24.S90 Exploring Language Technologies
aka Demystifying LLMs

Massachusetts Institute of Technology
Fall 2023, Tuesdays & Thursdays, 9:30–11:00am, 56-167

Teaching staff

Instructor:	Dr. Hadas Kotek (she/her)
Email:	hkotek@mit.edu
Office hours:	By appointment (Calendly link)
Course website:	Canvas Link

Description

This course explores the abilities and limitations of language models, focusing on state of the art tools such as ChatGPT. LLMs possess impressive language abilities, but they also occasionally fail in unpredictable ways. Our goal in this class will be to map the abilities and limitations of these models, focusing on complex reasoning and language abilities. We will attempt to discover systematicity in the models failures and to understand how they relate on the one hand to how the prompt is formulated and what we believe the training data and model architecture to be, and on the other hand how humans perform on the same tasks and how children acquire this knowledge. Along the way, we will learn about the development of language technologies and their capacities over time, as well as the state of the art linguistic theories that explain the phenomena of interest. Well ask ourselves whether LLMs appear to resemble humans in their approach to language and reasoning, and what this means for how we should understand what LLMs actually do (and how humans can and should interact with them).

The course also aims to develop skills and materials useful for non-academic jobs, such as behavioral study design, fundamentals of prompt engineering, knowledge engineering, and data annotation. Class assignments will build a scaffolding for a final paper or experiment to be presented at the end of the semester. No prior knowledge of programming languages or computational linguistics is expected.

Course expectations

1. **Attendance and participation:** I expect active participation from all members of the class—enrolled for credit or otherwise.

2. **Readings:** You are expected to do the readings prior to the class in which they will be discussed. Readings will be limited to 1–2 required readings per week.

3. **Homework** The final assignment for this class will be a paper testing the behavior of some LLM on some empirical linguistic phenomenon. I will assign 2–3 scaffolding

assignments during the semester to help you build toward this goal. These assignments will be optional but highly recommended to keep you on track for your final project.

4. **Final Presentations:** During the last week of classes, students will give a brief presentation of their final paper topic.

Rules of note

- **Talk to me:** I am committed to helping you succeed in this course. Please don't hesitate to contact me. For questions about content, homework, or readings, send me an email or set up an appointment with me.

- **Cooperation:** You may discuss assignments with other students. However, you must always submit your own write-up, and you must list the students who you worked with on your assignment.

- **Integrity:** The use of others' ideas or expressions without citation is **plagiarism**, and will not be tolerated. You must declare all sources in submitted work. Citations don't need to be in any particular format, but they have to be there. This policy also applies to the use of Large Language Models in the course of researching or writing up an idea. If you relied on an LLM in your work, please describe it accordingly.

- **Participation:** As the instructor, I will be doing a large portion of the talking in class, but the course will be vastly improved by you, the students, sharing your ideas and asking your questions. If you have a question, there is probably at least one other person with the same question. Ask it; others will be grateful you did.

- **Disabilities:** MIT is committed to the principle of equal access. Students who need disability accommodations are encouraged to speak with Disability and Access Services (DAS), prior to or early in the semester so that accommodation requests can be evaluated and addressed in a timely fashion. If you have a disability and are not planning to use accommodations, it is still recommended that you meet with DAS staff to familiarize yourself with their services and resources. Please visit the DAS website for contact information.

  If you have already been approved for accommodations, please inform the instructor as soon as possible.

- **Diversity and inclusion:** I am committed to making this class a safe and welcome space for all participants. If there are any concerns you wish to raise, please reach out to me directly, or via the anonymous feedback survey link provided here. As a participant of this course, I ask that you strive to maintain a respective environment and honor the diversity of your fellow classmates. For additional resources, please see:

  1. `https://hr.mit.edu/diversity-equity-inclusion`
  2. `https://studentlife.mit.edu/impact-opportunities/diversity-inclusion`
  3. `https://linguistics.mit.edu/diversity-statement/`

Course plan

The plan may be adjusted based on how the discussion develops and the participants preferences. Topics 3–5 can each be as large or small as we want them to be. We will likely interleave topics from the different areas depending on our progress and interests.

Topic 1 A brief history of the development of language technologies, NLP tasks, and models.

Topic 2 Language data
  - data sources, data sparsity
  - *practical topic 1: annotation*

Topic 3 Do LLMs 'know' language: an investigation of some linguistic phenomena

  - compositionality
  - binding and control
  - logical and pragmatic inferences
  - dialects and variation
  - model evaluation, benchmarks, and datasets
  - *practical topic 2: behavioral experiment design*

Topic 4 Ethics and safety

  - harm, bias and stereotypes, toxicity
  - misinformation and disinformation
  - hallucinations
  - data concerns I: who owns the training data
  - data concerns II: the work of annotators
  - *practical topic 3: constructing a benchmark*

Topic 5 Current affairs

  - are LLMs sentient, "AGI"
  - the environmental cost of using LLMs
  - LLMs in current real-world applications

Holidays and days of note:

  - September 6, 2023: First day of class
  - October 10: Student holiday
  - November 6–10: Hadas away at a conference
  - November 23: Thanksgiving
  - December 13: Last day of class

Some resources

Papers and other readings

Most readings in this class will be in the form of articles, blog posts, and slides. These materials will all be uploaded to Canvas.

As the field of LLMs is moving at an especially fast pace these days, we may choose to expand or replace some topics or readings with alternative topics as the semester progresses.

Books

No book is required for this course. If you are interested in pursuing topics in greater depth in this course, you may supplement your reading for class with these suggested books:

- Jurafsky and Martin (2023). Speech and Language Processing, Ed. 3.
- Eisenstein (2019). Introduction to Natural Language Processing.
- Bender (2013). Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax.
- Bender and Lascarides (2019). Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics.
- Gorman and Sproat (2021). Finite-State Text Processing.