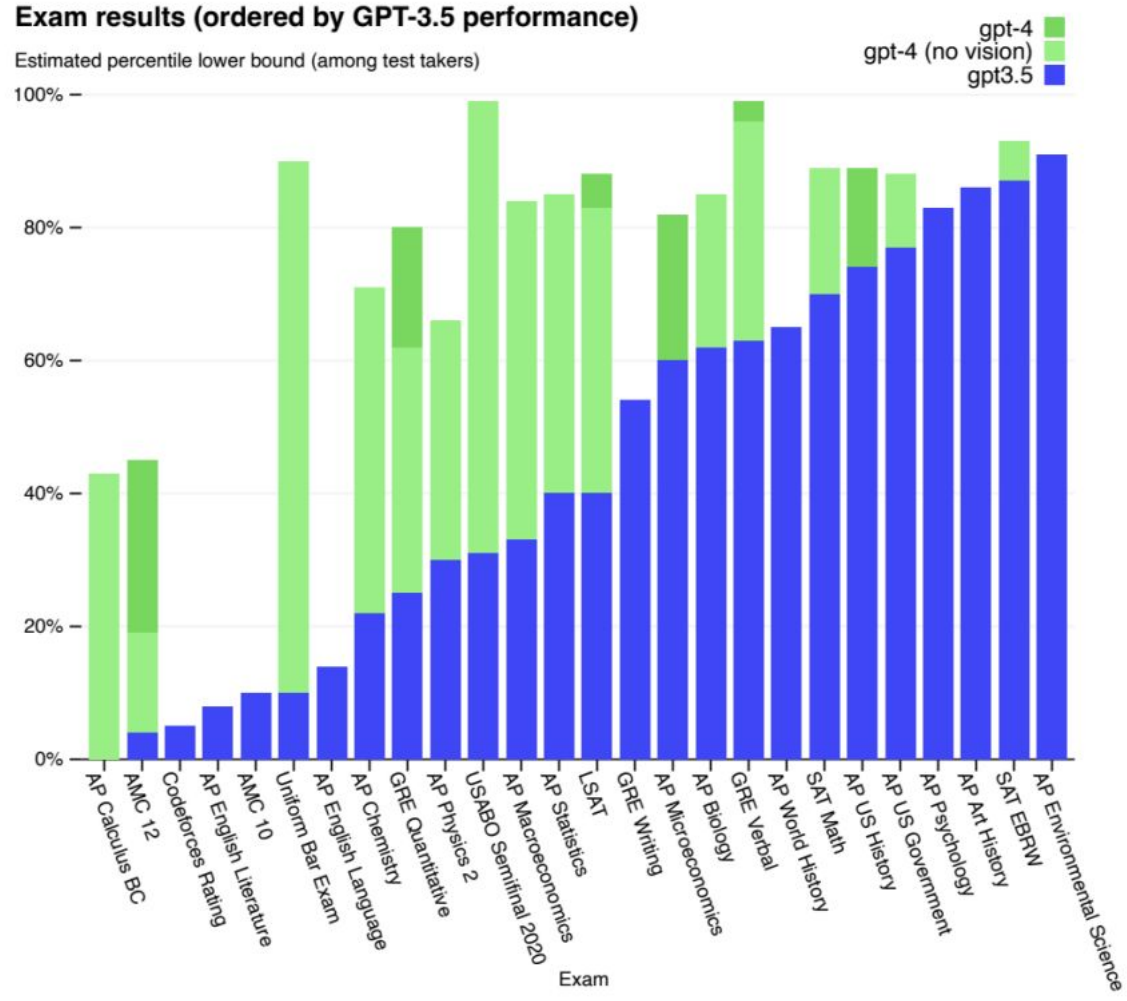


<https://arxiv.org/pdf/2108.07258.pdf>

- | | | | |
|--|--|---|---|
| Q&A
Answer questions based on existing knowle... | Grammar correction
Corrects sentences into standard English. | Spreadsheet creator
Create spreadsheets of various kinds of dat... | JavaScript helper chatbot
Message-style bot that answers JavaScript ... |
| Summarize for a 2nd grader
Translates difficult text into simpler concep... | Natural language to OpenAI API
Create code to call to the OpenAI API usin... | ML/AI language model tutor
Bot that answers questions about language... | Science fiction book list maker
Create a list of items for a given topic. |
| Text to command
Translate text into programmatic commands. | English to other languages
Translates English text into French, Spanish... | Tweet classifier
Basic sentiment detection for a piece of text. | Airport code extractor
Extract airport codes from text. |
| Natural language to Stripe API
Create code to call the Stripe API using nat... | SQL translate
Translate natural language to SQL queries. | SQL request
Create simple SQL queries. | Extract contact information
Extract contact information from a block of ... |
| Parse unstructured data
Create tables from long form text | Classification
Classify items into categories via example. | JavaScript to Python
Convert simple JavaScript expressions into ... | Friend chat
Emulate a text message conversation. |
| Python to natural language
Explain a piece of Python code in human un... | Movie to Emoji
Convert movie titles into emoji. | Mood to color
Turn a text description into a color. | Write a Python docstring
An example of how to create a docstring for ... |
| Calculate Time Complexity
Find the time complexity of a function. | Translate programming languages
Translate from one programming language ... | Analogy maker
Create analogies. Modified from a communi... | JavaScript one line function
Turn a JavaScript function into a one liner. |
| Advanced tweet classifier
Advanced sentiment detection for a piece o... | Explain code
Explain a complicated piece of code. | Micro horror story creator
Creates two to three sentence short horror ... | Third-person converter
Converts first-person POV to the third-pers... |
| Keywords
Extract keywords from a block of text. | Factual answering
Guide the model towards factual answering ... | Notes to summary
Turn meeting notes into a summary. | VR fitness idea generator
Create ideas for fitness and virtual reality g... |
| Ad from product description
Turn a product description into ad copy. | Product name generator
Create product names from examples word... | ESRB rating
Categorize text based upon ESRB ratings. | Essay outline
Generate an outline for a research topic. |
| TL;DR summarization
Summarize text by adding a 'tl;dr:' to the en... | Python bug fixer
Find and fix bugs in source code. | Recipe creator (eat at your own risk)
Create a recipe from a list of ingredients. | Chat
Open ended conversation with an AI assist... |

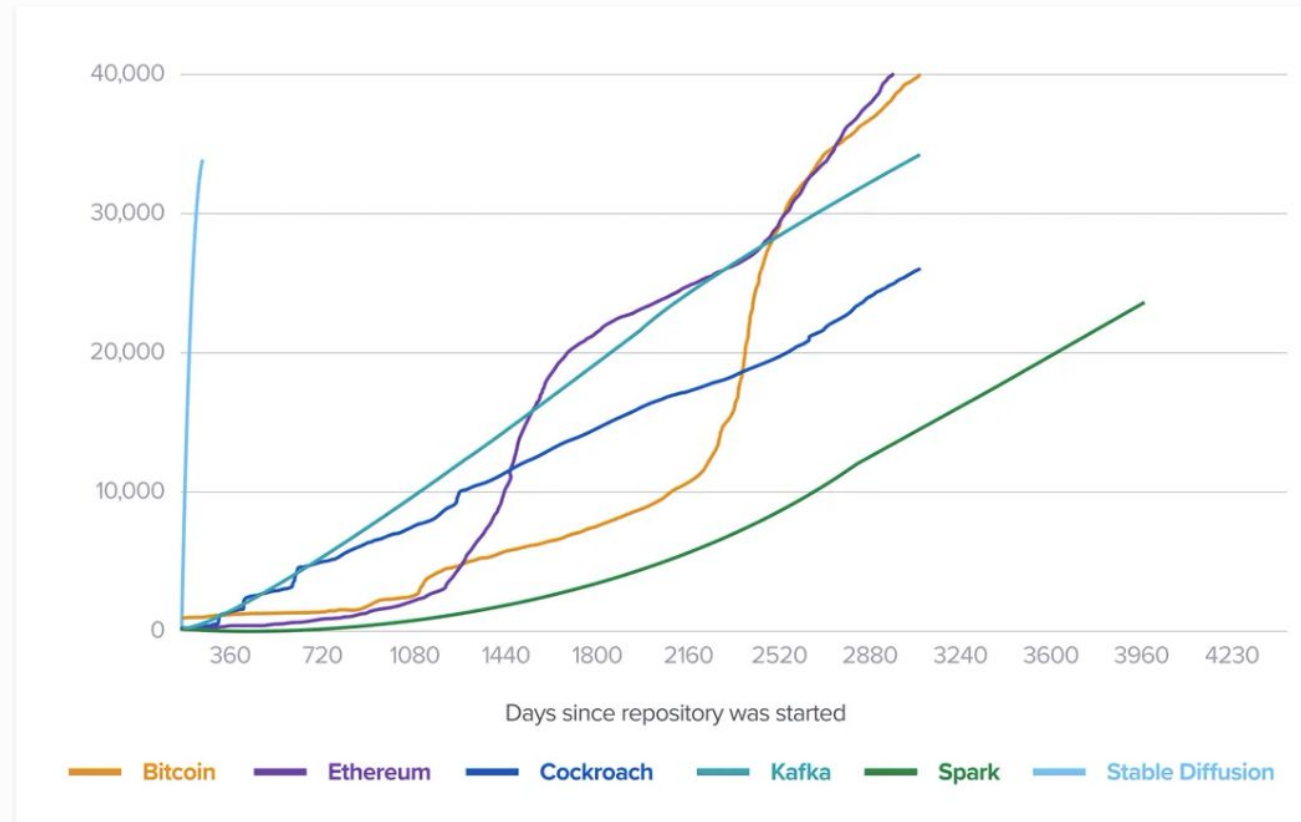
Source: openai

GPT-4 Excels in Standard Exams over Diverse Topics



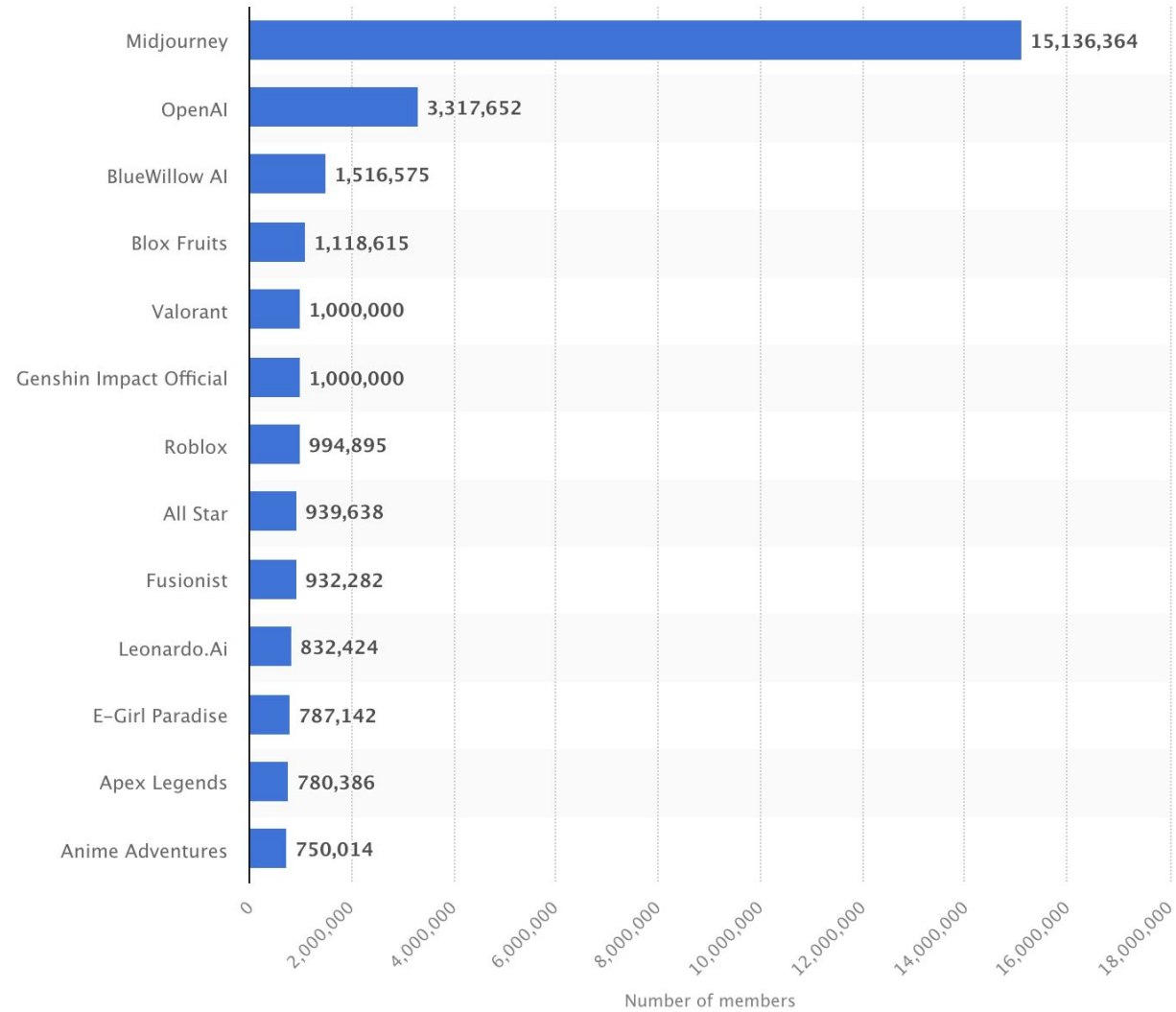
Stable Diffusion: Fastest Repo to Reach 35K Stars on GitHub

Stable Diffusion Developer Adoption

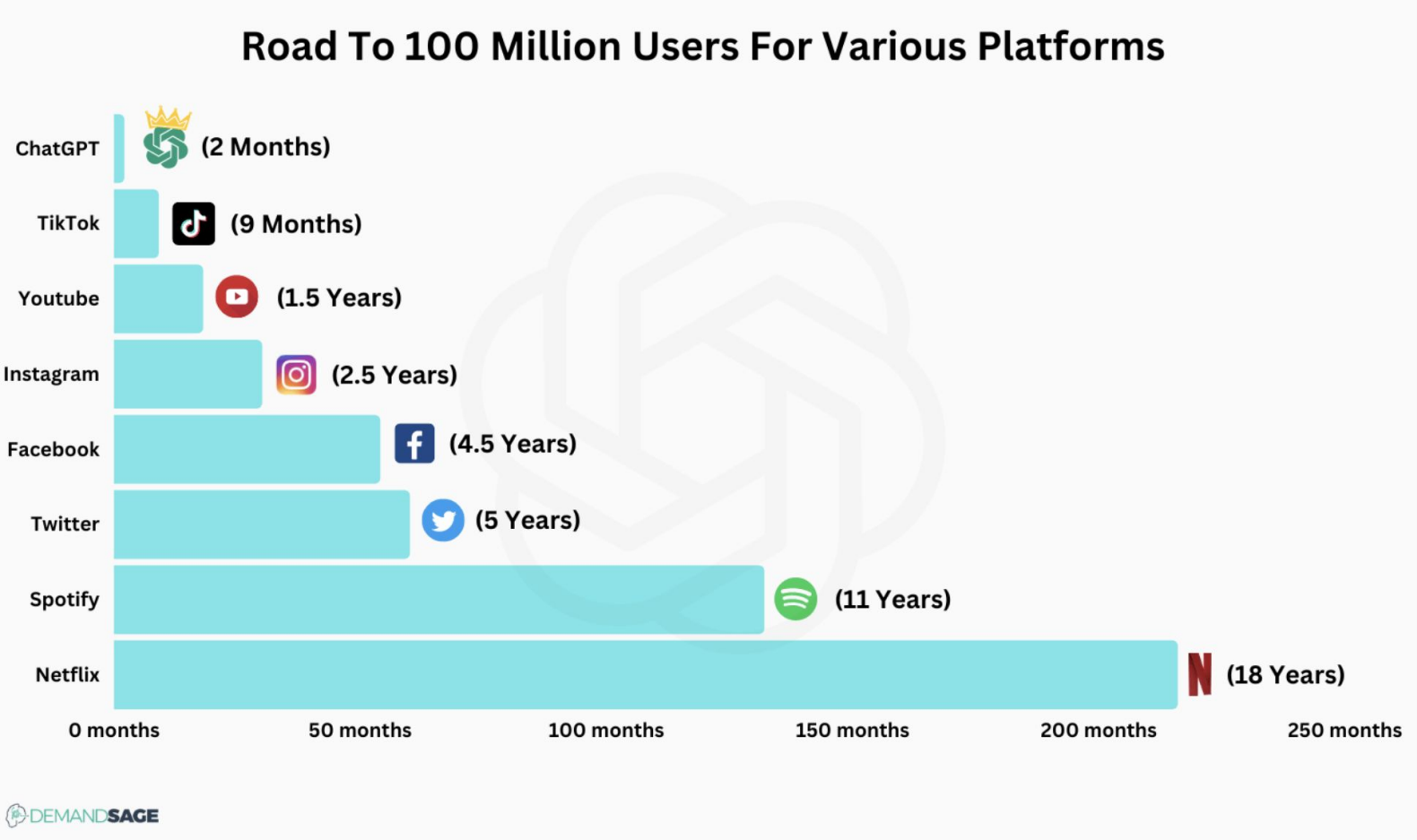


Source: GitHub

Midjourney: The Largest Discord

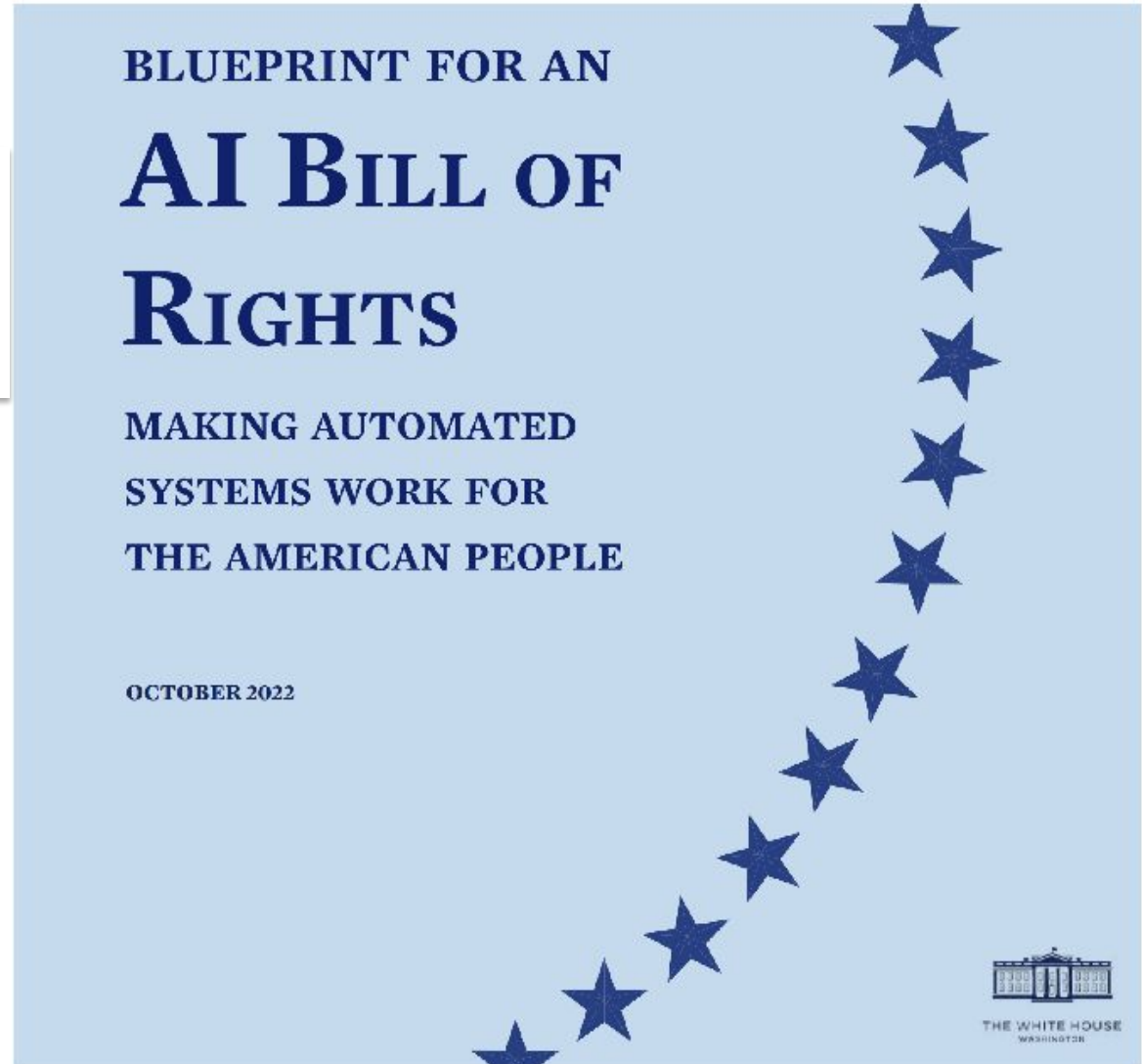


ChatGPT: Fastest Platform to 100M Users (Before Thread)



Importance of Responsible AI

- Robustness: Safe and Effective Systems
- Fairness: Algorithmic Discrimination Protections
- Data Privacy
- Notice and Explanation
- Human Alternatives, Consideration, and Fallback



White House Executive Order

(k) The term “dual-use foundation model” means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

(ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

(iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

Many Risks & Open Challenges for Responsible AI

- Who controls AI?
 - centralized vs. decentralized control; open vs. closed source
- Trustworthiness
 - Robustness
 - Adversarial robustness
 - Out-of-distribution robustness
 - Test-time attacks vs. training-time attacks
 - Privacy
 - Fairness
 - Toxicity
 - Stereotype
 - Machine ethics
- AI Safety
 - Misuse/abuse of AI
 - Super intelligence

Importance of Mitigating Risk of Extinction from AI

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.



The New York Times

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

Signatories:

AI Scientists Other Notable Figures

Geoffrey Hinton

Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio

Professor of Computer Science, U. Montreal / Mila

Demis Hassabis

CEO, Google DeepMind

Sam Altman

CEO, OpenAI

Dario Amodei

CEO, Anthropic

Dawn Song

Professor of Computer Science, UC Berkeley

Ted Lieu

Congressman, US House of Representatives

Bill Gates

Gates Ventures

Ya-Qin Zhang

Professor and Dean, AIR, Tsinghua University

Ilya Sutskever

Co-Founder and Chief Scientist, OpenAI

Shane Legg

Chief AGI Scientist and Co-Founder, Google DeepMind

Unique Aspects of AI

- AI capability already exceeds human-level performance on many tasks and progresses extremely fast
- Humans are highly incentivized to continue develop & enhance AI capabilities
- AI capability is extremely general, widely applicable to almost all areas
- AI agents interact directly with the world autonomously
- We have little understanding of how deep learning system works
- AI systems can create new capabilities that were not designed in and improve on its own
- AI capability can be easily misused
- **AI safety is different & much more challenging than safety for nuclear & bio tech**

Topics covered in this Course

- Understanding:
 - Foundations of LLMs
 - Interpretability
 - Scaling laws
 - Reasoning and mathematics
 - Agency and emergence
 - Evaluation and benchmarking
- AI Safety:
 - AI alignment and governance
 - Adversarial robustness
 - Trojans
 - Privacy, unlearning

Tasks for Different Units

In-person lecture participation + reading summaries & questions for Q&A
(due by 2pm before the day of lecture)

+

1 unit: article about the topic of a lecture (at least 2 pages)

2 units: lab + project (implementation not required)

3 units: lab + project with implementation

4 units: lab + project with significant implementation and end-to-end demo

(Groups of 4 students required for 2-4 units projects)

Grading

	1 unit	2 units	3/4 units
Lecture participation	25%	10%	10%
Reading summaries & questions for Q&A	25%	10%	10%
Article	50%		
Lab		20%	10%
Project			
- <i>Proposal</i>		10%	10%
- <i>Milestone</i>		10%	10%
- <i>Presentation</i>		20%	15%
- <i>Report</i>		20%	15%
- <i>Implementation</i>			20%

Lab and Project Timeline

	Released	Due
Project group formation	Jan 19	Jan 30
Project proposal	Jan 23	Feb 13
Lab	Jan 30	Feb 27
Project milestone	Feb 13	Mar 19
Project presentation	Mar 19	Apr 16
Project final report	Mar 19	Apr 30

Sample Project Ideas

- Developing new methods for interpretability
- Developing new frameworks for evaluating safety properties, including robustness, monitoring, controllability, etc.
- New methods for mitigating biases in LLMs
- New methods for unlearning hazardous or undesirable capabilities
- New methods for jailbreaking, adversarial attacks on LLMs & mitigations