Course introduction

CS 685, Spring 2025

Advanced Natural Language Processing

http://people.cs.umass.edu/~hschang/cs685/

Haw-Shiuan Chang

College of Information and Computer Sciences

University of Massachusetts Amherst

Most slides come from Mohit Iyyer

Survey

- Why do you want to take this course?
  - **Curiosity**: LLMs are very cool. I want to learn more about LLMs.
  - **Industry Jobs**: LLMs are very popular. Some related questions might appear in my interview
  - **Research**: I am studying (or want to) something related to LLMs
  - **Recommendation**: I heard that this is a good course from the previous students
- I will sometimes ask you if you understand what I said.
  - Please raise your hand if you do understand

Course logistics

- Follow along w/ the lectures in-person (or Zoom)
  - I will try and see if Zoom also works
  - No guarantee
- There might be a short quiz about the week's topics to be submitted on Gradescope (none for the first week!)
  - You will get full scores by trying to answer them
- Gradescope for all assignment submissions
  - Please find me after the class if you haven't been added to Gradescope

who?

TAs:

Ankita Gupta
Erica Cai
Nguyen Tran

Check out nlp.cs.umass.edu for news/info on NLP research going on at UMass!

email all of us (including me!) at
cics.685.instructors@gmail.com

course website:
https://people.cs.umass.edu/~hschang/cs685

Office hours (in-person and on zoom)

Tuesday w/ Erica: 4-5PM in CS207 Cube 2
Wednesday w/ Ankita: 4-5PM in CS207 Cube 2
Thursday w/ Haw-Shiuan: 11AM-12PM in CS207 Cube 2
Friday w/ Nguyen: 3-4PM in CS207 Cube 2

If necessary, TA office hours will be extended by one hour during homework / exam weeks

Office hours will begin next Monday 2/10 (none before then)

Readings

- No need to buy any textbooks!
- Readings will be provided as PDFs on website
  - Usually NLP research papers / notes

waitlist override pass/fail etc.

- don’t email us about getting into the class because we can’t help… please contact Eileen Hamel at hamel@umass.edu with such questions or requests
- Add/drop deadline is **Feb 12** for graduate students and **Feb 5** for undergrads

Questions / comments?

- Submit questions/concerns/feedback to Piazza
  - Try to ask LLMs first before you ask them at Piazza
  - Could be anonymous
  - Please find me after the class if you haven't been added to CS 685 Piazza
  - 

- FAQ
  - does this course require prior knowledge of NLP? *No, but basic ML/probability/stats/programming will help a lot*
  - Size of final project groups? *4*
  - Will we have notes? *Slides will be posted before the lecture, any notes will be posted after*

No official prereqs, but the following will be useful:

- comfort with programming
  - We'll be using Python (and PyTorch) throughout the class
- comfort with probability, linear algebra, and mathematical notation
- Some familiarity with matrix calculus
- Excitement about language!
- Willingness to learn

Please brush up on these things as needed!

What if I don’t understand what you said

- Case 0: You are talking about unrelated/abstract concepts
  - Usually, I will explain these high-level concepts multiple times
    - These concepts might guide you even after you forget most of materials
- Case 1: Only a few questions (or you disagree)
  - Please raise your hand
  - Look forward to discussions
- Case 2: Some questions
  - Please check Mohit’s corresponding class from last year
    - I will explain more insights for some parts and skip some parts
  - I will also try to record the lecture
- Case 3: Many questions
  - You can consider reading the materials I provide on the course website
    - Please don’t ask many detailed questions in the papers. Like in page x, they do this and that

Previous class videos / material

- Fall 2020: https://people.cs.umass.edu/~miyyer/cs685_f20
- Fall 2021: https://people.cs.umass.edu/~miyyer/cs685_f21/
- Fall 2022: https://people.cs.umass.edu/~miyyer/cs685_f22/
- Spring 2023: https://people.cs.umass.edu/~miyyer/cs685_s23/
- Spring 2024: https://people.cs.umass.edu/~miyyer/cs685/
  - Feel free to use these materials / videos to study!
  - This course will have a lot of overlap with the S24 edition
  - That said, there will be quite a bit of interesting new stuff later in the semester!

Grading breakdown

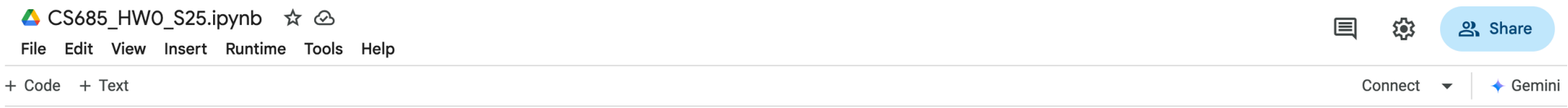
- 5% quizzes
- 30% problem sets (hw0, hw1, hw2)
  - Written: math & concept understanding
  - Programming: in Python
- 25% exam (**in-class exam**)
- 40% final projects (groups of 4)
  - Choose any topic you want
  - Project proposal (10%)
    - If lower than the final, same score
  - Final report / presentation? (30%)
- See more details at https://people.cs.umass.edu/~hschang/cs685/grading.html

Extra credit

- There will be many seminar/job talks related to NLP this semester
  - https://nlp.cs.umass.edu/seminar/
  - The first one is Wed Feb 5, 12pm-1pm
  - Remotely attend up to **five** of these talks (or watch their recordings) and then complete a writeup about each
  - In total, earn up to 3% on top of your final grade

Homework

- Strongly recommend to do the homework by yourself
  - You can use LLMs to help you do the homework
    - Please provide all of your prompts that allow us to reproduce your answer



- Homework 0, CS685 Spring 2025

- Plagiarism
  - If we find that your answers are the same or very similar to those of other people, we might report your behavior
    - e.g., copying from others or from last year's homework

Late Policy

- Everyone will get **three** late days to use for homework assignments or quizzes.
  - After all three late days have been exhausted, no more late submissions will be accepted.
    - No late days for project submissions
- For unforeseen health and personal emergencies, please contact the instructors at cics.685.instructors@gmail.com .
  - Job interviews / other schoolwork are **not** excuses for late homework.

Midterm

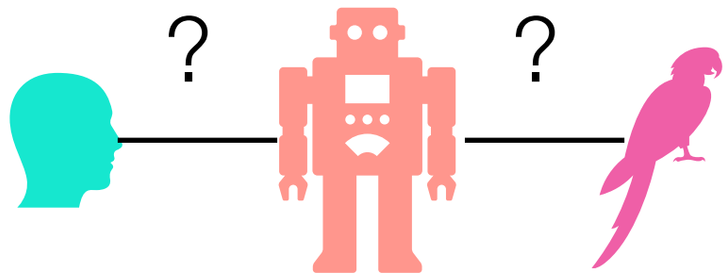
- In-class exam
  - An A4 double-sided note is allowed
    - Although I don’t think you will need one
- Questions would be centered on the classes, quizzes, and homework
  - I plan to have some easy and hard questions
  - You can choose not to come to the class
  - After the midterm, I think whether you come to the class or not won’t affect your score, but we will talk about many practical topics that might be useful for your project or interview

What do you want to learn from this course?

- How to do excellent in my interviews?
- How to judge if NLP/LLM papers make sense?
- What LLMs can do?
  - Why do LLMs work so well?
- What LLMs cannot do?
  - What are the fundamental limitations of LLMs?
    - e.g., Why do LLMs hallucinate
    - How could we overcome them?
  - How far are we toward AGI? Will we all lose our jobs?
    - What are the main barriers?
- What's the difference between LLMs and Humans
  - How should we collaborate with LLMs?

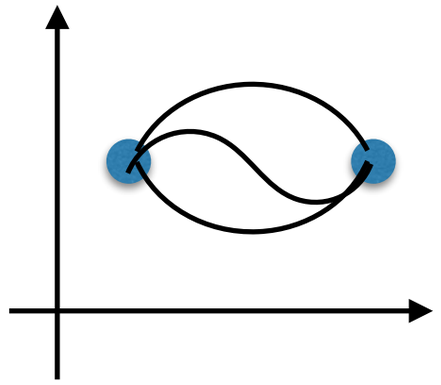
~~Facts~~ Perspectives

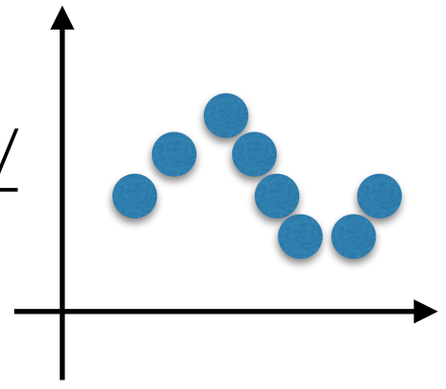
- Many materials are based on our **interpretation/ perspective** of the latest findings
  - Or even just insights
  - No good textbook on this
- Perspectives are debatable
  - Could be even controversial
    - You often see lots of debate between experts
      - You can learn different perspectives in talks, Mohit's videos, or just ask ChatGPT
- Uncertainty could lead to creativity
  - Challenge me! Just like I challenge some mainstream perspectives



NLP/AI History

- Why NLP/AI development history?
- What will you study if GPU does not exist?
- What is good NLP/AI research?
  - For industry, good performance
  - Structure vs Data
    - Interpolation?
    - Making training data closer to testing data?
      - e.g., Test-time training?
    - Bitter Lesson (http://www.incompleteideas.net/IncIdeas/BitterLesson.html)
- Why LMs need to be large?





natural language processing

natural language processing

languages that evolved naturally through human use

e.g., Spanish, English, Arabic, Hindi, etc.

natural language processing

supervised learning: *map text to* **X**
unsupervised learning: *learn* **X** *from text*
generate *text from* **X**

Levels of linguistic structure

Discourse

Semantics

Syntax: Constituents

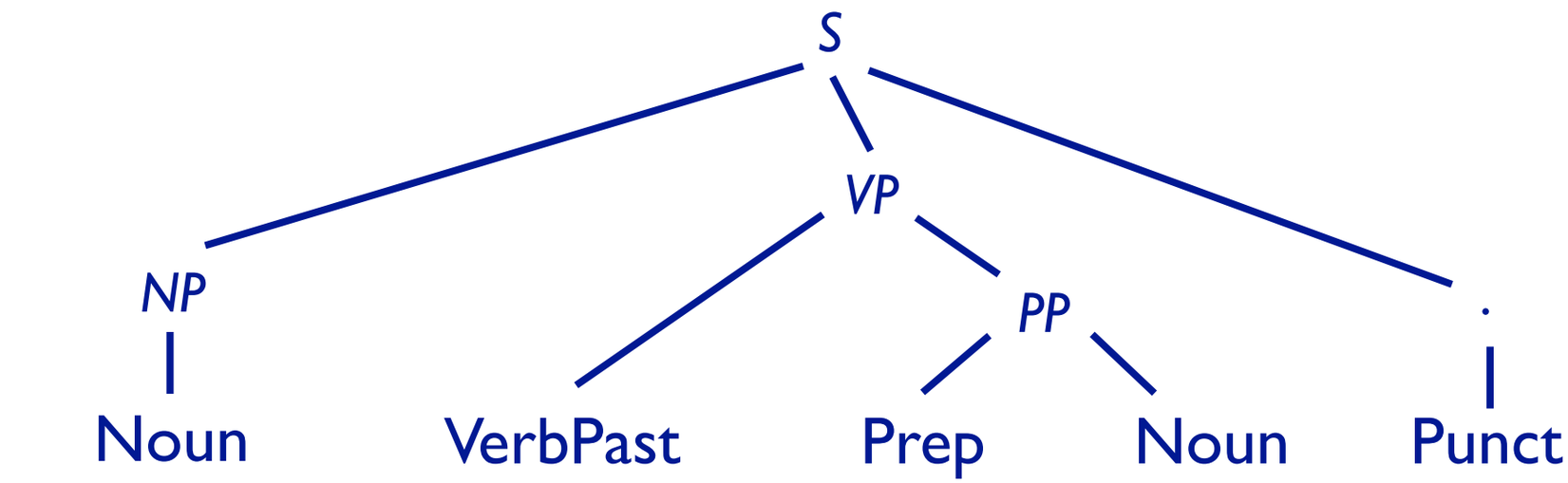
Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)



Alice talked to Bob .

talk -ed [VerbPast]

Alice talked to Bob.

supervised learning: given a collection of labeled examples (where each example is a text *X* paired with a label *Y*), learn a mapping from *X* to *Y*

Example: given a collection of 20K movie reviews, train a model to map review text to review score (*sentiment analysis*)

self-supervised learning: given a collection of *just text,* without extra labels, create labels out of the text and use them for *pretraining* a model that has some general understanding of human language

- **Language modeling**: given the beginning of a sentence or document, predict the next word
- **Masked language modeling**: given an entire document with some words or spans masked out, predict the missing words

How much data can we gather for these tasks?

transfer learning: first *pretrain* a large self-supervised model, and then *fine-tune* it on a small labeled dataset using supervised learning

Example: pretrain a large language model on hundreds of billions of words, and then fine-tune it on 20K reviews to specialize it for sentiment analysis

in-context learning: first *pretrain* a large self-supervised model, and then *prompt* it in natural language to solve a particular task without any further training

Example: pretrain a large language model on hundreds of billions of words, and then feed in “what is the sentiment of this sentence: <insert sentence>”

Language models

api.together.xyz

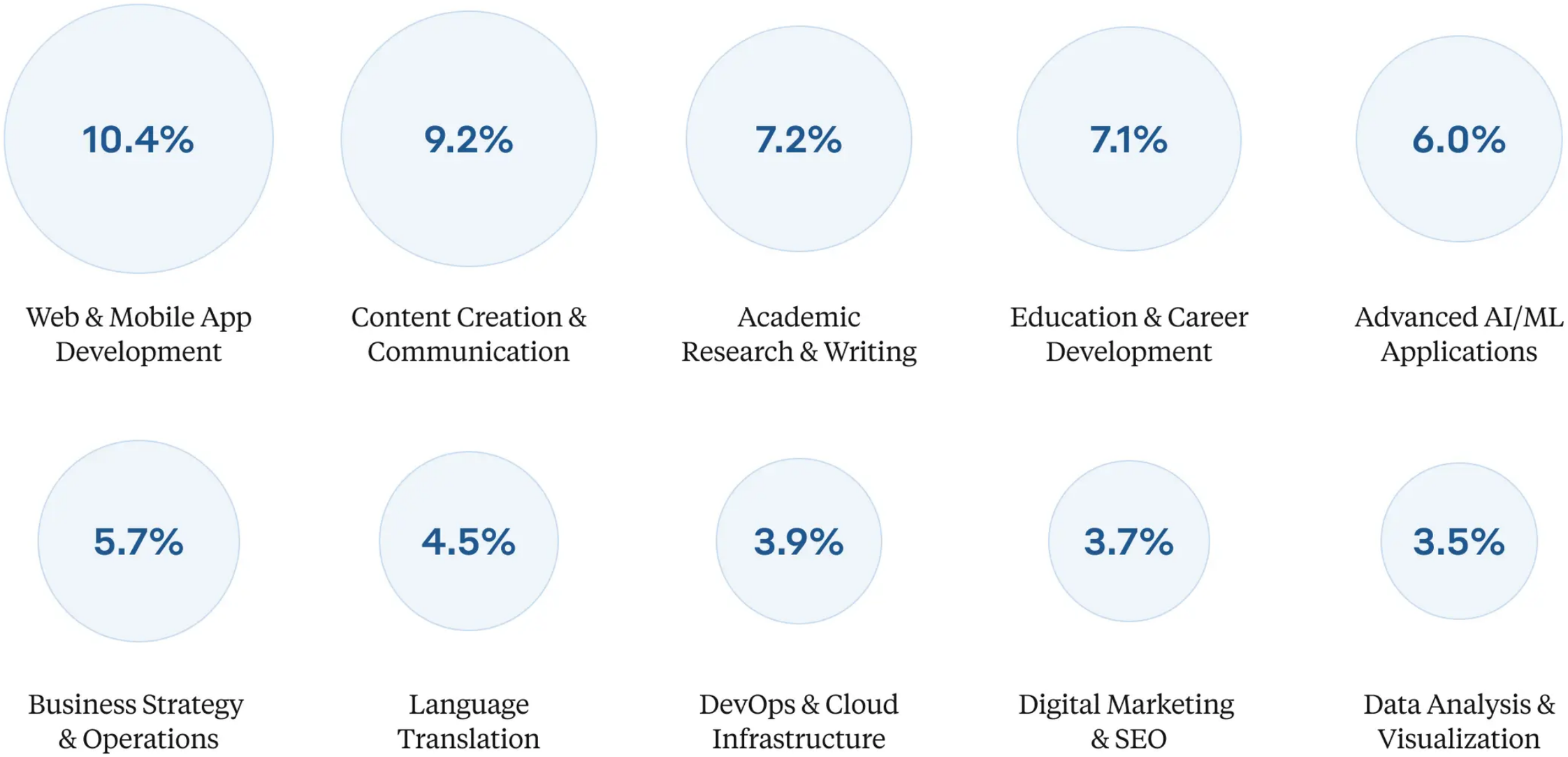
What are people using LLMs for?

Figure 3: Topic distribution of 100K sampled conversations. Manual inspection of cluster centroids

[29] https://arxiv.org/pdf/2309.11998.pdf

What are people using LLMs for?

Top use cases on Claude.ai



https://www.anthropic.com/research/clio

Rough list of topics

- This year, the language models do not change too much, so I will use most of the materials from Mohit
  - I will add some content while de-emphasizing others
- **Background**: language models and neural networks
- **Models**: Transformers
  - RNN > BERT > GPT3 > ChatGPT > today's LLMs
- **Tasks**: text generation (e.g., translation, summarization), classification, retrieval, etc.
- **Data**: annotation, evaluation, artifacts
- **Methods**: pretraining, finetuning, preference tuning, prompting, reasoning?
- **Notice**: NLP != LLMs

Final projects

Timeline

- All groups should be formed by **2/14**
  - https://forms.gle/PKvJRxZkUMgFrkVG8
  - Groups of 4, either form them yourselves and tell us, or we will randomly assign you on 2/15
- Only two deliverables:
  - project proposal: 3+ pages, due **3/7**
  - final report/code: 8+ pages, due last day of classes (**5/9**)
- Almost completely open-ended!
  - All projects must involve natural language data
  - We strongly recommend a significant coding component of every project

Project

- Either *build* natural language processing systems, or *apply* them for some task.
- Use or develop a dataset. Report empirical results or analyses with it.
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task

Resources and Grades

- Resources are the same as last year
  - Won’t provide GPU resources
  - We only have $500 budget for API calls
  - Money for textbook -> API credits
- Some directions that require less resources
  - Evaluation
  - Applications
  - Dataset building
  - Re-implementation
  - Interpretation/Visualization
  - Prompt engineering
  - Survey (not recommended)
- Good Project Criteria
  - Effort or Novelty or Usefulness or Implication

Sample projects

Taller, Stronger, Sharper:
Probing Comparative Reasoning Abilities of Vision-Language Models

Examining Medical Narratives of Eating Disorder Recovery on Reddit

Replication of TagRec, a Hierarchical Taxonomy Tagging Model

Learning Schematic and Contextual Representations for Text-to-SQL Parsing

Syllamo: Generating Keyword Mnemonics for Vocabulary Acquisition

Formulating a proposal

- What is the ***research question***?
- What's been done before?
- What experiments will you do?
- How will you know whether it worked?
  - If data: held-out accuracy
  - If no data: manual evaluation of system output. Or, annotate new data

Feel free to be ambitious (in fact, we explicitly encourage creative ideas)! Your project doesn't necessarily have to "work" to get a good grade.

NLP Research

- All of the best NLP publications are open access!
  - The ACL Anthology (https://aclanthology.org/) contains papers from all of the top NLP conferences (e.g., ACL, EMNLP, NAACL) spanning many decades
  - Machine learning conferences (ICLR, NeurIPS, ICML)
  - Check out arXiv CS-CL (https://arxiv.org/list/cs.CL/recent) for the most recent papers!
  - This is a fast-moving field, so follow NLP researchers on Twitter for discussion on the latest advances
- Use Google Scholar and Semantic Scholar to search for relevant papers

Broader ideas

https://2024.aclweb.org/calls/main_conference_papers/#call-for-main-conference-papers

https://colmweb.org/cfp.html

- At homework 0, we will ask you to summarize a paper.

An example proposal

- Introduction / problem statement
- Motivation (why should we care? why is this problem interesting?)
- Literature review (what has prev. been done?)
- Possible datasets
- Evaluation
- Tools and resources
- Project milestones / tentative schedule

Be on the lookout for

- **HW0:** released today, due 2/14 (11:59pm) on Gradescope
- **Final project:** Organize into groups of 4 by 2/14
  - https://forms.gle/PKvJRxZkUMgFrkVG8
- **Final project:** project proposal due 3/7

Having issues accessing Piazza/Gradescope/videos? Email the instructors account!