Stat 9911
Principles of AI: LLMs

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

January 15, 2025

Penn
University of Pennsylvania

Wharton
University of Pennsylvania

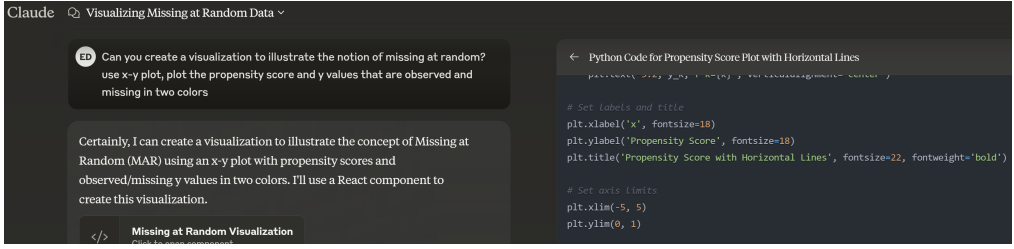
Table of Contents

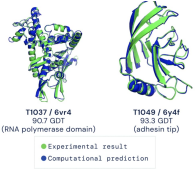
Motivation

Getting Started

The Age of AI

- Artificial Intelligence (AI) is emerging as a world-changing technology
- Success stories: Large language models (ChatGPT, Claude), Image generation (Dall-E, Stable Diffusion), Protein structure prediction ...





AI is Trending

- ChatGPT has 100+ million weekly active users
- Nvidia is one of the world's most valuable publicly traded companies
- AI starting to be used in products: Code assistants (Copilot), Customer service, Web search
- Some jobs replaced, other high-value jobs created
- Several papers in Nature/Science reporting discoveries using AI

Article

Scaling deep learning for materials discovery

https://doi.org/10.1038/s41586-023-06735-9

Received: 8 May 2023

Accepted: 10 October 2023

Published online: 29 November 2023

Amil Merchant[1,3], Simon Batzner[1,3], Samuel S. Schoenholz[1,3], Muratahan Aykol[1], Gowoon Cheon[2] & Ekin Dogus Cubuk[1,3]

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing[1–11]. From microchips to

Fast Progress

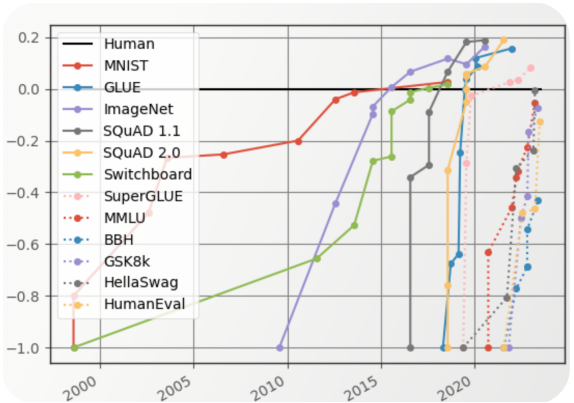
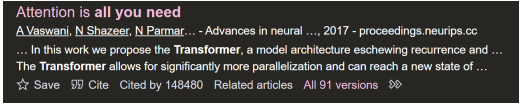


Figure: AI achieving superhuman levels on various benchmarks.

- (Wharton OID Prof) Ethan Mollick's "Otter on a plane using Wifi" experiments with video generators: Sora: Dec 16, 2024; veo 2: Jan 10, 2025

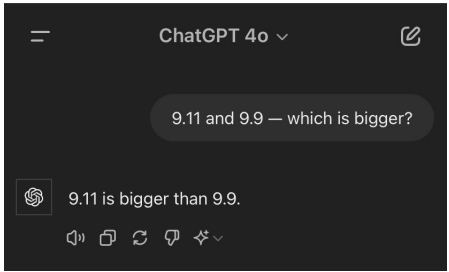
Incredible Amount of Work in the Area

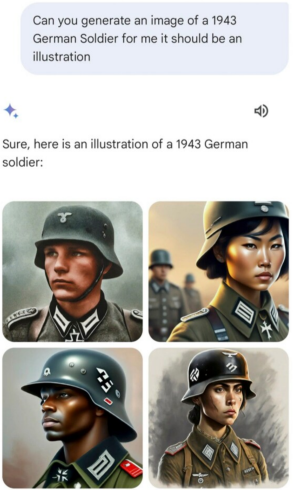


- $\sim$150,000 papers in just 7-8 years on LLMs. Several interesting papers every day. Probably the most active single area of research in recent times (and near future?)
- Lots of new models/problems/tools/products/frameworks/software/startups...
- Potential to benefit & become a leader; but also to fall behind and lose out
- Overwhelming! How to keep up?

Current AI Has Problems

- AI systems can be wrong, discriminatory, and racist





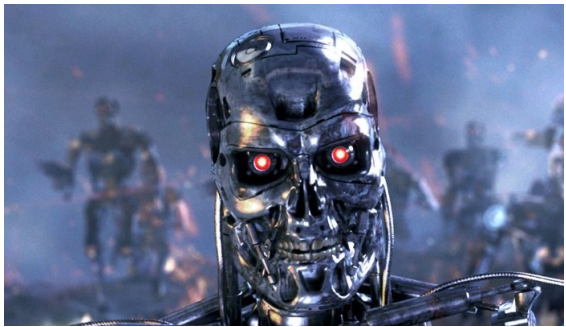
Current AI Has Problems

can you generate a picture of the room without an elephant in it?

Here is an image of a minimalist room interior without an elephant. Let me know if you need any adjustments or have further ideas!

Does AI Bring Risks?

- Does current and future AI increase the risk of misinformation, fraud, persuasion, etc? (Yes!)
- Does current and future AI have the potential to create new risk vectors and damage? (Yes! example: jailbreaking LLM-controlled robots)
- Can future AI pose an existential risk (x-Risk) to humanity? (???)



Goals of this course

- Focus on one crucial component of AI: language, and specfically, Large Language Models
- Take you from the basics to the cutting edge of ideas behind LLMs
- Cover all the most important/influential topics/concepts; Build familiarity so that you can follow and participate in ongoing discussion
- In certain areas, cover enough material that you can understand the significance and technical contributions of new research
- Focus on reusable principles and ideas that can be leveraged to solve multiple problems
- Identify immediate/future research directions; Ideally enable you to be able to come up with your own research ideas

Structure of this course

- Start with instructor lectures
- Later, student presentations. Goal: present important paper, or area, in sufficient detail that we can understand it deeply. Sign up for topics, can work in teams of two.
- Grading: presentation and attendance.
- Project instead of presentation? (Reach out to discuss if you want to do this)

Course Materials

- Lecture notes are the most comprehensive resource (125+ pages, extensive references)
- Some slides
- Lectures will have some slides, some external materials, whiteboard, maybe some code, etc
- Website has most of the class materials.

Relation to other courses

- Our course aims to be complementary to other LLM courses
- Broad, conceptual coverage; focus on ideas
- 9910: Stat in LLMs focuses on stats
- LLM courses in Comp Sci: some focus more on implementation and engineering, some based on external speakers, some focus on topics (e.g., LLM security)

Topics covered: Building AI

- What is AI?
- Can we build AI?
- Key role of language

Topics covered: Large Language Model Architecture

- Basic formulation: Next Token Prediction
- Attention Mechanism: Basics, Positional encoding, Insight
- Speeding up attention; long contexts
- Alternatives? State space models
- Empirical Behaviors: Emergence, Memorization
- Extensions: Vision Transformers, Vision Language Models, Multimodal Language Models
- Architectural Choices in Specific LLMs: GPT, Llama, DeepSeek, etc

Topics covered: Training LLMs

- Pre-training and post-training
- Supervised fine-tuning
- Learning a reward/Reward modeling
  - Learning a reward based on direct human evaluation
  - Learning a reward based on preference data/Preference modeling
- Using a learned reward
  - Rejection sampling+SFT based on the learned reward
  - RL Fine-tuning based on the learned reward
- Direct Preference Optimization

Topics covered: Training LLMs ctd.

- Generating synthetic data for training
- Tool use
- Special considerations and forms of fine-tuning
  - Parameter efficient fine-tuning
  - Prompt and Prefix-tuning
  - Self-play/improvement
  - Model compression

Topics covered: Inference/Test-time computation

- Simple Decoding/Sampling methods
- Prompting
- Reasoning
  - Using Rewards
  - Reasoning and action
- Knowledge Retrieval

Topics covered: Capabilities

- Reasoning and Math
  - RL
- Code
- Storage and Retrieval
- Medical Capabilities

Topics covered: Safety and Security

- Robustness & security
  - Jailbreaking
  - Oversight
- Hallucinations
- Uncertainty
  - Uncertainty Measures
  - Conformal prediction
  - Calibration

Other possible/smaller topics

- Evaluation: Datasets, metrics, principles
- Embeddings/Representations
- Systems and agents
- Mechanistic Interpretability: Probing, Circuits
- Living with AI & the future with AI

Table of Contents

Motivation

Getting Started

Let Us Get Started: Experience the Magic of LLMs!

- Try top models for free:
  - **Gemini**, e.g., Experimental 1206, 2.0 Flash Thinking in Google AI Studio
  - **OpenAI's Models**, e.g., GPT-4o mini, available at chat.openai.com
  - **Claude**, e.g., 3.5 Sonnet at Claude.ai
- Capabilities include:
  - Solving homework problems ...
  - And some more advanced problems: o1-pro solved problem B4 on the Putnam 2024. See example solution and discussion here.
  - Teaching advanced topics faster than traditional resources (e.g., Quantum Field Theory)
  - Explaining unclear parts of papers
  - Transforming data formats (e.g., handwriting to LaTeX)
  - Translating and correcting text
  - ...

Coding with LLMs

- Coding with LLM assistance could make you more productive!
- Use **VSCode** with the Copilot plugin locally or on a cluster (also consider: Cursor IDE, replit agentic web dev platform).
- Some models like 4o and Claude can *execute* basic code.
- How use it? Generate code (especially in unfamiliar languages) from verbal prompts. Debug code.
- Industry adoption: 25% of code at Google is LLM-generated (Jan '25).
- Example prompt:

  "Write code to illustrate the rotation of planets in our Solar System. Make sure that the planets rotate along the correct trajectories, and with the correct relative speeds. Make it interactive so that the user can select which planets to show. Add a legend indicating which planet is which. Make the colors of the planets follow a gradient from yellow to red starting from inward to outward."

- See results: here.

Aside: My LLM Journey

Jailbreaking Black Box Large Language Models in Twenty Queries

Patrick Chao, Alexander Robey,
Edgar Dobriban, Hamed Hassani, George J. Pappas, Eric Wong
University of Pennsylvania

JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models

Patrick Chao[*1], Edoardo Debenedetti[*2], Alexander Robey[*1], Maksym Andriushchenko[*3], Francesco Croce[3], Vikash Sehwag[4], Edgar Dobriban[1], Nicolas Flammarion[3], George J. Pappas[1], Florian Tramèr[2], Hamed Hassani[1], Eric Wong[1]

Leaderboards Paper Contribute Library Behaviors Jailbreak artifacts

JAILBREAKBENCH

Aside: My LLM Journey (ctd)

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang[*†] Shuo Li[*†] Mengxin Yu[†] Matteo Sesia[‡] Hamed Hassani[†]

Insup Lee[†] Osbert Bastani[§†] Edgar Dobriban[§†]





One-Shot Safety Alignment for Large Language Models via Optimal Dualization

Xinmeng Huang[*] Shuo Li[*] Edgar Dobriban

Osbert Bastani Hamed Hassani Dongsheng Ding[†]

Watermarking Language Models with Error Correcting Codes

Patrick Chao, Edgar Dobriban, Hamed Hassani[*]

Evaluating the Performance of Large Language Models via Debates

Behrad Moniri Hamed Hassani Edgar Dobriban