Syllabus | Language Models as Cognitive Models

Course number: DSC 291 C00
Instructor: Professor Alex Warstadt
TA: Aditya Yadavalli
Dates: Every Monday, Wednesday, and Friday, March 30 – June 5
Time: 1:00-1:50 pm
Location: Pepper Canyon Hall, Room 121 (in-person only)
Alex's office hours: Mondays 10:30-12, Applied Physics and Mathematics #4220
Aditya's office hours: Monday 16:00-17:00, Zoom: [link]
Slides: [link]

1 Course Overview

We can all see how Large Language Models have transformed technology, the classroom, and society, but what about their impact on how we understand our own minds? In this class, we learn how LLMs are revolutionizing the linguistic and behavioral sciences by serving as test subjects and model organisms. We cover the latest research developments in three primary areas: (1) BabyLMs – training small language models from scratch to simulate human language acquisition, (2) Language Processing – predicting human neural and behavioral responses to language processing using LLMs predictions, and (3) Mechanistic Interpretability – exploring techniques from machine learning to unveil the inner workings of connectionist models. In addition to lectures providing an overview of research areas, students in teams will lead weekly reading group discussions in which we engage closely with recent landmark papers. Throughout, we will revist the key philosophical question underlying this endeavor: What can we learn about ourselves from an intelligence that is so different from our own?

2 Prerequisites

There are no formal prerequisites. You must know how basic Python. Other concepts/frameworks that will be relevant but not taught in detail include:

- Some familiarity with libraries like Pandas, Matplotlib, Transformers, Pytorch, Numpy.
- Introductory-level knowledge of cognitive science and linguistics.
- Basic technical understanding of neural networks, language models, machine learning.
- Basic use of supercomputer clusters, shell scripting, kubernetes, jupyter, and docker.

The above are **not** required prior knowledge. Many of these things will be covered briefly, either directly or indirectly, many can be easily learned on one's own, and the use of code generation models is permitted to assist with these tools.

3 Schedule

The schedule below is subject to change.

Week	Session	Date	Topic	Activities	Suggested Readings
Week 1	1	Mar 30	**Intro**: Welcome		
	2	Apr 1	**Intro**: Background on language models		Jurafsky and Martin, 2025. "Speech and Language Processing (3rd ed.)." Chapter 9, Chapter 10
	3	Apr 3	**Intro**: Linguistics & Cognitive Science		
Week 2	4	Apr 6	**Intro:** Theoretical Discussion on LMs in Linguistics and Cogsci		1. Futrell & Mahowald, 2025. "How Linguistics Learned to Stop Worrying and Love the Language Models." 2. Millière, 2024. "Philosophy of cognitive science in the age of deep learning."
	5	Apr 8	**Intro (Reading Group 1):** Theoretical Discussion on LMs in Linguistics and Cogsci		1. Piantadosi, 2023. "Modern language models refute Chomsky's approach to language." 2. Kodner et al., 2023. "Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023)."
	6	Apr 10	**Baby LMs**: Background on language acquisition		
Week 3	7	Apr 13	**Baby LMs**: Background on language acquisition		
	8	Apr 15	**Baby LMs (Reading Group 2):** Evaluating LMs		1. Warstadt et al., (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English. 2. Hu et al., (2025). "What Can String Probabilities Tell Us About Grammaticality?"
	9	Apr 17	**Baby LMs**: Data-efficient pretraining		
Week 4	10	Apr 20	**Baby LMs**: Controlled rearing	Project groups due	1. Wilcox et al., 2025. "Bigger is not always better: The importance of human-scale language modeling for psycholinguistics." 2. Eldan & Li, 2023. "TinyStories: How

					Small Can Language Models Be and Still Speak Coherent English?"
	11	Apr 22	**Baby LMs (Reading Group 3)**: Corpus editing and controlled rearing		1. Misra & Mahowald, 2024. Language Models Can Learn Rare Phenomena from Less Rare Phenomena. 2. Patil & Jumelet, et al., 2024. Filtered Corpus Training (FiCT) Shows that Language Models Can Generalize from Indirect Evidence.
	12	Apr 24	**Baby LMs**: Controlled rearing		Kallini et al., 2024. "Mission: Impossible Language Models."
Week 5	13	Apr 27	**Baby LMs:** Developmentally plausible pretraining (word learning & vision)		1. Chang & Bergen, 2022. "Word Acquisition in Neural Language Models" 2. Zhuang et al., 2024. "Lexicon-Level Contrastive Visual-Grounding Improves Language Modeling"
	14	Apr 29	**Baby LMs (Reading Group 4)**: Developmentally plausible pretraining (vision & audio)		1. Vong et al., 2024. "Grounded language acquisition through the eyes and ears of a single child." 2. Lavechin et al., 2025. "Simulating Early Phonetic and Word Learning Without Linguistic Categories."
	15	May 1	**Baby LMs**: Developmentally plausible pretraining (interaction)	HW1 (Part 1) due: BabyLMs	1. Lazaridou et al., 2020. "Multi-agent communication meets natural language: Synergies between functional and structural language learning." 2. Nikolaus & Fourtassi, 2023. "Communicative Feedback in Language Acquisition."
Week 6	16	May 4	**Language Processing**: Reading times and Surprisal theory		1. Smith & Levy, 2012. "The effect of word predictability on reading time is logarithmic."
	17	May 6	**Language Processing (Reading Group 5)**: Reading times and Surprisal theory		1. Wilcox et al., 2023. "Testing the Predictions of Surprisal Theory in 11 Languages." 2. Oh & Schuler, 2023. "Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?"
	18	May 8	**Language Processing**: Reading times and Surprisal theory	HW1 (Part 2) due: BabyLMs	1. Kuribabyashi et al., 2025. "Large Language Models Are Human-Like Internally."
Week 7	19	May 11	**Language Processing**: Neurological response data	Project Proposals	1. Michaelov et al., 2024. "Strong prediction: Language model surprisal

				Due	explains multiple N400 effects." 2. Hosseini et al., 2023. "Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training."
	20	May 13	**Mechanistic Interpretability (Reading Group 6)**: Probing		1. Tenney et al., 2019. "BERT rediscovers the classical NLP pipeline." 2. Hewitt & Manning, 2019. "A structural probe for finding syntax in word representations."
	21	May 15	**Mechanistic Interpretability**: Subspaces and causal intervention		1. Ravfogel et al., 2020. "Null it out: Guarding protected attributes by iterative nullspace projection." 2. Finlayson et al., 2021. "Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models."
Week 8	22	May 18	**Mechanistic Interpretability**: Circuits and causal intervention		1. Hu et al., 2025. Between circuits and Chomsky: Pre-training on formal languages imparts linguistic biases 2. Fehlauer et al, 2025. "Convergence and Divergence of Language Models under Different Random Seeds"
	23	May 20	**Mechanistic Interpretability (Reading Group 7)**: Circuits and causal intervention	HW 2 due: Language Processing & Mechanistic interpretability	1. Mueller et al., 2025. "Sparse feature circuits: Discovering and editing interpretable causal graphs in language models."
	24	May 22	Buffer		
Week 9	25	May 25	NO CLASS, Memorial Day		
	26	May 27	Buffer		
	27	May 29	Buffer		
Week 10	28	June 1	Project Presentations		
	29	June 3	Project Presentations		
	30	June 5	Project Presentations		
Week	NO	Thur	Papers due		

11 (exams)	CLASS	day June 11			

Bonus topics

- Inductive bias
- Simulating human behavioral response data
    - Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design
- Theory of mind
- Multiagent interaction

4 Grading

Participation: 10 points
Reading group leading: 15+15 points
Homework: 25+25 points
Project proposal: 10 points
Project presentation: 20 points
Project paper: 30+20 points
Attendance: 0 points or 30 points

4.0 Homeworks

Homework 1 (Part 1): DSC 291 -- HW1 -- Evaluating BabyLMs.ipynb

4.1 Reading Group

Every student will lead a reading group session. Sessions will be led in assigned groups of 2-5, depending on enrollment. Groups will meet with the instructor one week before to review the paper(s) and presentation content. Grades will be out of 30 and reflect both shared and individual work (see rubric below).

Reading Group Grades will be determined as follows:

- Presentation: 15 points
    - Preparation, comprehension, clarity, organization (10 points)
    - Individual contribution (5 points): Submit a group contribution statement
- Individual write-up: 15 points
    - 2 pages, single-spaced, prose (no bullet points)
    - Paper summary (5 points): Big question, hypothesis, methods, conclusion
    - Critique (5 points): What did you like/dislike about the paper?
    - Future work (5 points): How would you extend this work? What questions are left open?

Note: See policy of use of generative AI in writing (5.1).

4.2 Final project

The final project will be done in groups (based on enrollment) and should be an ACL-style paper (8 pages, not including references) + an individual contribution statement. Projects must relate to language models and cognitive modeling broadly construed (all topics must be approved), but besides that there are no restrictions. However, some sample topics are listed at the bottom of this document.

Timeline:

- **Week 3**: Form project groups of 2-5 (size depends on class enrollment)
- **Monday of Week 4**: Project groups due
- **Weeks 4 & 5:** Meet with instructor & TA (~30 mins) to discuss project.
- **End of Week 5**: Submit project proposal.
    - Target: 1000-2000 words
    - Motivation / research question
    - Brief lit review
    - Overview of methods (data/models) & planned analysis
    - Discussion of possible outcomes
    - (optional: Possibilities for future extensions)
- **Week 10**: Project presentations.
- **Thursday, June 11**: Papers due

Grades

- Proposal (10 points):
    - Meet with the instructor to discuss project (5 points)
    - Written proposal (5 points)
        - Target: 1000-2000 words
        - Motivation / research question
        - Brief lit review
        - Overview of methods (data/models) & planned analysis
        - Discussion of possible outcomes
        - (optional: Possibilities for future extensions)
- Paper (30 points):
    - The whole group submits one final paper (30 points)
    - Paper content
        - Aim for an 8-page ACL-style paper. You can use the ACL template.
        - Typical structure: Intro, Related work, Methods, Results, Discussion.
            - There is no *required* structure or content. Just write the best paper you can.
        - You may submit supplemental materials (additional plots, data, code) alongside your paper.
    - Include a brief shared contribution statement (just a few bullet points is good)
- Individual contribution (20 points):
    - Each individual submits a ~1-page contribution statement
    - Describe in a bit more detail your individual contribution.
    - If everyone in the group makes contributions of similar size and value, the grades will all be

proportional to the shared grade.

4.3 Late assignments

- You have a budget of **2 late days** for individual work for the entire quarter (not per assignment).
- Late group work is not accepted.
- Late work will not be accepted if you've already used 2 late days. There is not much individual work, so this is non-negotiable. The only exception is for incompletes.

4.4 Attendance policy

This course allows you to **opt in to mandatory attendance**.

- After the two weeks of class, you will be asked to choose whether you want attendance to count towards your grade.
- If you choose mandatory attendance:
    - Your attendance grade starts at 30/30.
    - Attendance will be taken by answering a discussion question in the first 5 minutes of class.
    - 1 point is subtracted for every class session you skip or are >5 minutes late.
    - You have 2 sick days for free, otherwise a doctor's note is required
    - A total of 30 possible attendance points is added to the 170 assignment points, meaning your final grade is out of 200.
- If you do not choose mandatory attendance:
    - Attendance has no (direct) effect on your grade.
    - Your final grade is out of 170.

4.5 Final grades

If you're taking the course for a letter grade, your grade will be determined according to the scale below.

- Note that the number on the right-hand side of the range is not included in that range (except for 100%): that is, an "A-" ranges from 90% all the way to 92.99% but does not include 93%.
- There is no rounding of percentages.

Percentage	**Grade**
97% –	A+
93% – 97%	A
90% – 93%	A-
etc.	
60% – 70%	D
0% – 60%	F

5 Course Conduct

5.1 Academic Integrity

When working in a group, all members of the group must contribute as equally as possible. However, this is not always realistic, so groups will write shared and individual contribution statements and self-assessments. All members of the group must agree to the shared contribution statement (if there are absentee group members, please reach out to the instructor or TA). UCSD's academic integrity policies are available here.

The use of AI assistants in your work is permitted ONLY in some specific ways:

- **Disclosure**: All use of LLMs must be disclosed and described in submitted work.
- **Code**: You may use LLMs to debug and write code; however, code must be reviewed and tested.
- **Writing**: You may use LLMs to correct grammar and style issues; however, you may not use LLMs to generate entire sentences or paragraphs. All writing must be in your own voice.
- **Research**: You may use LLMs to suggest prior literature or research methods; however, you must consult the primary sources yourself and properly cite any methods you use.

Any unauthorized or undisclosed uses of LLMs could result in disciplinary action.

5.2 Personal Conduct

The classroom is a place of learning but it is also a community. We follow UC San Diego Principles of Community. When you enter the classroom, you agree to respect the individual dignity and rights of your fellow community members regardless of race, ethnicity, sex, gender identity, age, disability, sexual orientation, religion, and political beliefs.

5.3 Screen Policy

Computers, tablets, and phones are permitted for course-related activities only, e.g. note-taking or referring to a paper. Laptops and handwriting are both fine for notetaking, but laptops allow for more distractions to yourself and others (see Voyer et al., 2022).

6. Project Ideas

Some project ideas (with potential mentors) are listed in this Google Doc.

Below is a sample of some general project ideas. We are happy to elaborate on any of them. You do NOT have to work on one of these topics, they are just to help suggest what some reasonable projects might look like.

- **Training BabyLMs from scratch**

    - Innovations in data-efficient or human-like architectures
    - Reproducible data-efficient techniques
    - Testing learnability claims: Corpus editing and controlled training
    - The “Motherese Hypothesis”: The effect of prosody on syntax learning
    - Multimodality: Training with images and text
    - Critical periods in language acquisition
- **Evaluations**
    - New minimal pair benchmarks
        - Very easy English syntactic contrasts for BabyLMs
        - Low resource languages (NOT English or Mandarin)
        - Discourse/Pragmatics
        - Commonsense reasoning
        - Text & prosody benchmark
    - Benchmarking human-like word learning (in multimodal models)
    - Evaluations for human-like behavioral judgments
    - Evaluating learning trajectories for diverse evaluations
- **Processing**
    - Predicting reading times / N400s from model internal layers
    - Evaluating information density from model internal layers
    - Uniform information density hypothesis in long texts and discourse
    - Decoding linguistic features from neural response data
- **Interpretability**
    - Tracking concepts or circuits throughout training
    - Searching for concepts or circuits for grammatical features or behavioral responses
    - Measuring the effect of ablating concepts or circuits