

Welcome to CS 294-288

(Class starts at 2:10)

Course website: sewonmin.com/courses/cs294_288

Today's Plan

- Class intro
- Paper discussion
- Course project
- Course logistics
- — QnA —
- Qualification quiz (2:55–3:30)

Course Staff

- Instructor: Sewon Min
 - Faculty in EECS since 2025, co-PI of BAIR and Berkeley NLP
 - Research: Foundation model architecture, training, data, auditing
 - PhD 2024 from University of Washington
 - Previously: Meta (part-time, 4 yrs), Google (part-time, 0.5 yr), currently Ai2 (part-time, 2 yrs)
 - Teaches CS 294-288 (this class) every fall, CS 288 (Grad-level NLP) every spring
- Course admin: Jimin Ha

Class info

Website: sewonmin.com/courses/cs294_288 (all materials, schedule, instructions for presentations and projects, grading details, etc)

Communication: Slack! (You'll be invited once your enrollment is confirmed)

- Please use Slack DM, not email

Data-centric LLM

Advances in LLMs have been driven by the growing availability of large and diverse datasets.

- But where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems **emerge** in the **2026+** model-data ecosystem?

We will explore these questions through paper readings, group discussions, & an open-ended project.

Data-centric LLM

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

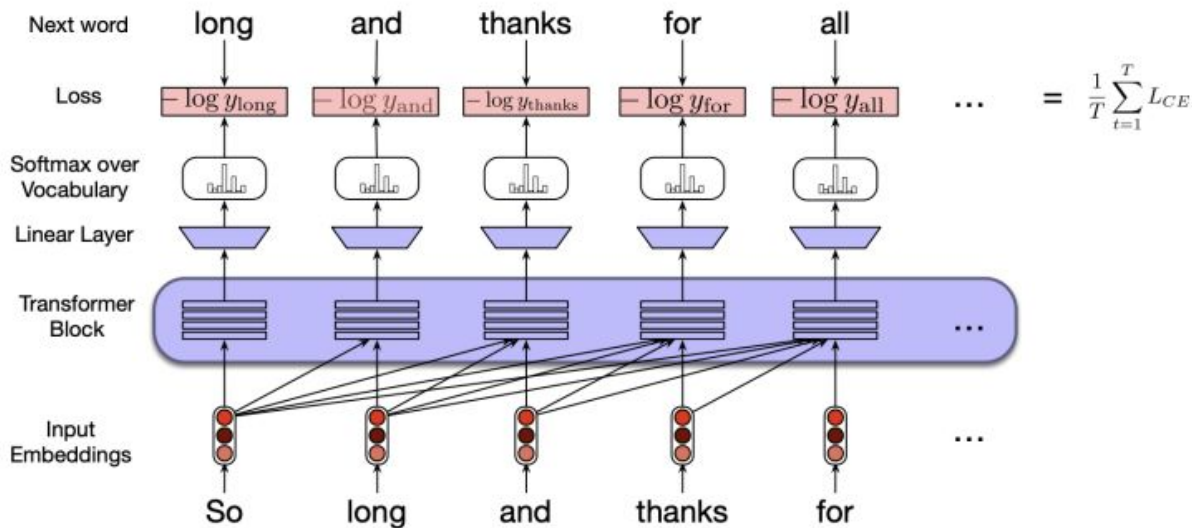
Data-centric LLM

- Where do these datasets come from?
- **How are they being used?**
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

Data-centric LLM: Modeling

- Where do these datasets come from?
- **How are they being used?**
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

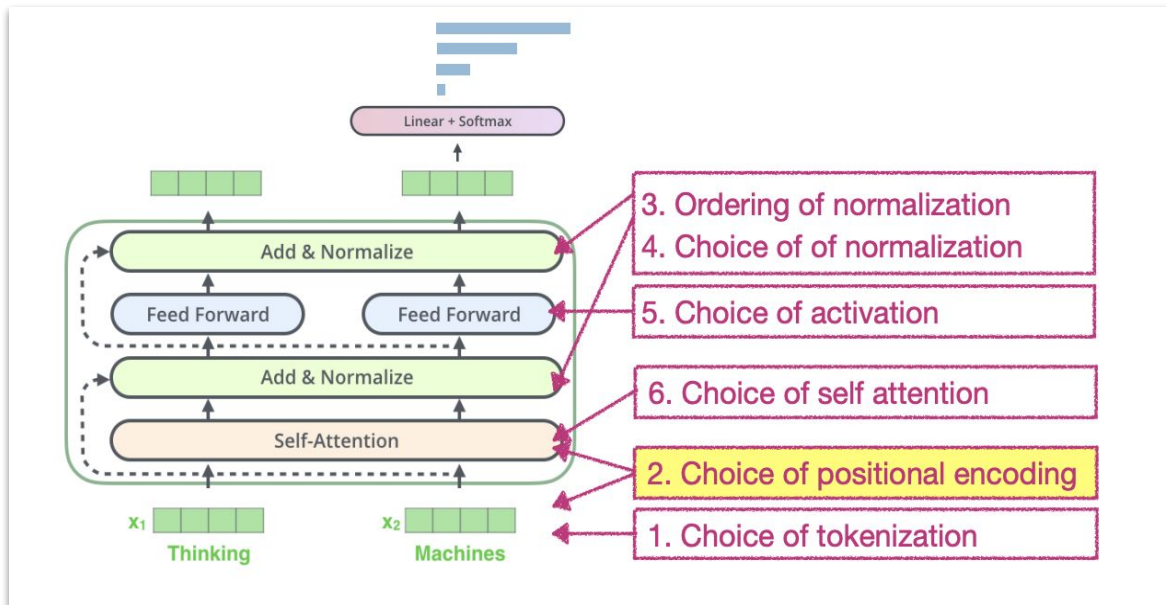
You learned them already from CS 288 (cal-cs288.github.io/sp26)



Data-centric LLM: Modeling

- Where do these datasets come from?
- **How are they being used?**
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

You learned them already from CS 288 (cal-cs288.github.io/sp26)



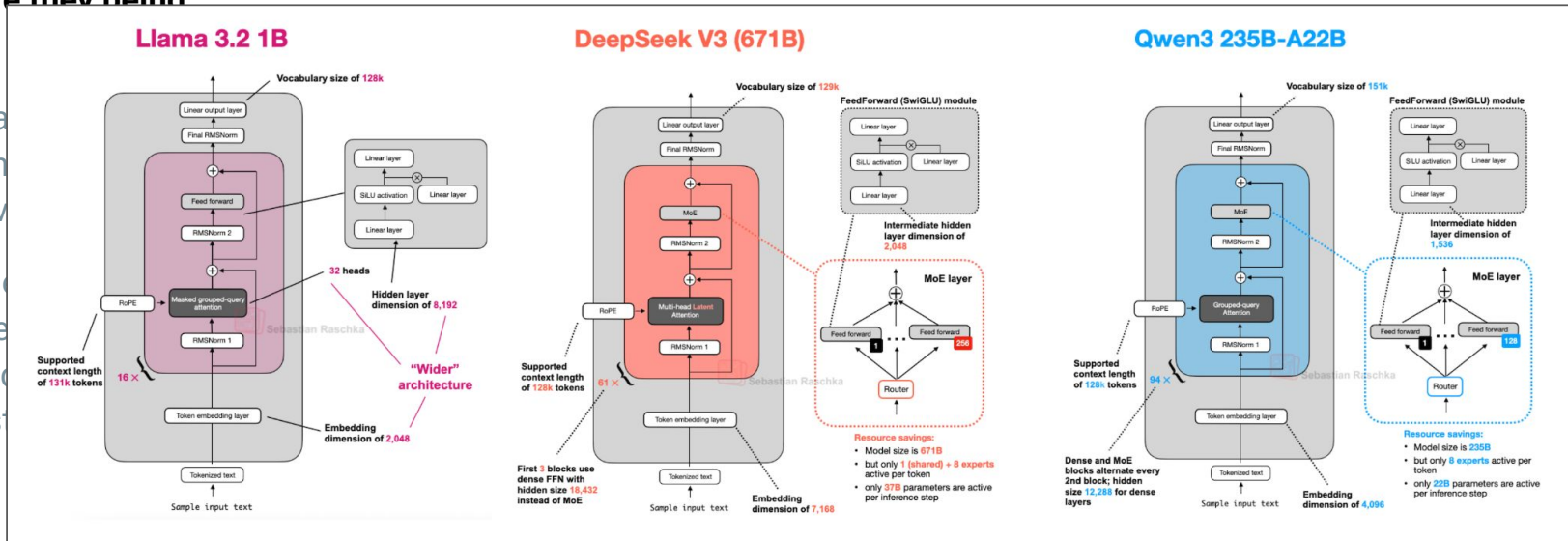
Data-centric LLM: Modeling

- Where do these datasets come from?

You learned them already from CS 288 (cal-cs288.github.io/sp26)

- How are they being used?

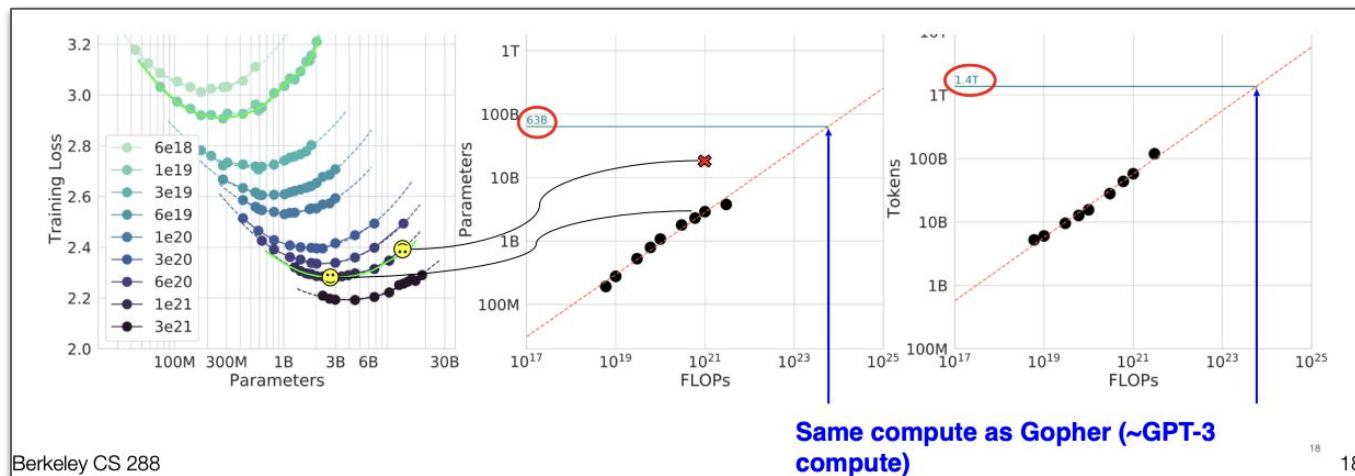
- How can we use them in a more effective way?
- What new emergent model capabilities can we build in this ecosystem?



Data-centric LLM: Modeling

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

You learned them already from CS 288 (cal-cs288.github.io/sp26)

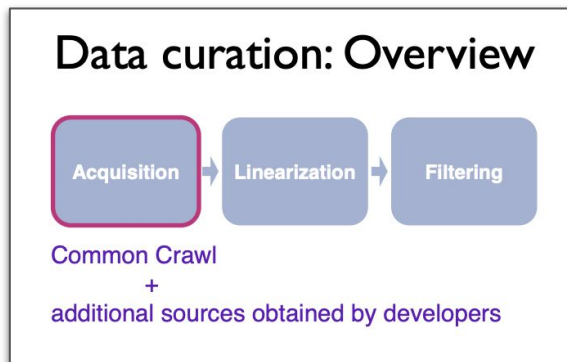
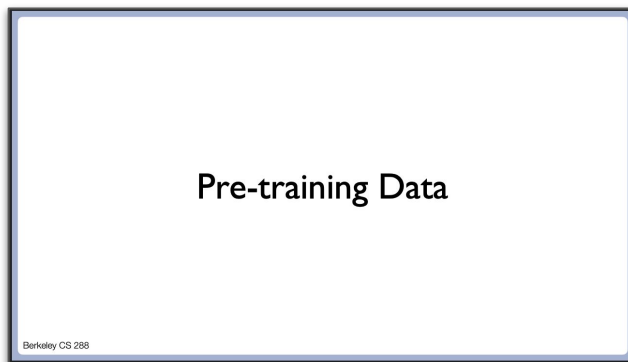


We assume you already know all of these

Data-centric LLM: Data curation

- **Where do these datasets come from?**
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

You learned them (briefly) from CS 288 (cal-cs288.github.io/sp26)

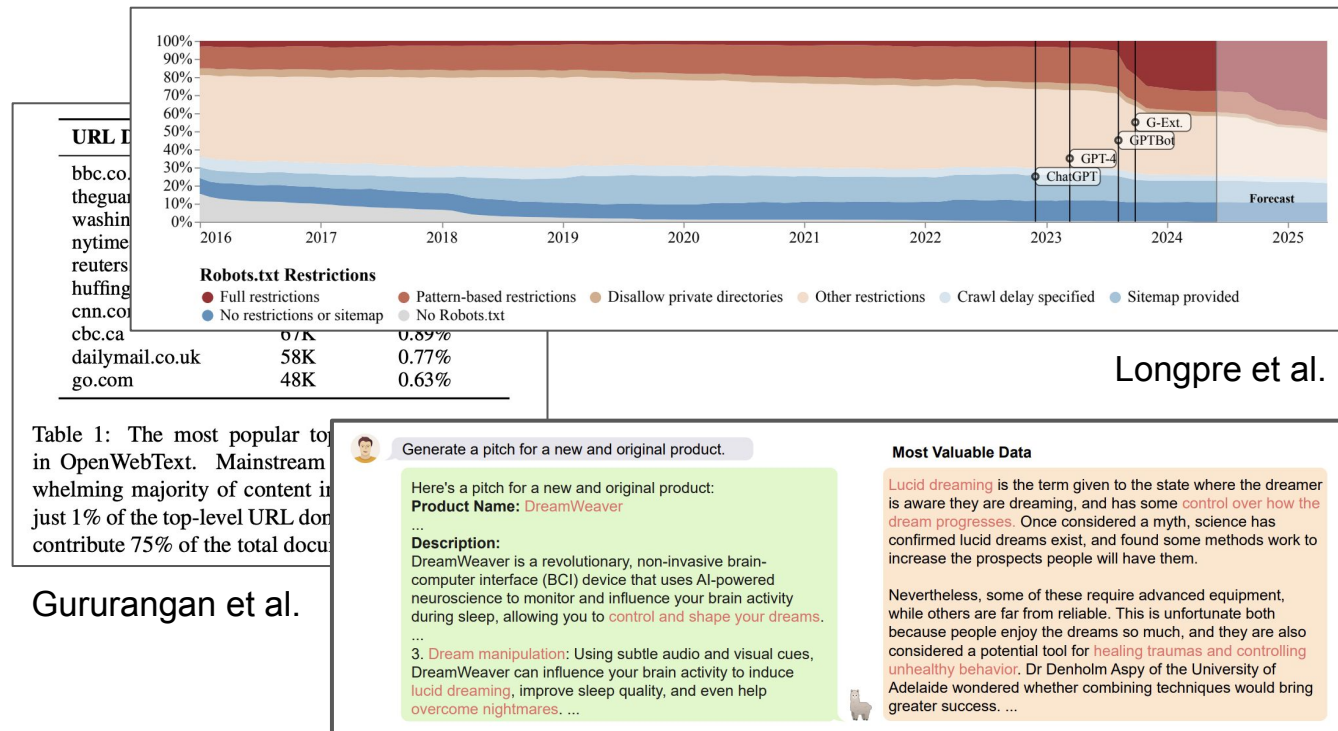


But you'll learn more about these

Goal: Suppose you are given nothing but GPUs & Internet. How will you turn it into a usable language model?

Data-centric LLM: Data provenance

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?



Longpre et al.

Gururangan et al.

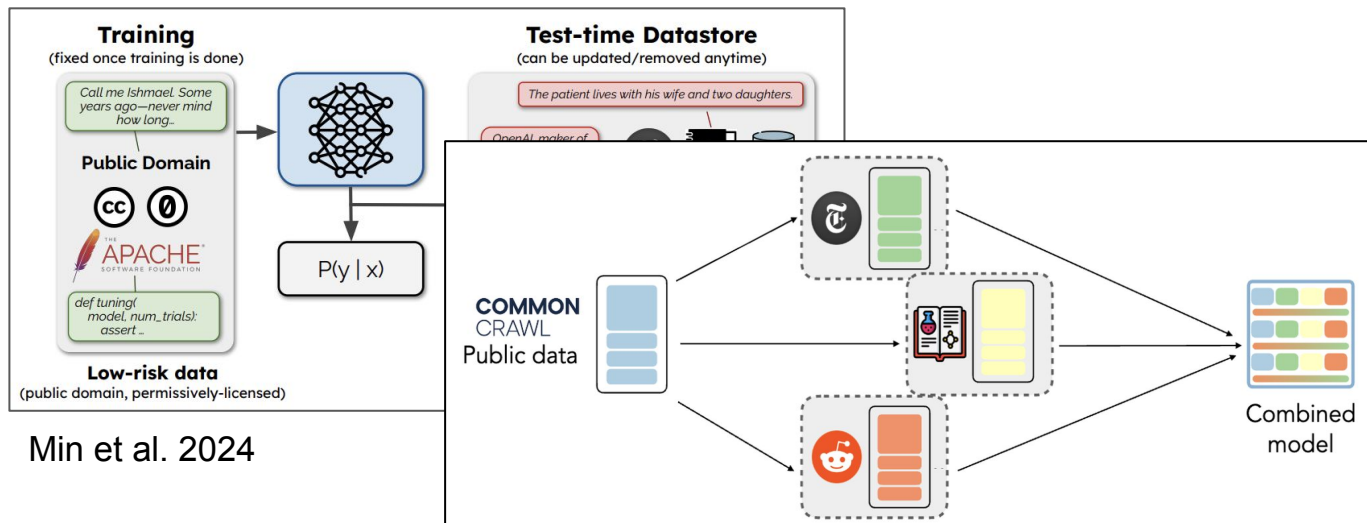
(We will discuss all these papers in the class)

Choe et al.

Data-centric LLM: Future modeling

- Where do these datasets come from?
- How are they being used?
- **How can we leverage them in better, more effectively?**
- What new problems emerge in the 2026+ model-data ecosystem?

Models motivated by concrete data problems



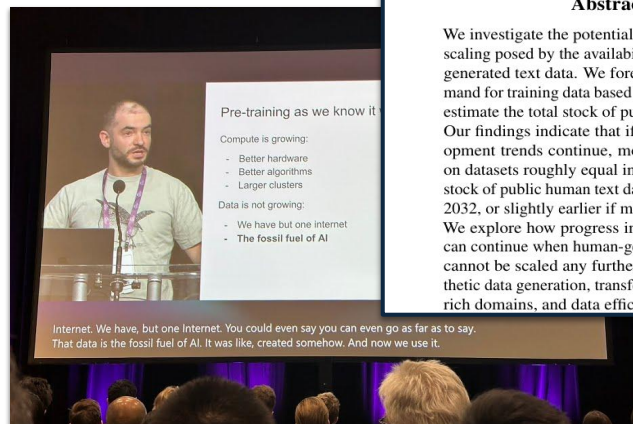
Min et al. 2024

Shi et al. 2025

Emerging model-data challenges

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- **What new problems emerge in the 2026+ model-data ecosystem?**

Example 1: “Data is



“We have but one Internet.”

Ilya Sutskever @ NeurIPS 2024

Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos¹ Anson Ho¹ Jaime Sevilla^{1,2} Tamay Besiroglu^{1,3} Lennart Heim^{1,4} Marius Hobbhahn^{1,5}

Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling can continue when human-generated text datasets cannot be scaled any further. We argue that synthetic data generation, transfer learning from data-rich domains, and data efficiency improvements

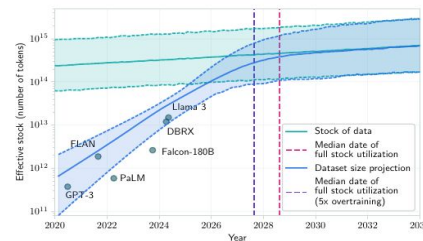


Figure 1. Projections of the effective stock of human-generated public text and dataset sizes used to train notable LLMs. The intersection of the stock and dataset size projection lines indicates the median year (2028) in which the stock is expected to

Forecast: Human text may be exhausted in 2028.
(We will discuss this paper in the class)

Emerging model-data challenges

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- What new problems emerge in the 2026+ model-data ecosystem?

Example 2: AI-generated text



The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text

Yanzhu Guo¹, Guokan Shang^{2,4*}, Michalis Vazirgiannis^{1,2}, Chloé Clavel³

¹École Polytechnique, Institut Polytechnique de Paris

²Mohamed bin Zayed University of Artificial Intelligence

³Inria ⁴Linagora

Chantal Shaib^{1,*} Tunin Chakrabarty² Diego Garcia-Olano³ Byron C. Wallace¹

¹Northeastern University, ²Stony Brook University, ³Meta AI

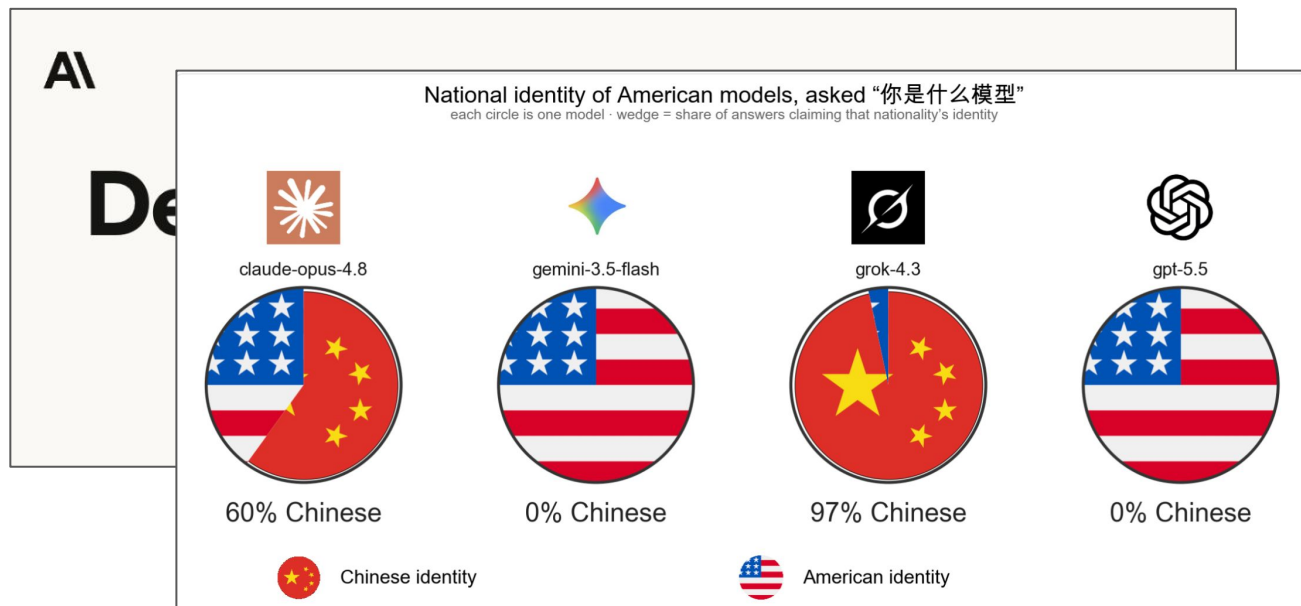
shaib.c@northeastern.edu

(We will discuss all papers in the class)

Emerging model-data challenges

- Where do these datasets come from?
- How are they being used?
- How can we leverage them in better, more effectively?
- **What new problems emerge in the 2026+ model-data ecosystem?**

Example 3: Distillation



Class topics

It's a LOT of topics, as you can see, so we have slightly different focuses each year.sure

Class topics

It's a LOT of topics, as you can see, so we have slightly different focuses each year.

Last year: Broad coverage of data / models / evals, both pre-training & post-training
sewonmin.com/courses/cs294_288_fa25

Class topics

It's a LOT of topics, as you can see, so we have slightly different focuses each year.

Last year: Broad coverage of data / models / evals, both pre-training & post-training
sewonmin.com/courses/cs294_288_fa25

This year: Focus on **pre-training / foundation model training**, with more time on emerging data-model ecosystem problems (1–2 classes for other topics)

- Projects on agents, reasoning models, or VLM/VLA are eligible only when they're tightly connected to a data-model ecosystem problem – discuss with the instructor if unsure

Class info

We will study these questions through paper reading, discussion, and the project.

We have

- ~2 lectures
- Your paper presentation & discussion
- Project presentations
- ~2 breakout discussion sessions

Paper discussion (1/3)

Detail: sewonmin.com/courses/cs294_288/paper_discussion

For each class, we have

- 2 main presenters (authors)
- 2 critics
- 2 auditors

Each class: Main presentation + Panel discussion + Open-ended discussion

Paper discussion (2/3)

Everyone: Read papers **in detail** before class (presentations assume everyone has read them)

Main presenters

- Present papers as if you were the authors, reflecting on your work in 2026. 30-min, up to 45-min after QnA.
 - What motivated the work at the time?
 - What was valuable / changed the field?
 - What did we miss or what was later contradicted?
 - If we redid this work in 2026, what would we do differently?
 - (Basically take the authors' side)
- Share slides in the dedicated Slack channel 72 hours before class. Instructors may provide feedback/suggestions. **Encouraged:** Upload a practice recording or schedule a practice presentation for feedback.

Paper discussion (3/3)

Critics vs. Proponents (main presenters)

- Panel discussion: 5-min opening remarks each, then back-and-forth discussion
- Critics challenge the work, proponents defend it.
- 1 slide each for opening remarks (extra slides only for concrete experimental evidence from their own experiments or later work)
- Critics share slides 24 hours before class

Open-ended discussion

- Main presenters prepare 2-3 open-ended questions that would have meaningful opposing viewpoints and disagreements.
- The class discusses them in breakout groups.

Project

Research project directly related to course topics (detail: [Project doc](#))

Deliverables: Proposal (Due 9/4 Fri), midpoint report/presentation, final report/presentation

The level of quality expected for a full-score final report is **comparable to that of a conference paper accepted** at ICLR/NeurIPS/ICML/COLM/ACL. This is an ambitious timeline. Consider feasibility. Consult with your PhD adviser before enrolling.

In short, **treat this as a real research project.**

All projects should be completed in teams of 2–4 students.

OH before proposal ddl: calendar.app.google/AT4dEgBPt71VSYEd9

Some other course info

- Grading: In-class activity (50%) + Project (50%) – see website
- Auditing: You still need to pass the qualification quiz
- Attendance: Required
 - Absences reported in advance (on the form) allowed. No additional notice needed.
 - Otherwise, absences are generally not allowed. If necessary, DM Sewon before the day of class (deduction will apply)
- Laptop policy
 - Can be used for course-related purposes (checking details in the paper, note-taking)
 - No laptop use while classmates are presenting. You will receive a warning if this happens.

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

NOT ALLOWED:

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

NOT ALLOWED:

- For ALL presentations
 - AI is good for creating beautiful slides
 - In our class, we want **your own** interpretation of the paper, opinions, & comments.

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

NOT ALLOWED:

- For ALL presentations
 - AI is good for creating beautiful slides
 - In our class, we want **your own** interpretation of the paper, opinions, & comments.
- Any written assignments (proposal, reports, auditor notes) that require interpretation, your own summarization, or analysis.
 - Only allowed for unambiguous/deterministic tasks like reporting results.
 - We'll not evaluate grammar/fluency – only the substance of ideas, clarity, & logical flow.

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

NOT ALLOWED:

- For ALL presentations
 - AI is good for creating beautiful slides
 - In our class, we want **your own** interpretation of the paper, opinions, & comments.
- Any written assignments (proposal, reports, auditor notes) that require interpretation, your own summarization, or analysis.
 - Only allowed for unambiguous/deterministic tasks like reporting results.
 - We'll not evaluate grammar/fluency – only the substance of ideas, clarity, & logical flow.
- For all readings
 - People intentionally choose what to read closely vs. what to skim. It influences & shapes your research taste. Don't let AI to take control over your research taste.
 - Only allowed for fact-checking after you've read it yourself.

AI policy

Can be assisted for writing codes, running experiments, & literature search assistance

NOT ALLOWED:

- For ALL presentations
 - AI is good for creating beautiful slides
 - In our class, we want **your own** interpretation of the paper, opinions, & comments.
- Any written assignments (proposal, reports, auditor notes) that require interpretation, your own summarization, or analysis.
 - Only allowed for unambiguous/deterministic tasks like reporting results.
 - We'll not evaluate grammar/fluency – only the substance of ideas, clarity, & logical flow.
- For all readings
 - People intentionally choose what to read closely vs. what to skim. It influences & shapes your research taste. Don't let AI to take control over your research taste.
 - Only allowed for fact-checking after you've read it yourself.

Unauthorized AI use will result in a zero score for the entire course.

Next classes

9/1: First lecture on
pre-training data curation

9/3: Guest lecture by
Shayne Longpre (MIT PhD
→ Anthropic)

9/8: Your presentations
start! (assignment
announced by 9/1)



Any questions?