24.S90 Demystifying Large Language Models

Massachusetts Institute of Technology
Fall 2024, 10:00am–12:30pm, 32-D461

Teaching staff

Instructor:	Dr. Hadas Kotek (she/her)
Email:	`hkotek@mit.edu`
Office hours:	By appointment (T/W mornings preferable)
Course website:	Canvas Link

Formal description

This course explores the abilities and limitations of language models, focusing on state of the art tools such as GPT-4, Gemini, and LLaMA. Large Language Models (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. We will additionally entertain the various costs associated with the deployment and use of LLMs, be they due to privacy breaches, environmental costs, security risks, copyrights abuses, the environment, or the entrenchment and amplification of biases and stereotypes at scale. Along the way, we will investigate 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. We'll ask ourselves whether it is reasonable to conclude that the LLMs use a similar sort of approach as humans do to complex language reasoning, and what this means for how we should understand what LLMs actually do (and how humans can and should interact with them).

Alternative description

There are several papers on LLMs that I'd like to write, but I don't have the time or capacity to do it all alone. So depending on students' interests and goals, we may choose to do less reading and more experimenting and group work. If so, we may focus on one topic or split into groups and spend some class sessions on exploration and progress reports.

(Topics will be shared with class participants.)

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 per week.

   Some weeks, I may ask you to write short (>1 page) responses to the readings, including (i) a brief summary of the argumentation in the article and (ii) a question that occurred to you while you were reading. These summaries will be due the day before the relevant class: I will use the questions that you raise to guide our class discussion.

3. **Class presentation** Enrolled students will be asked to lead the discussion during one week of class. This could entail leading the discussion for that week, presenting on readings on a topic of your choosing, or presenting on your final paper/project.

4. **Final Paper:** Enrolled students will submit a final paper at the end of the semester. Possible topics to be discussed in class later in the semester. Group work on final projects is allowed and encouraged. Groups can submit a single paper, where the contributions of each group member are clearly laid out. In that case, a group presentation in class will be required.

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 any class content or requirements, send me an email or set up an appointment with me. I work on campus on Tuesday/Wednesday/Thursday and am occasionally also around on Mondays.

- **Cooperation:** Collaboration and discussion with other class members is allowed and encouraged. However, please list the students who you worked with on any work you submit in this class.

- **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 this anonymous feedback survey link. 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 2–4 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. Connecting modern NLP and linguistics/cognitive science

- compositionality
- binding and control
- logical and pragmatic inferences
- dialects and variation
- model evaluation, benchmarks, and datasets
- (your topic here)
- *practical topic 1: annotation, behavioral experiment design*

Topic 3. 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 2: constructing a benchmark*

Topic 4. Current affairs

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

Holidays and days of note:

- September 5, 2024: First class of the semester
- October 4: Add date
- October 10: Hadas away at a conference
- November 20: Drop date
- November 28: Thanksgiving
- December 5: Last class of the semester

NLP conference deadlines:

- Oct 15, 2024: North American Chapter of the Association for Computational Linguistics (NAACL)
- Likely ~Feb 2025: Association for Computational Linguistics (ACL), ICML, KDD
- Likely ~June 2025: NeurIPS, EMNLP

Some Readings

Credit for large portions of this list: joint effort with Katrin Erk, Naomi Feldman, Dan Jurafsky, Tal Linzen, Zoey Liu, Isabel Papadimitriou.

Books

No book is required for this course. If you are interested in pursuing technical NLP-related topics in greater depth, 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.

Papers and other readings

0. Survey articles

- Pater (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. Language 95(1).
  - Survey article on 60 years of development in Linguistics and in Neural Nets.
- Linzen (2018). What can linguistics and deep learning contribute to each other? Response to Pater. Language.
- Boleda (2020). Distributional Semantics and Linguistic Theory. Annual Review of Linguistics. Annual Review of Linguistics, 6(1), 213-234.
  - Survey article on lexical semantics.
- Linzen and Baroni (2021). Syntactic structure from deep learning. Annual Reviews of Linguistics.
- Baroni (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. Book chapter.

1. Meta-commentary on NLP and LLMs

- Bender et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proceedings of FAccT.
- Jernite et al. (2022). Data Governance in the Age of Large-Scale Data-Driven Language Technology
- Atari et al. (2023). Which humans?.
- Bubeck et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.
- McCoy et al. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve.
- Moro et al. (2023). Large languages, impossible languages and human brains. Cortex.

- Van Rooij et al. (2023) Reclaiming AI as a theoretical tool for cognitive science.
- Gebru and Torres (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday, 29(4).
- Grieve et al. (2024). The Sociolinguistic Foundations of Language Modeling.
- Hicks et al. (2024). ChatGPT is bullshit. Ethics and Information Technology.
- Kotek, Hadas. 2024. Language and technology. Routledge Handbook of Linguistics.
  - Survey article with a focus on LLMs, their uses, and ethical considerations.
- Opitz et al. (2024). Natural Language Processing RELIES on Linguistics.

2. Blog posts and news media articles

- Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. Time Magazine, 2022.
- How the AI industry profits from catastrophe. MIT Technology Review.
- Stochastic Parrots Day reading list (March 17, 2023); community-generated list.
- ChatGPT is amazing and everything that's wrong with the world (blog post, 2023)
- China's AI boom depends on an army of exploited student interns. Rest of World, 2023.
- Google's AI Chatbot Is Trained by Humans Who Say They're Overworked, Underpaid and Frustrated. Bloomberg, 2023.
- OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. The Verge, 2023.

3. LLMs and philosophy of language

These papers rehash arguments from philosophy of language under the perspective of whether LLMs can be said to refer. They provide a nice summary of different positions on reference.

- Bender and Koller (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of ACL 2020.
- Merrill et al. (2021). rovable limitations of acquiring meaning from ungrounded form: What will future language models understand?. ransactions of the Association for Computational Linguistics 9.
- Piantadosi and Hill (2022). Meaning without reference in large language models.
- Mandelkern and Linzen (2024). Do Language Models' Words Refer?. Computational Linguistics.
- Lederman and Mahowald (2024). Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs.
- Baggio and Murphy (2024). On the referential capacity of language models: An internalist rejoinder to Mandelkern & Linzen

4. Language learning/learnability

- McCoy et al. (2020). Universal linguistic inductive biases via meta-learning. Cognitive Science Society.
  - A proof of concept idea for using meta-learning to give the model a linguistic inductive bias (by generating an appropriate distribution of synthetic languages).
- Portelance et al. (2024). Learning the meanings of function words from grounded language using a visual question answering model. Cognitive Science 48:5.

- Kodner et al. (2023a). Re-Evaluating the Evaluation of Neural Morphological Inflection Models. In Proceedings of the 45th Annual Conference of the Cognitive Science Society (CogSci), 3259-3267.
- Constantinescu et al. (2024). Do Language Models Have a Critical Period for Language Acquisition?

5. Lexical semantics

- Chronis and Erk (2020). When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. Proceedings of CoNLL 2020.
  - Suggests that type-level semantic similarity judgments are sensitive to polysemy, and proposes representing a lemma using LLM token clusters even for type-level similarity judgments.
  - Finds that LLMs can differentiate between taxonomic similarity and topical similarity.
- Petersen and Potts (2023). Lexical Semantics with Large Language Models: A Case Study of English break. Proceedings of EACL.
  - Comparing a theoretical analysis of different senses of "break" to what an LLM can see.
  - To what extent can LLMs be used as data for lexical semantics?
- Erk and Chronis (2022). Word Embeddings are Word Story Embeddings (and That's Fine). Chapter in *Algebraic Structures in Natural Language.*
  - Against viewing LLMs as "compact corpus"; In a small qualitative study, we find the clusters to be sensitive to narrative schemas.
  - This matches the Potts' hypothesis (2019) that what we get from LLMs is "a record of utterances rather than idealized linguistic objects".
- Katinskaia and Yangarber (2024). Probing the Category of Verbal Aspect in Transformer Language Models. Proceedings of EMNLP 2022.
  - Do LLMs encode verbal aspect in Russian?
  - With LLM as "compact corpus", one can analyze how the LLM perceives aspectual features to change in different context.

6. Syntax

- Potts (2023). Characterizing English Preposing in PP Constructions
  - How can humans and LMs learn the constraints around the PiPP construction ("Happy though we were that..."), even though the construction is quite rare and gets quite complicated?
  - This paper is cool because it is very clearly a paper with a linguistics concern, but it casually uses LMs as one part of a diverse evidence base.
- Yedetore et al. (2023). Aditya Yedetore, Tal Linzen, Robert Frank & R. Thomas McCoy (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. Proceedings of ACL.
- Wilcox et al. (2023). Using Computational Models to Test Syntactic Learnability. Linguistic Inquiry.

  - Do LMs learn island constraints? Clearly addresses innateness questions.
- Papadimitriou and Jurafsky (2023). Injecting structural hints: Using language models to study inductive biases in language learning
- Papadimitriou et al. (2021). Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT

7. Syntactic Typology

- Hahn and Xu (2022). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. PNAS.
- Hahn et al. (2020). Universals of word order reflect optimization of grammars for efficient communication. PNAS.
- Dyer et al. (2020). Predicting cross-linguistic adjective order with information gain. Findings of ACL-IJCNLP.

8. Sentence Processing

- Hahn et al. (2022). A resource-rational model of human processing of recursive linguistic structure. PNAS.
- Arehalli and Linzen (2024). Neural networks as cognitive models of the processing of syntactic constraints. Open Mind.
- Van Schijndel and Linzen (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. Cognitive Science.
  - The authors use language models as a baseline: if surprisal from good LMs doesn't explain a certain aspect of human sentence processing, it probably means that we need something other than surprisal to explain it.
- Li and Ettinger (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. Cognition.
- Ryu and Lewis (2021). Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. CMCL.

9. Semantics

- Potts (2019). A case for deep learning in semantics: Response to Pater. Language 95(1).
  - Sketch of what a LLM-focused semantics could look like.

Several theoretical linguists have argued for some sort of descriptive content, or concepts, in addition to intensional semantics, notably Nick Asher and Louise McNally. But then, how can we simulate the kind of rich descriptive representations needed for that? Word embeddings!

- Louise et al. (2016). Conceptual vs . Referential Affordance in Concept Composition.
- Sadrzadeh and Muskens (2018). Static and Dynamic Vector Semantics for Lambda Calculus Models of Natural Language. Journal of Language Modelling.
- Emerson (2020). Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics.

10. Pragmatics

- Ko et al. (2022). Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections. Proceedings of ACL 2022.
  - Discourse comprehension via questions and answers — precursor to their later work on obtaining Questions Under Discussion with LLMs.
  - Automatically generating QUDs using LLMs.
- White et al. (2020). Learning to refer informatively by amortizing pragmatic reasoning
  - Rational Speech Acts theory has a problem with cognitive plausibility as the approach as originally formulated won't scale up. This paper proposes that instead of doing pragmatic reasoning from scratch every time, cognizers memorize. There is a proof-of-concept implementation using LLMs.
- Hu et al. (2023). Expectations over unspoken alternatives predict pragmatic inferences. Transactions of the Association for Computational Linguistics.

11. Tools for helping documentary linguists

Resources on bridging computational linguistics and NLP with language documentation (some resources are position papers)

- Lots of papers in SIGMORPHON like this:
  Ginn et al. (2023) .Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing.
- ComputEL: `https://aclanthology.org/2024.computel-1.0/`
- AmericasNLP: `https://aclanthology.org/volumes/2024.americasnlp-1/`
- Lane and Bird (2020). Bootstrapping Techniques for Polysynthetic Morphological Analysis. Proceedings of ACL.
- Liu et al. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. Proceedings of ACL.
- Bird and Yibarbuk (2024). Centering the Speech Community. Proceedings of ACL.

12. Model evaluation with Linguistics

- Bender (2011). On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology
- Jeretic et al. (2020). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. Proceedings of ACL.
- Kodner et al. (2023b). Morphological Inflection: A Reality Check. Proceedings of ACL.

13. Neuroscience

Toneva et al. (2022). Combining computational controls with natural text reveals aspects of meaning composition. Nature computational science.

- The authors propose a technique for removing lexical effects from LLM token representations, and use the resulting representations to probe where in the brain supra-lexical semantics might be processed.