Cognitive science of large language models (Psych 711)

Gary Lupyan

Fall 2025

E-mail: lupyan@wisc.edu
Web: sapir.psych.wisc.edu
Office Hours: Mondays 1pm-2pm.
Class Hours: Wednesdays 9:00am-11:30am
Office: Psych 526
Class Room: Psych 338

Course Description

The development of large language models (LLMs) — large artificial neural networks trained on large amounts of natural language — is arguably the biggest thing to happen to Cognitive Science in decades. Beyond whatever uses and misuses these models come to have in our society, their very existence acts as a stress test of many theories and frameworks that have been developed to explain the human mind.

For example, language was thought to be unlearnable from data. Strong embodiment theories have argued that that much of our semantic knowledge is represented in terms of bodily states. What does it mean then that LLMs do learn language from data and come to possess rich semantic knowledge without any body or direct sensory experiences?

We will also be touching on the long tradition of mechanistic interpretability. How useful are methods developed to study the human mind for studying LLMs? Can studying LLMs inspire new methods for studying human cognition?

Course Learning Outcomes

Students will…

1. be familiarized with the intellectual history that led to large language models, to their basic operating principles, and relate them to classic connectionist ideas of distributed representations and error-driven learning.
2. consider competing claims of what it means to "understand" and how behavioral evidence can be used to make sense of these competing claims.
3. explore the consequences LLMs have for core issues in cognitive science including poverty of the stimulus, concept learning, the role of context, and "general" intelligence.
4. complete a final project that stress-tests an existing cognitive theory/framework using existing data, or an empirical project where you use an LLM as a model organism to help understand the ways a given cognitive problem may be solved.

Expectations

Students are expected to read the 2-5 readings assigned for each class. I expect students to critically engage with each reading. Be on the lookout for ideas that challenge your intuitions and that provoke you into thinking differently. To help you engage with the readings, each student will be randomly assigned to a 2-3 person group each week. After reading the papers individually, the group should get together to discuss the readings and – working together – fill out a *Response Sheet* (see Template. Each group should upload the completed template to Canvas by 8pm on Monday. I will grade each response sheet on a 1-5 scale and will use your responses to help organize the discussion for the following class. Please note that some weeks have more/longer readings than other weeks. Plan your schedule accordingly. **Final projects** will be completed in groups of two. Each group has the option of choosing to write either a review/synthesis paper, or an empirical paper. Review/synthesis papers should involve selecting a cognitive theory/framework and stress testing it using previously published evidence involving LLMs. Empirical papers can involve collecting your own data (in the lab, online, by prompting an LLM, as appropriate), and/or using an LLM as a model organism to help understand how people may be solving some problem or, more abstractly, the computational principle by which a certain problem can be solved.

Use of AI for classwork

Despite large language models being in the name of this course, please engage with the readings using your own brain. I recommend you minimize interruptions, cut yourself off from the Internet and grapple with the readings. Write down questions as you read. It's fine to look things up or chat with an LLM to help you understand things, but it is helpful to keep in mind that the work you are doing for the course is not for impressing me or getting a good grade. It is in the service of helping *you* do good science, to create useful knowledge.

How credit hours are met by the course

(Time estimates may vary for students)

- 35 hrs class-time
- 50 hrs reading
- 13 hrs weekly assignments
- 20 hrs final project
  = 118 hrs total

Grading Policy

- **30%** Attendance & in-class participation
- **30%** Reading responses & essays
- **10%** Final project presentation
- **30%** Final project paper

Class Schedule[1]

I recommend reading the papers in the order in which they are listed.

[1] Subject to revision

Week 01, Wednesday, 09/03

What is learnable?

- **Readings (in class):** Sutton (2019); Aguera y Arcas (n.d.);
- *Related resources*: Halevy et al. (2009); talk by Alyosha Efros

Week 02, Wednesday, 09/10

How do LLMs work?

- **Readings:** Rumelhart (1989); Lee (2025); Vaswani et al. (2017);
- **Other reasources for understanding transformers (very useful; strongly recommended!)**: The illustrated transformer; Transformers, the tech behind LLMs video and the next chapter focusing on attention
- Extra: McCoy et al. (2024)

Week 03, Wednesday, 09/17

Learning language from language

Special guest - Steven Piantadosi

- **Readings:** Elman (1990); Boleda (2020); Piantadosi (2024)

Week 04, Wednesday, 09/24

Learning about the world from language

Special guest - Marina Bedny

- **Readings:** Plato: The allegory of the cave; Lupyan & Lewis (2017); Yildirim & Paul (2024a); Wang et al. (forthcoming)
- Recommended: response to Yildirim and& Paul: Goddu et al. (2024); Counter-response: Yildirim & Paul (2024b)

Week 05, Wednesday, 10/01

Do large language models understand us?

- **Readings:** Mitchell & Krakauer (2023); Aguera y Arcas (2022); Piantadosi & Hill (2022)
- **Short reflection essay**: So what do you think? Do large language models understand us? (~800 words).
- Extra: Feel free to also browse the talks in this series organized by Stevan Harnad, e.g., the different kinds of understanding by Chalmers.

Week 06, Wednesday, 10/08

Stress testing embodiment

- Lots of reading this week!
- **Readings:** Barsalou (1999); Mollo & Millière (2023); Chalmers (2024); Pavlick (2023)
- Extra: A phliosophical take by David Chalmers on the possibility of "pure thinkers"

Week 07, Wednesday, 10/15

Stress testing the language of thought hypothesis: learning to learn and represent

- **Readings:** Quilty-Dunn et al. (2022) (also read at least 2 positive and 2 negative commentaries); Lake & Baroni (2023); Watch this talk on learning compositionality by Pavlick
- Extra: Griffiths et al. (2025); Binz et al. (2024);

Week 08, Wednesday, 10/22

Stress testing numerical cognition

- **Readings:** Leslie et al. (2008); Banerjee et al. (2025); O'Shaughnessy et al. (2023); Nanda et al. (2023) (don't worry about all the technical details. Focus on the big picture. The authors do have a useful video walkthrough).
- Extra: Gordon (2004); O'Shaughnessy et al. (2021);

Week 09, Wednesday, 10/29

Stress testing concepts: representational format; flexibility; data efficiency

- **Readings:** *Tentative:* Casasanto & Lupyan (2014); Barsalou (2016);
- *Related resources*: This ~2010 talk by Efros on the role of categories in recognition,

Week 10, Wednesday, 11/05

Stress testing reasoning: the role of context

- **Readings:** *Tentative:* Lampinen et al. (2024)

Week 11, Wednesday, 11/12

Stress testing theories of intelligence

- **Readings:** TBA: Legg & Hutter (2007a); Chollet (2019); chapter from Aguera y Arcas
- Extras: Legg & Hutter (2007b); van der Maas et al. (2006)

Week 12, Wednesday, 11/19

Stress testing methodology of cognitive psychology/cognitive (neuro)science

- **Readings:** Bower & Clapper (1989) (read or skim depending on your level of familiarity); Jonas & Kording (2017); Read 2 of the 12 case studies from Lindsey et al. (2025)
- Extras: Nickels et al. (2022); Yarkoni (2020)

Week 13, Wednesday, 11/26

Thanksgiving - no class

- **Project proposals due** by 8pm.

Week 14, Wednesday, 12/03

Revisiting old and new questions

- **Readings:** Review previous readings in light of what you've learned
- **Short reflection essay**: Is all the AI madness of the last few years good for or bad for understanding the human mind? (~800 words).

Week 15, Wednesday, 12/10

Final presentations

Group project presentations [20 min presentation + 8 min Q&A].

STUDENTS' RULES, RIGHTS & RESPONSIBILITIES

During the global COVID-19 pandemic, we must prioritize our collective health and safety to keep ourselves, our campus, and our community safe. As a university community, we must work together to prevent the spread of the virus and to promote the collective health and welfare of our campus and surrounding community.

UW-MADISON BADGER PLEDGE

COURSE EVALUATIONS

Students will be provided with an opportunity to evaluate this course and your learning experience. Student participation is an integral component of this course, and your feedback is important to me. I strongly encourage you to participate in the course evaluation.

Digital Course Evaluation (AEFIS)

UW-Madison now uses an online course evaluation survey tool, AEFIS. In most instances, you will receive an official email two weeks prior to the end of the semester when your course evaluation is available. You will receive a link to log into the course evaluation with your NetID where you can complete the evaluation and submit it, anonymously. Your participation is an integral component of this course, and your feedback is important to me. I strongly encourage you to participate in the course evaluation.

ACADEMIC CALENDAR & RELIGIOUS OBSERVANCES

- See: https://secfac.wisc.edu/academic-calendar/#religious-observances

Ethics of Being a Student in the Department of Psychology

The members of the faculty of the Department of Psychology at UW-Madison uphold the highest ethical standards of teaching and research. They expect their students to uphold the same standards of ethical conduct. By registering for this course, you are implicitly agreeing to conduct yourself with the utmost integrity throughout the semester.

In the Department of Psychology, acts of academic misconduct are taken very seriously. Such acts diminish the educational experience for all involved – students who commit the acts, classmates who would never consider engaging in such behaviors, and instructors. Academic misconduct includes, but is not limited to, cheating on assignments and exams, stealing exams, sabotaging the work of classmates, submitting fraudulent data, plagiarizing the work of classmates or published and/or online sources, acquiring previously written papers and submitting them (altered or unaltered) for course assignments, collaborating with classmates when such collaboration is not authorized, and assisting fellow students in acts of misconduct. Students who have knowledge that classmates have engaged in academic misconduct should report this to the instructor.

Academic Integrity

The members of the faculty of the Department of Psychology at UW-Madison uphold the highest ethical standards of teaching and research. They expect their students to uphold the same standards of ethical conduct. By registering for this course, you are implicitly agreeing to conduct yourself with the utmost integrity throughout the semester.

In the Department of Psychology, acts of academic misconduct are taken very seriously. Such acts diminish the educational experience for all involved – students who commit the acts, classmates who would never consider engaging in such behaviors, and instructors. Academic misconduct includes, but is not limited to, cheating on assignments and exams, stealing exams, sabotaging the work of classmates, submitting fraudulent data, plagiarizing the work of classmates or published and/or online sources, acquiring previously written papers and submitting them (altered or unaltered) for course assignments, collaborating with classmates when such collaboration is not authorized, and assisting fellow students in acts of misconduct. Students who have knowledge that classmates have engaged in academic misconduct should report this to the instructor.

Accommodations Policy

McBurney Disability Resource Center syllabus statement: "The University of Wisconsin-Madison supports the right of all enrolled students to a full and equal educational opportunity. The Americans with Disabilities Act (ADA), Wisconsin State Statute (36.12), and UW-Madison policy (Faculty Document 1071) require that students with disabilities be reasonably accommodated in instruction and campus life. Reasonable accommodations for students with disabilities is a shared faculty and student responsibility. Students are expected to inform faculty [me] of their need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized. Faculty [I], will work either directly with the student [you] or in coordination with the McBurney Center to identify and provide reasonable instructional accommodations. Disability information, including instructional accommodations as part of a student's educational record, is confidential and protected under FERPA." http://mcburney.wisc.edu/facstaffother/faculty/syllabus.php

UW-Madison students who have experienced sexual misconduct (which can include sexual harassment, sexual assault, dating violence and/or stalking) also have the right to request academic accommodations. This right is afforded them under Federal legislation (Title IX). Information about services and resources (including information about how to request accommodations) is available through Survivor Services, a part of University Health Services: https://www.uhs.wisc.edu/survivor-services/ .

Diversity and Inclusion

Institutional statement on diversity: "Diversity is a source of strength, creativity, and innovation for UW-Madison. We value the contributions of each person and respect the profound ways their identity, culture, background, experience, status, abilities, and opinion enrich the university community. We commit ourselves to the pursuit of excellence in teaching, research, outreach, and diversity as inextricably linked goals.

The University of Wisconsin-Madison fulfills its public mission by creating a welcoming and inclusive community for people from every background – people who as students, faculty, and staff serve Wisconsin and the world." https://diversity.wisc.edu/

Complaints

Occasionally, a student may have a complaint about a TA or course instructor. If that happens, you should feel free to discuss the matter directly with the TA or instructor. If the complaint is about the TA and you do not feel comfortable discussing it with the individual, you should discuss it with the course instructor. Complaints about mistakes in grading should be resolved

with the TA and/or instructor in the great majority of cases. If the complaint is about the instructor (other than ordinary grading questions) and you do not feel comfortable discussing it with the individual, make an appointment to speak to the Associate Chair for Graduate Studies, Professor Yuri Saalmann (saalmann@wisc.edu).

If you have concerns about climate or bias in this class, or if you wish to report an incident of bias or hate that has occurred in class, you may contact the Chair of the Department, Professor Shawn Green (cshawn.green@wisc.edu) or the Chair of the Psychology Department Climate & Diversity Committee, Martha Alibali (martha.alibali@wisc.edu). You may also use the University's bias incident reporting system, which you can reach at the following link: https://osas.wisc.edu/report-an-issue/bias-or-hate-reporting/.

Concerns about Sexual Misconduct

All students deserve to be safe and respected at UW-Madison. Unfortunately, we know that sexual and relationship violence do happen here. Free, confidential resources are available on and off campus for students impacted by sexual assault, sexual harassment, dating violence, and stalking (regardless of when the violence occurred). You don't have to label your experience to seek help. Friends of survivors can reach out for support too. A list of resources can be found at https://www.uhs.wisc.edu/survivor-resources/

If you wish to speak to someone in the Department of Psychology about your concerns, you may contact the Chair of the Department, Professor Shawn Green (cshawn.green@wisc.edu) or the Associate Chair of Graduate Studies, Professor Yuri Saalmann (saalmann@wisc.edu). Please note that all of these individuals are Responsible Employees (https://compliance.wisc.edu/titleix/mandatory-reporting/#responsible-employees).

References

Aguera y Arcas, B. (n.d.). *What is Intelligence?* Retrieved August 19, 2025, from https://whatisintelligence.antikythera.org/introduction/

Aguera y Arcas, B. (2022, February 16). *Do large language models understand us?* Medium. https://medium.com/@blaisea/do-large-language-models-understand-us-6f881d6d8e75

Banerjee, A. V., Bhattacharjee, S., Chattopadhyay, R., Duflo, E., Ganimian, A. J., Rajah, K., & Spelke, E. S. (2025). Children's arithmetic skills do not transfer between applied and academic mathematics. *Nature*, *639*(8055), 673–681. https://doi.org/10.1038/s41586-024-08502-w

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences*, *22*(4), 577-609; discussion 610-660. http://www.ncbi.nlm.nih.gov/pubmed/11301525

Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review*, 1–21. https://doi.org/10.3758/s13423-016-1028-3

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147. https://doi.org/10.1017/S0140525X23003266

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, *Accepted*. https://doi.org/10.1146/annurev-linguistics-011619-030303

Bower, G. H., & Clapper, J. P. (1989). Experimental methods in cognitive science. In *Foundations of cognitive science* (pp. 245–300). The MIT Press.

Casasanto, D., & Lupyan, G. (2014). All Concepts are Ad Hoc Concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: New Directions* (pp. 543–566). MIT Press.

Chalmers, D. J. (2024, August 18). *Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models*. https://doi.org/10.48550/arXiv.2408.09605

Chollet, F. (2019, November 25). *On the Measure of Intelligence*. https://doi.org/10.48550/arXiv.1911.01547

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. ISI:A1990DK92500001

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: Reply to Yildirim and Paul. *Trends in Cognitive Sciences*, *28*(11), 963–964. https://doi.org/10.1016/j.tics.2024.06.008

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496–499.

Griffiths, T. L., Lake, B. M., McCoy, R. T., Pavlick, E., & Webb, T. W. (2025, August 7). *Whither symbols in the era of advanced neural networks?* https://doi.org/10.48550/arXiv.2508.05776

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121. https://doi.org/10.1038/s41586-023-06668-3

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233

Lee, T. B. (2025, May 19). *Large language models, explained with a minimum of math and jargon*. https://www.understandingai.org/p/large-language-models-explained-with

Legg, S., & Hutter, M. (2007a, June 25). *A Collection of Definitions of Intelligence*. https://doi.org/10.48550/arXiv.0706.3639

Legg, S., & Hutter, M. (2007b, December 20). *Universal Intelligence: A Definition of Machine Intelli-*

gence. https://doi.org/10.48550/arXiv.0712.3329

Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, *12*(6), 213–218.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025). *On the Biology of a Large Language Model*. Transformer Circuits. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Lupyan, G., & Lewis, M. (2017). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Mollo, D. C., & Millière, R. (2023, April 3). *The Vector Grounding Problem*. https://doi.org/10.48550/arXiv.2304.01481

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023, October 19). *Progress measures for grokking via mechanistic interpretability*. https://doi.org/10.48550/arXiv.2301.05217

Nickels, L., Fischer-Baum, S., & Best, W. (2022). Single case studies are a powerful tool for developing, testing and extending theories. *Nature Reviews Psychology*, *1*(12), 733–747. https://doi.org/10.1038/s44159-022-00127-y

O'Shaughnessy, D. M., Cruz Cordero, T., Mollica, F., Boni, I., Jara-Ettinger, J., Gibson, E., & Piantadosi, S. T. (2023). Diverse mathematical knowledge among indigenous Amazonians. *Proceedings of the National Academy of Sciences*, *120*(35), e2215999120. https://doi.org/10.1073/pnas.2215999120

O'Shaughnessy, D. M., Gibson, E., & Piantadosi, S. T. (2021). The cultural origins of number. *Psychological Review*. http://colala.berkeley.edu/papers/oshaughnessy2021cultural.pdf

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. *From Fieldwork to Linguistic Theory*, 353–414. https://zenodo.org/records/12665933

Piantadosi, S. T., & Hill, F. (2022, August 12). *Meaning without reference in large language models*. https://doi.org/10.48550/arXiv.2208.02957

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. *The Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002849

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In *Foundations of cognitive science* (pp. 133–159). The MIT Press.

Sutton, R. (2019). *The Bitter Lesson*. https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/

0033-295X.113.4.842

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/arXiv.1706.03762

Wang, Z., Akshi, Keil, S., Kim, J. S., & Bedny, M. (forthcoming). Constructing meaning from language: Visual knowledge in people born blind and in large language models. *Annual Review of Linguistics*. Forthcoming.

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Yildirim, I., & Paul, L. A. (2024a). From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences*, *28*(5), 404–415. https://doi.org/10.1016/j.tics.2024.02.008

Yildirim, I., & Paul, L. A. (2024b). Response to Goddu et al.: New ways of characterizing and acquiring knowledge. *Trends in Cognitive Sciences*, *28*(11), 965–966. https://doi.org/10.1016/j.tics.2024.08.004