Course Overview

CSCI 601-771 (NLP: Advances in Self-Supervised Models)

https://self-supervised.cs.jhu.edu/fa2025/

Why This Course?

- Let's start with an example: **ChatGPT DeepResearch**

Why This Course?

OpenAI

- Let's start with an example: **ChatGPT DeepResearch**

> Create a table where the rows are recent language model and the columns are different details about their alignment process.

Why This Course?

OpenAI

- Let's start with an example: **ChatGPT DeepResearch**

> Create a table where the rows are recent language model and the columns are different details about their alignment process.

Could you specify which language models you'd like to include in the table? Also, what specific details about their alignment process should be covered—such as training methods, RLHF (Reinforcement Learning from Human Feedback), safety measures, dataset curation, bias mitigation strategies, etc.?

Why This Course?

OpenAI

Could you specify which language models you'd like to include in the table? Also, what specific details about their alignment process should be covered—such as training methods, RLHF (Reinforcement Learning from Human Feedback), safety measures, dataset curation, bias mitigation strategies, etc.?

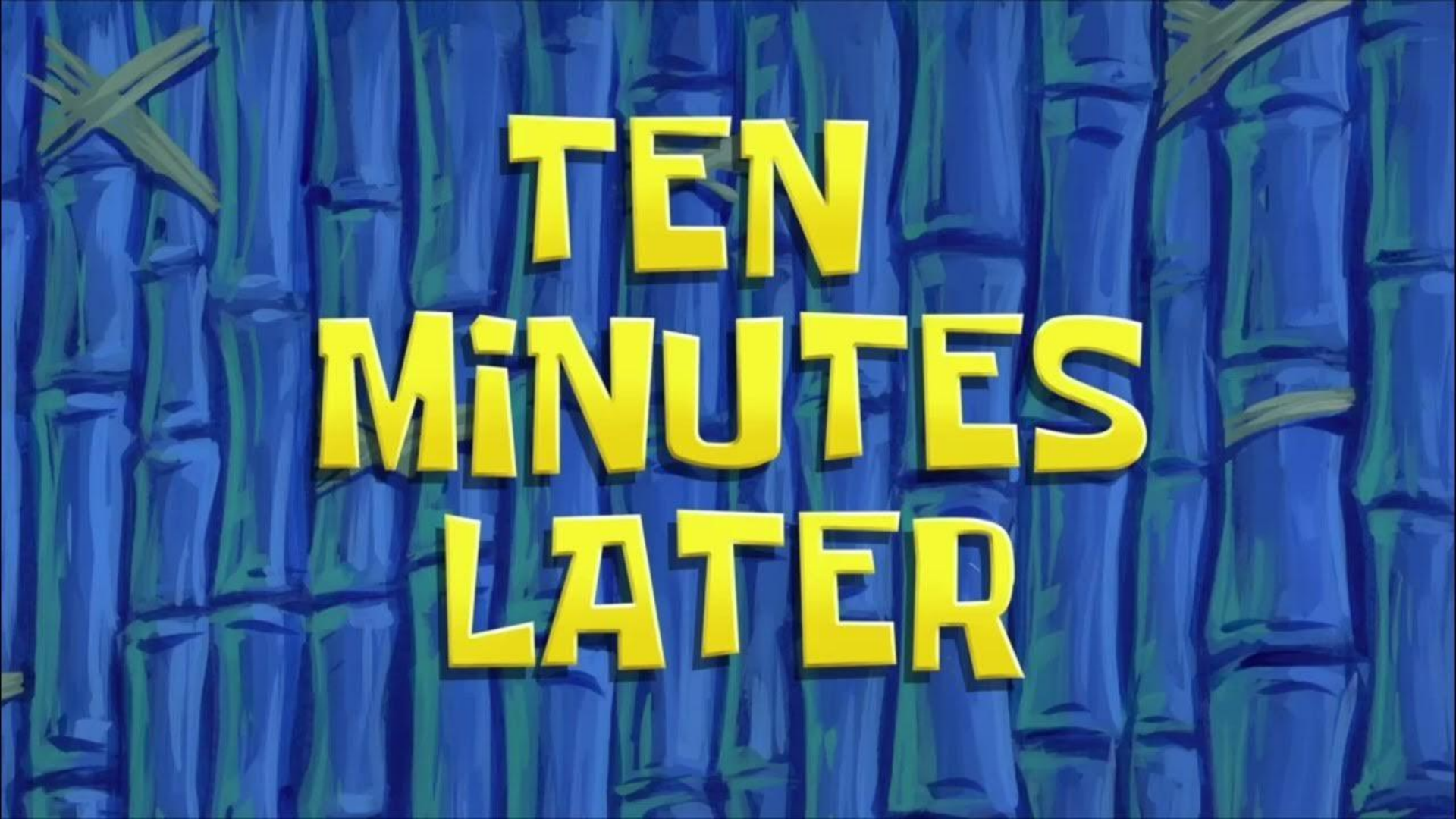
Pick a few recent models and specify steps about their alignment process.

Why This Course?

OpenAI

Pick a few recent models and specify steps about their alignment process.

I'll gather information on recent language models and their alignment processes, including Reinforcement Learning from Human Feedback (RLHF), dataset curation, bias mitigation strategies, and other safety measures. I'll provide a structured table summarizing these details for each model. I'll update you once the research is complete.


TEN
MINUTES
LATER

Create a table where the rows are recent language model and the columns are different details about their alignment process.

ChatGPT **Deep Search (~10 mins)**

Research completed in 11m · 30 sources

Looks impressive; doesn't it?

Model	Alignment Methodology (Key Steps)
OpenAI InstructGPT (GPT-3.5, 2022)	- **Supervised fine-tuning** on a curated set of human-written demonstrations of desired behavior (labelers provided prompts and ideal answers) ARXIV.ORG . - **Reward model training** on human preference data: labelers ranked multiple model outputs for the same prompt, producing a dataset of comparisons ARXIV.ORG . - **RLHF**: The model was further fine-tuned with **reinforcement learning from human feedback** (using PPO) to maximize the reward model's score, aligning outputs with human-preferred responses ARXIV.ORG . - **Safety & bias mitigation**: Human labelers followed guidelines to prefer helpful, truthful, and inoffensive outputs, leading to a model that generated *more truthful and far less toxic* responses than the original GPT-3 ARXIV.ORG .
OpenAI ChatGPT / GPT-4 (GPT-3.5 & GPT-4, 2022–2023)	- **Initial fine-tuning**: Both ChatGPT (based on GPT-3.5) and GPT-4 use a similar alignment pipeline as InstructGPT, starting with SFT on demonstration data and then RLHF fine-tuning CDN.OPENAI.COM . GPT-4's post-training *"alignment process"* notably improved factuality and adherence to desired behavior CDN.OPENAI.COM . - **Safety-focused data**: OpenAI augmented GPT-4's training with an *additional set of safety-relevant prompts* during RLHF CDN.OPENAI.COM . These are adversarial or sensitive queries designed to teach the model to refuse or handle harmful requests appropriately. - **Rule-based reward modeling**: GPT-4 introduced *rule-based reward models (RBRMs)* – essentially GPT-4 classifier models that evaluate outputs against a set of written safety rules CDN.OPENAI.COM CDN.OPENAI.COM . During RLHF, the model received an extra reward signal from these classifiers, encouraging it to **refuse disallowed content** in the correct style and **not refuse** innocuous queries CDN.OPENAI.COM . - **Bias & safety mitigations**: The above methods, along with extensive red-team testing by domain experts, significantly improved GPT-4's safety. GPT-4 is far less likely than GPT-3.5 to produce disallowed content or toxic language (an 82% reduction in improper responses) CDN.OPENAI.COM , and it more often gives a *polite refusal or safe completion* when faced with harmful requests.

Example retrieved on 03/13/2025

ChatGPT

- Can write essays, gather related ideas, expand your research idea, debug your code, etc.

How is it able to do these? 🤔

- What are the principles that govern this technology?
- Can we replicate it?
- Can we extend it?

Meanwhile …

Create a table where the rows are recent language model and the columns are different details about their alignment process.

ChatGPT **Deep Search (~10 mins)**

Model	Alignment Methodology (Key Steps)
OpenAI InstructGPT (GPT-3.5, 2022)	- **Supervised fine-tuning** on a curated set of human-written demonstrations of desired behavior (labelers provided prompts and ideal answers) ARXIV.ORG . - **Reward model training** on human preference data: labelers ranked multiple model outputs for the same prompt, producing a dataset of comparisons ARXIV.ORG . - **RLHF**: The model was further fine-tuned with **reinforcement learning from human feedback** (using PPO) to maximize the reward model's score, aligning outputs with human-preferred responses ARXIV.ORG . - **Safety & bias mitigation**: Human labelers followed guidelines to prefer helpful, truthful, and inoffensive outputs, leading to a model that generated *more truthful and far less toxic* responses than the original GPT-3 ARXIV.ORG .
OpenAI ChatGPT / GPT-4 (GPT-3.5 & GPT-4, 2022–2023)	- **Initial fine-tuning**: Both ChatGPT (based on GPT-3.5) and GPT-4 use a similar alignment pipeline as InstructGPT, starting with SFT on demonstration data and then RLHF fine-tuning CDN.OPENAI.COM . GPT-4's post-training *"alignment process"* notably improved factuality and adherence to desired behavior CDN.OPENAI.COM . - **Safety-focused data**: OpenAI augmented GPT-4's training with an *additional set of safety-relevant prompts* during RLHF CDN.OPENAI.COM . These are adversarial or sensitive queries designed to teach the model to refuse or handle harmful requests appropriately. - **Rule-based reward modeling**: GPT-4 introduced *rule-based reward models (RBRMs)* – essentially GPT-4 classifier models that evaluate outputs against a set of written safety rules CDN.OPENAI.COM CDN.OPENAI.COM . During RLHF, the model received an extra reward signal from these classifiers, encouraging it to **refuse disallowed content** in the correct style and **not refuse** innocuous queries CDN.OPENAI.COM . - **Bias & safety mitigations**: The above methods, along with extensive red-team testing by domain experts, significantly improved GPT-4's safety. GPT-4 is far less

GPT 3.5 has two rows?

Incorrect information (**precision** issues)

Lots of text, not much structure/insight.

Example retrieved on 03/13/2025

No reference to Llama2 paper itself?

Create a table where the rows are recent language model and the columns are different details about their alignment process.

ChatGPT **Deep Search (>15 mins)**

Meta LLaMA-2 Chat (2023)	- **Supervised instruction tuning**: LLaMA-2 Chat was initialized by **supervised fine-tuning** on a large collection of instruction-following data. This included prompt-response pairs from public sources and safety-centric data, so the model learned to produce helpful answers and follow user instructions while adhering to content guidelines VISO.AI VISO.AI. The SFT data covered both general helpful responses and safer completions (e.g. how to politely refuse inappropriate requests).
(continued)	- **Two-step RLHF with separate rewards**: Meta then applied **RLHF** in two dimensions HEIDLOFF.NET. They trained **two reward models** – a *helpfulness* reward model (judging how well the assistant's answer satisfies the user's request) and a *safety* reward model (judging harmful or policy-violating content) HEIDLOFF.NET. The chat model was optimized via Proximal Policy Optimization, balancing these rewards to produce responses that are both useful *and* harmless. They also employed **rejection sampling**: the model generates multiple responses and selects the one with the best combined helpfulness/safety score, which further improves alignment HEIDLOFF.NET. - **Safety context distillation**: An additional safety technique was used wherein, if the model started to produce unsafe outputs, the behavior from a higher-precision safety model or human-written safe responses were distilled back into the chat model (as extra fine-tuning) ARXIV.ORG VISO.AI. This way, the model learns to internalize safer responses for problematic prompts. - **Bias and toxicity mitigation**: The **safety reward model** was explicitly trained on detecting toxic, biased, or harmful content VISO.AI. By optimizing against this model's feedback, LLaMA-2 Chat greatly reduces toxic or biased generations. The model card reports strong performance on safety evaluations compared to previous open models HEIDLOFF.NET HEIDLOFF.NET. However, like other LLMs, it can still be adversarially prompted to reveal unsafe behavior in edge cases VISO.AI, so ongoing evaluation is necessary.

???

The table has these models: GPT-3.5, GPT-4, Claude, Sparrow, LaMDA, LLaMA-2
What about the rest? **Recall issue.**

Example retrieved on 03/13/2025

The table I created on my own

Create a table where the rows are recent language model and the columns are different details about their alignment process.

What I manually created for my course

Language Model	Release	Base	Alignment Algorithm(s) Used	Alignment Data Sources for alignment
GPT-3-instruct	2020	GPT-3	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
GPT-4	2023	GPT-4 pre-trained?	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
Gemini	2023	Gemini pre-trained?	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
LLaMA2	2023	LLaMA2 pre-trained	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
LLaMA3	2024	LLaMA3 pre-trained	Iterate: Rejection sampling -> SFT -> DPO	10 million human-annotated examples. The alignment process was conducted over multiple rounds, with each round involving the collection of new preference annotations and SFT data.
Alpacha	2023	LLAMA 1	SFT	Self-Instruct, 52,000 input-output pairs
Qwen2.5	2024	Qwen2.5 pre-trained	SFT -> DPO -> GRPO	1 million samples
Tulu 3	2024	Llama 3.1	SFT -> DPO -> RLVR	near 1 million samples
DeepSeek (V3)	2024	DeepSeek pre-trained	SFT -> GRPO	1.5 million samples (reasoning + non-reasoning tasks). Reasoning data was generated by specialized models. Non-reasoning data was produced by DeepSeek-V2.5 and validated by human reviewers.

AI (e.g., Deep Research) is far from perfect

- LLMs know about many interesting connections. But can they be more precise?
- LLMs are fluent and verbose.
  - Remarkable on the first sight.
  - But quickly becomes a problem for information overload.
- Note that this was an easy prompt — extensively discussed online (papers, Twitter, Reddit, etc.) and hence, during pre-training.

Why does it make such surprisingly simple mistakes? 🤔

- Can we explain these?
- Can we predict them?
- Can we mitigate (or even better, solve) them?

Course Learning Objectives

- Broad research question:
  - **Reasoning:** Understanding LLMs ability and limitations in performing complex, multi-step inference.
  - **Long inputs/output:** Understanding LLMs ability and limitations in processing long inputs and generating long outputs.
  - **Efficient inference:** Understanding the latest advances and hurdles in efficient inference.
- Skills:
  - **Technical**—understanding of the algorithms and implementing them.
    - Gaining intuitions about **capabilities** and **limitations** of models.
  - **Soft skills**—intuition about capabilities, teamwork.

Focus on Natural/Human Language

- **Most** of the class revolves around **natural language**.

- Why natural language?
  - It is a **convenient medium of communication**.
  - Natural language is our species' best attempt to encode **everything about the world** as **efficiently** as possible.
  - A huge archive of natural language is **freely available** (e.g., on the web).

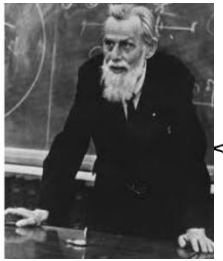
How did we get here?

Progress in AI

- Many advances are due to neural networks
- How old are neural networks?

[Slide credit: Arman Cohan]

Progress in AI



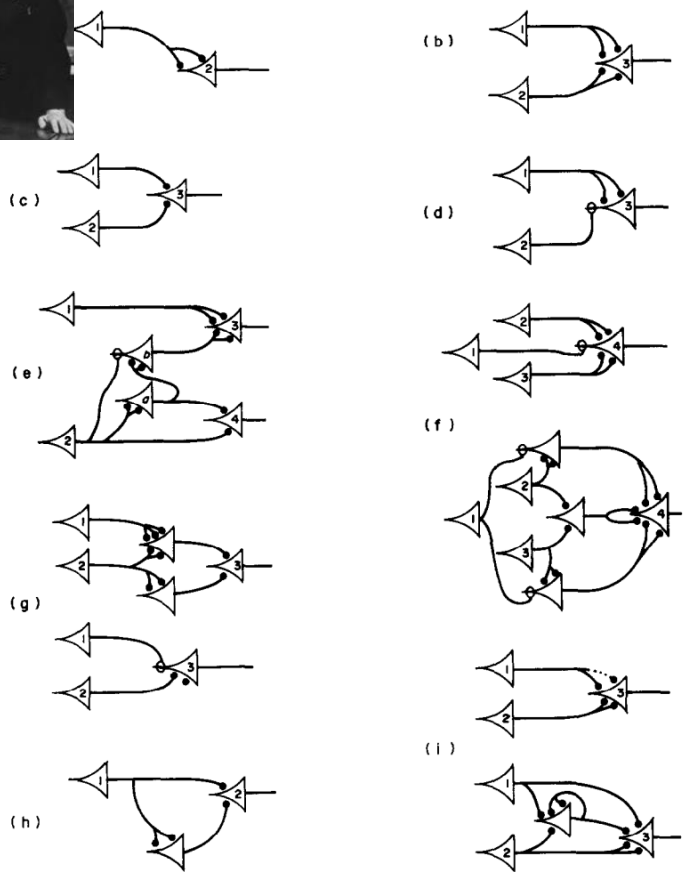
- Many advances are due to neural networks
- How old are neural networks?

McCulloch & Pitts (1943)

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

- WARREN S. MCCULLOCH AND WALTER PITTS
  University of Illinois, College of Medicine,
  Department of Psychiatry at the Illinois Neuropsychiatric Institute,
  University of Chicago, Chicago, U.S.A.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.



[Slide credit: Arman Cohan]

Progress in AI

- Many advances are due to neural networks
- How old are neural networks?
  - They've been around since the 1940s
  - But why have only recently we seen breakthroughs?
    - **3 necessary forces had to come together!**

[Slide credit: Arman Cohan]

Force 1: Massive Amount of Data

- Internet provided us with a massive repository of data.

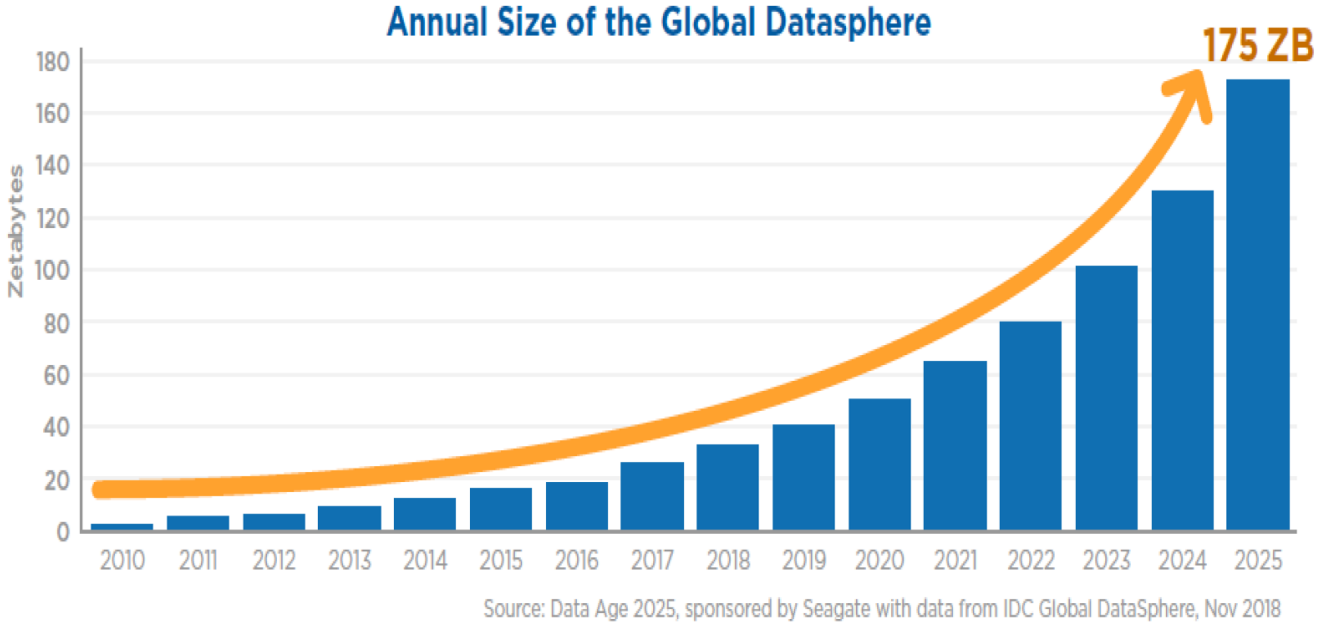
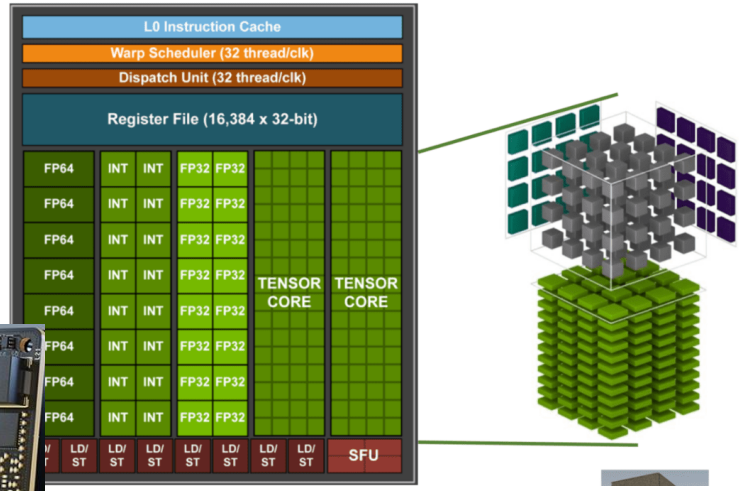
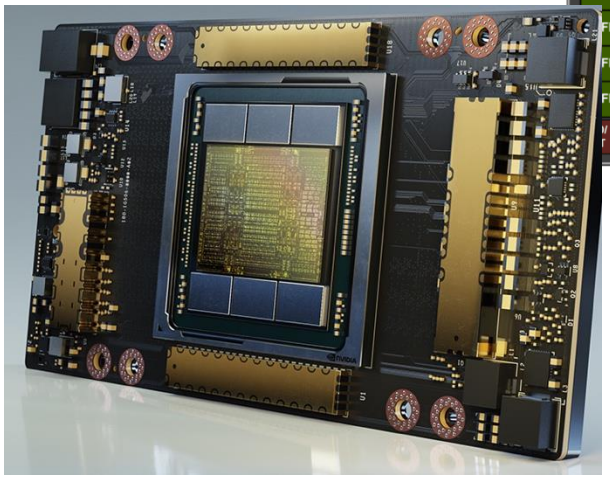


Image source: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/?sh=1e6ca8455459

Force 2: Computing Power

- Fast processors for deep learning!



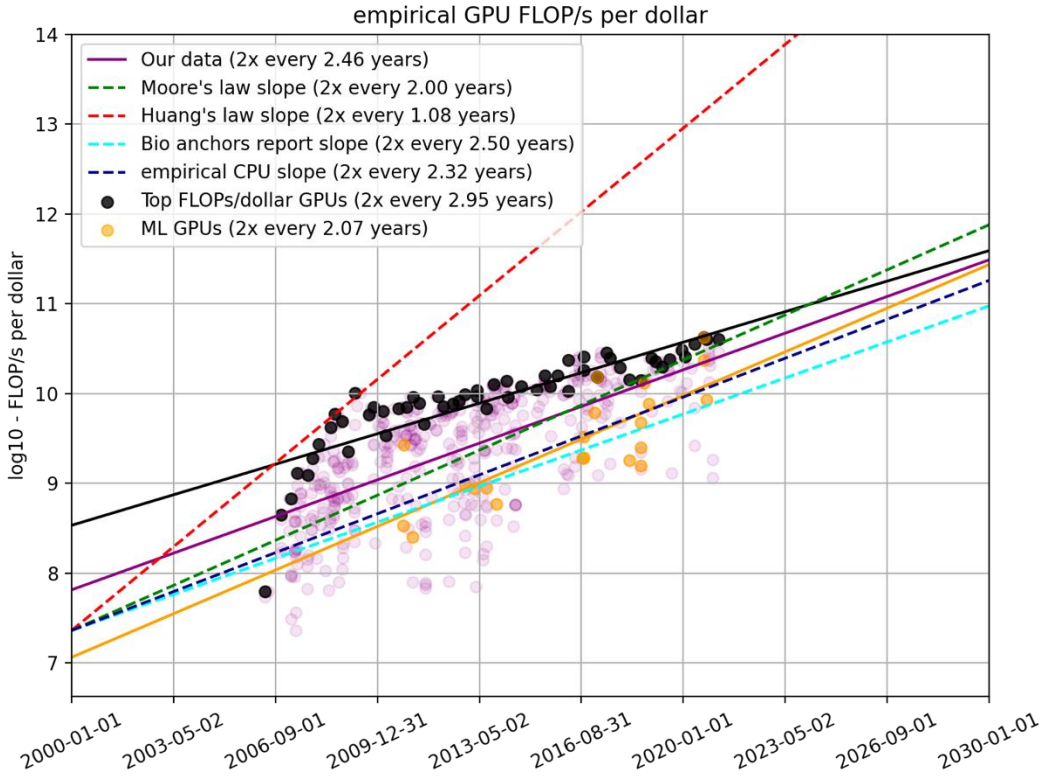


A100 GPU

Force 2: Computing Power

- Fast processors for deep learning!
- Cheaper computing power over time.

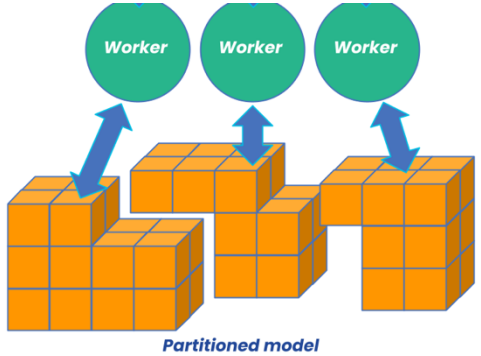
The amount of computing power, per dollar



https://www.lesswrong.com/posts/c6KFvQcZggQKZzxr9/trends-in-gpu-price-performance

Force 2: Computing Power

- Fast processors for deep learning!
- Cheaper computing power over time.
- Distributed training/inference allows us to scale to a larger set of processors.



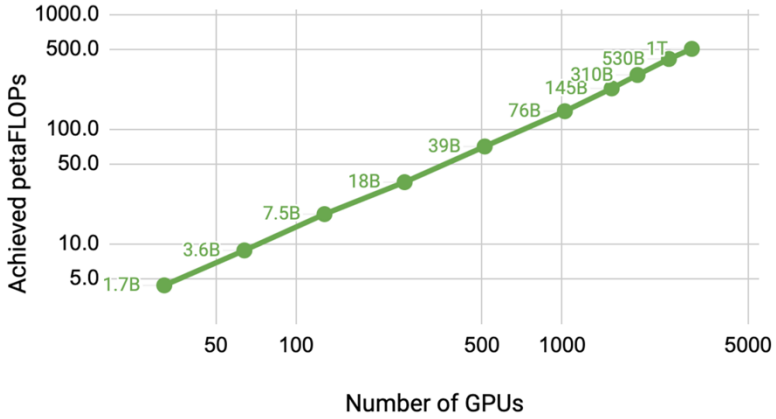
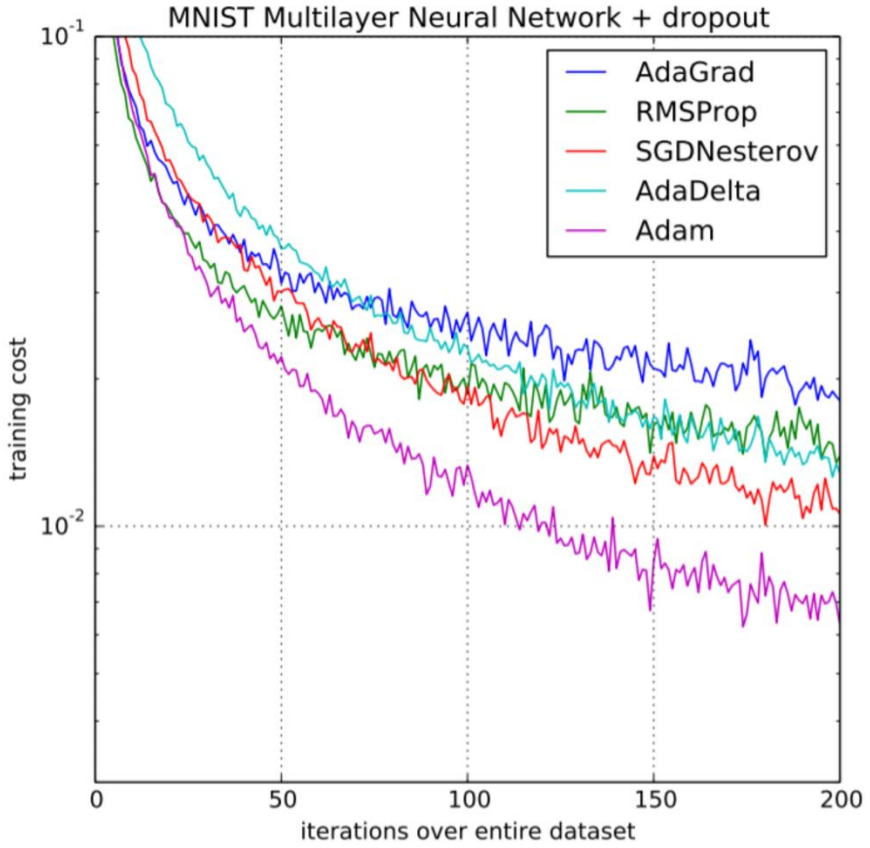


Image source: https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/

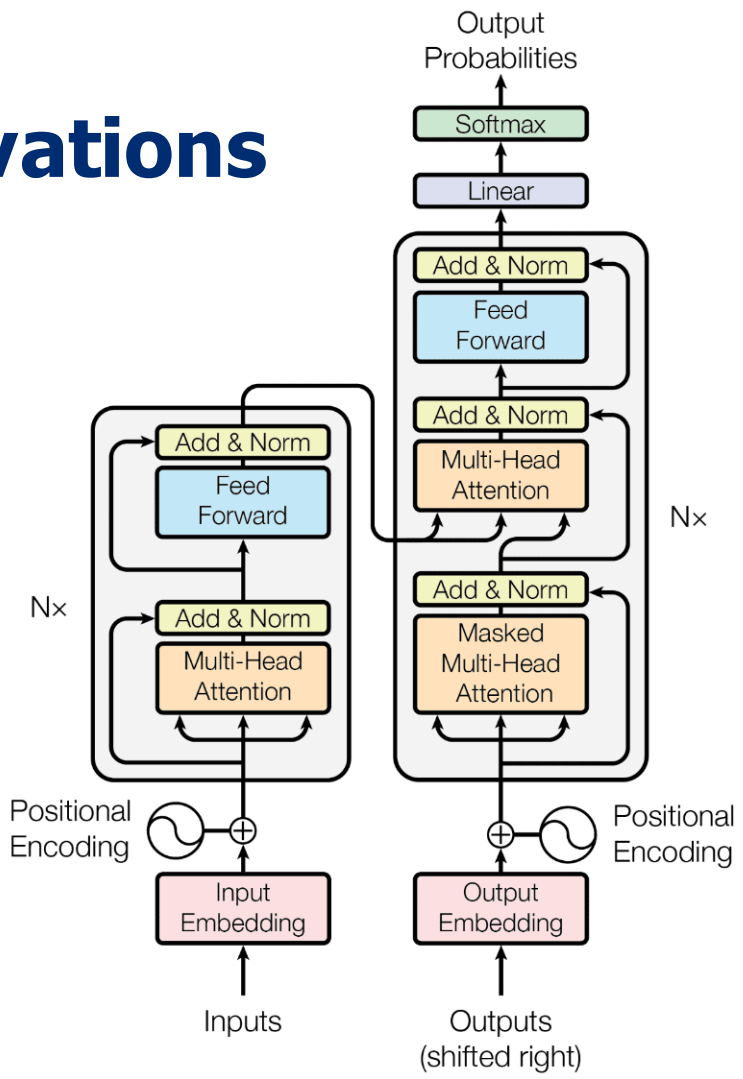
Force 3: Algorithmic innovations

- Advances in optimization



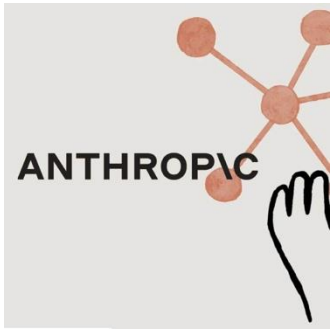
Force 3: Algorithmic innovations

- Advances in optimization
- Innovations in model architectures
- ….



The success we dreamed of

Gemini
Google



OLMoE
Ai2

Language models that are remarkably capable at solving many important NLP benchmarks.

Course Learning Objectives

- Broad research question:
  - **Reasoning:** Understanding LLMs ability and limitations in performing complex, multi-step inference.
  - **Long inputs/output:** Understanding LLMs ability and limitations in processing long inputs and generating long outputs.
  - **Efficient inference:** Understanding the latest advances and hurdles in efficient inference.
- Skills:
  - **Technical**—understanding of the algorithms and implementing them.
    - Gaining intuitions about **capabilities** and **limitations** of models.
  - **Soft skills**—intuition about capabilities, teamwork.

Course Logistics

CSCI 601-471/671 (NLP: Self-Supervised Models)

https://self-supervised.cs.jhu.edu/sp2025/

Course Logistics Brief

- **Instructor:** Daniel Khashabi
  - You can call me "Daniel", as long as we act mutually respectfully.

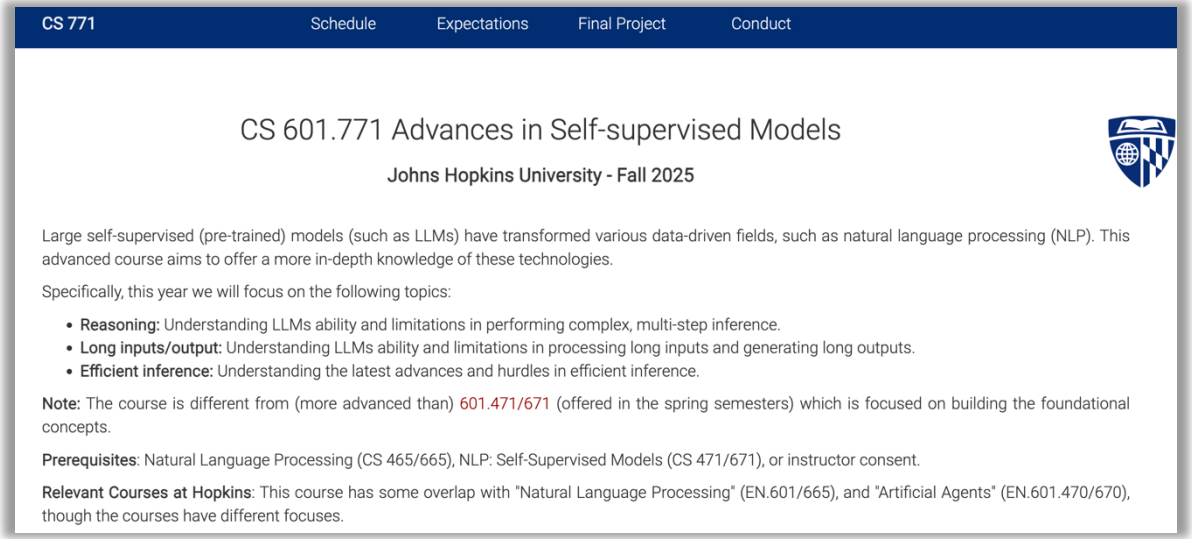
- **Course Assistant:** Anushri Suresh

Daniel Khashabi
Instructor

Anushri Suresh
Course Assistant

Course Website

- Lots of important information on the **website**:
  - https://self-supervised.cs.jhu.edu/fa2025



Course Prerequisites

- Comfortable with programming, particularly Python
- Comfortable with the foundations:
  - Transformer, pre-training, fine-tuning
  - Some basic understanding of alignment
- HW1 should give a sense how prepared you are!

During the Class

- **A small team** (usually 2 students) present 1-2 papers (20-25 minutes).
- We (collectively as a class) will critically discuss the paper (remaining 30 mins).
  - This may also turn into open-ended conversation.

Before the Class

- **The presenter students:**
  - You will know that you're presenting 10 days in advance.
    - Anushri is in charge of assignments. Negotiate with her! :)
  - You need to share your slides with us 48 hours before the class.
    - We will give you feedback so that you can improve your slides.

- **The non-presenter students:**
  - Write a one-pager summary of what was discussed last time.
  - A forcing mechanism to get us to think about what learned during the class.
  - At the beginning of each class, one of the students (randomly selected) will remind us what we talked about in the previous class based on their one-pager.

During the Class [Revised]

- **Beginning of the class:** A student will remind us what we talked about in the previous class based on their one-pager.

- **A small team** (usually 2 students) present 1-2 papers (20-25 minutes).

- We (collectively as a class) will critically discuss the paper (remaining 30 mins).
  - This may also turn into open-ended conversation.

Grading Policy

- **One** Homework (individual): 10%
- In-class participation (individually): 20%
  - Discussions: Step Up/Step Back
  - Punctuality: being at class before we start
- One-pager summary and its presentation: 20%
- Paper presentation and critique: 20%
- Final project (team): 30%

- Skip days: You may skip up to 3 classes (during the dates that you're not presenting) with instructor consent ahead of time.

Final Project

- Must be exploring a topic related to the focus of the class.
- This is your chance to gain research experience on a topic of interest.
- Topic choice will be (relatively) free. We will help you develop your ideas!
- **Deliverables:**
  1. Submit project **proposal** outline (for our formal review and suggestions)
     - To make sure that the project is scoped reasonably and doable in your limited time.
  2. Get excited  and work on the project
  3. Midway report
  4. Final project presentation, poster session, report

Communication Mechanism

- A Slack channel where:
  - We will have a room for sharing general discussions
  - Direct communication with me, if needed.
  - You can also create channels for your teamwork.

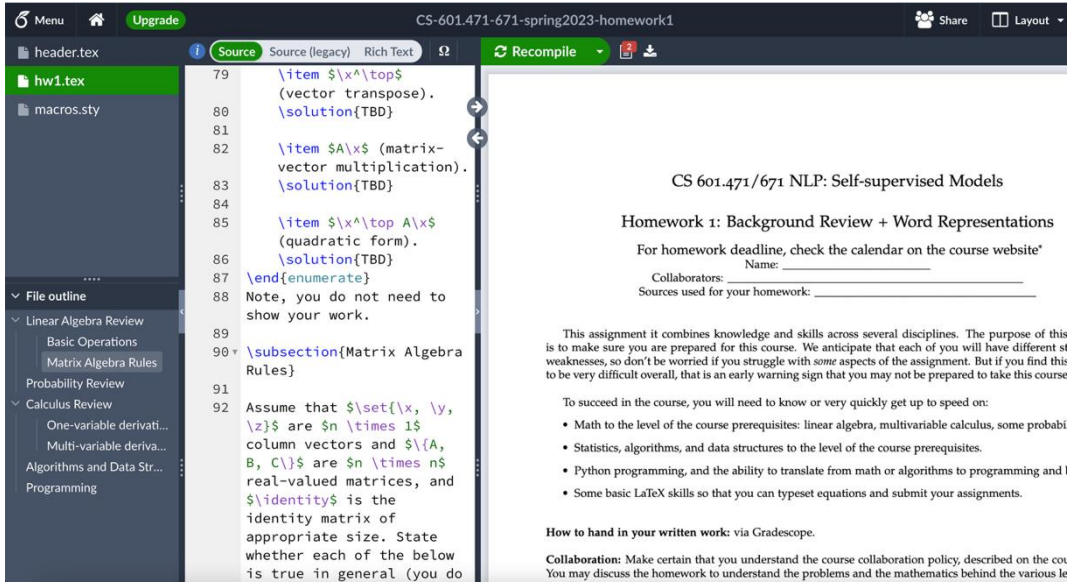
- Are people comfortable with Slack?

HW1 is released!

- Due Thursday next week.
  - Has both theory (background on algebra, etc.) and programming (building a classifier).
- A **baseline** for self-assessment.

"Is Typesetting Mandatory?"

- Yes!



Quick pulse check (1)

- I have understood the course expectations!
  - Yes
  - No

Quick pulse check (2)

- I am on the waitlist. Will I be able to register?
  - I don’t know.
  - However, typically, 20-30% of students drop the course within the first week. If you’re among the top 20 on the waitlist, there’s a good chance you’ll get in.

Quick pulse check (3)

- I am currently on the waitlist and very eager to take this course. I'm optimistic about getting in. Should I submit the assignments even though I'm not officially registered?
  - **Yes!**
  - If you're on the waitlist and excited to join, you should go ahead and submit the homework assignments.
  - Everyone—whether they joined late or were already registered—must adhere to the same deadlines.
  - No HW extensions for late joiners.

Wrapping it up!

- HW1 is released!
- If you're not going to take this, drop the course!