FINAL PROJECT

You (and, if applicable, 1 to 3 other students on your team) will investigate the cognitive abilities or one or more Large Language Models (LLMs), such as ChatGPT, Claude, or Gemini.

Project Description

As we discussed in class, LLMs are massive neural networks that can do amazing things which they were not explicitly trained to do. Yet, we do not understand their full range of capabilities (and limitations), nor do we have a deep understanding of why they are able (and unable) to do these things.

Your team is tasked with examining the abilities of one of more LLMs in the cognitive task of your choice. You can do so using the user interface or using an API if you are comfortable doing so. However, no sophisticated computer skills are expected for this assignment.

How do I select a cognitive ability to study? There are three main ways you can do this:

- You could select something discussed in class or in one of our readings (either optional or required).
- You could think of phenomena you have learned about in other Psychology courses (if you are a Psychology student) or drawn from the table of contents of a Psychology textbook.
- You could browse through issues of psychology or cognitive science journals to find an article reporting a novel phenomenon in *human* cognition, and then test whether LLMs do the same thing. Some examples of good journals to check out would include *Cognition*, *Cognitive Science*, *Psychological Science*, *Journal of Experimental Psychology: General*, *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *Judgment & Decision-Making*, and *Journal of Experimental Social Psychology*. This option is perhaps a bit tougher to start with, but a benefit is that the journal article will have lot of references you can use to help you write your paper.

Note that "cognitive abilities" take many forms and can include both successes and failures (e.g., "cognitive biases") on the part of humans. On Learn, there are several examples of papers from the academic literature that attempt to benchmark LLMs against human performance; some are asking questions like "Are LLMs as *skilled* at this task as humans?" while others are asking "Are LLMs as *biased* at this task as humans?" (or "Are LLMs biased in the *same way* as humans?"). For instance, the papers on analogy generally assume that humans are good at analogy and ask whether LLMs are equally good; the paper on metacognition assumes that people are overconfident and asks whether LLMs are as 'bad' as humans on this dimension.

How should I test the LLM(s)? I suggest you begin by just playing around and seeing what sorts of answers the model provides. Although you can start with questions drawn directly from academic articles, you would ideally want to come up with your own examples as well to test whether the model can generalize (e.g., in case the original article is contaminating the training data). Try to think like an experimentalist and to give the model different variations on similar tasks in order to probe the areas where it succeeds and fails. In the end, you will want to use an approach that is as systematic as possible.

This paper (mentioned in the lecture and posted to Learn) is very short but has a lot of great suggestions for how to generate good tests of cognitive abilities in LLMs:

> Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*, 451-452.

You must include an appendix with all prompts and outputs from your model(s), although you do not need to describe all of this in your report (and odds are, it would be confusing to do so).

One technical issue to bear in mind is that the context window (i.e., the number of tokens of input the model uses to generate the output) may include some of your previous prompts if you are testing the LLM on multiple prompts in the same conversation. You should be aware of this and ensure that you are clearing the context window so that you are getting a clean test of the model's ability to perform the specific task in a given prompt. How to clear the context window varies across models and you should ensure you know how to do this for whichever model(s) you are using. For example:

> **ChatGPT:** Go to settings > Personalization > make sure "Reference saved memories" and "Reference chat history" settings are turned off. Now you can create a new chat to clear the context window. Without these steps, ChatGPT will have access to saved memories and recent chats.
>
> **Claude:** Start a new chat.
>
> **Gemini:** Start a new chat.
>
> **Grok:** Go to X settings > Privacy and Safety > Grok Settings > turn off "Personalize Grok with your conversation history" and delete conversation history. Now you can start a new interaction to clear the context window.

In your appendix, please clarify when you are starting a new conversation or clearing the chat history.

What should I include in my report? The main things to include are (i) a brief introduction explaining what the cognitive ability is that you are testing (perhaps contextualizing your paper in light of what is known about that capacity in humans); (ii) an explanation of your overall strategy for testing this ability in the LLM; (iii) some examples of prompts you used, the output received, and how you interpret the output; (iv) an overall assessment of the LLM's capability; and (v) any limitations you can think of that might qualify the certainty of your conclusions. In addition, please look through the academic literature (e.g., using Google Scholar) to see whether others have attempted to benchmark a similar ability. It's totally fine if they have, but if so, please contextualize your work relative to those other attempts (what is different in your approach and your results).

How can I make this project "more ambitious"? Below, you will see that our expectations for "ambition" are higher for larger groups. Several factors can make your project more ambitious (you certainly do not need to do all of these):

- Selecting a task or ability using the wider research literature rather than testing something directly discussed in class (or in the other LLM papers on Learn).
- Developing more distinct tests of your ability in order to demonstrate converging evidence and examine whether the model can generalize.
- A more original or innovative method for testing the LLM, or one which relies on a deeper understanding of existing literature.
- Including enough prompts/tasks/items that you could calculate descriptive or inferential statistics (and perhaps supplying a graph) rather than merely reporting qualitative results.
- Testing multiple models to compare to one another (perhaps quantitatively)
- Really anything creative that your group can think of to think outside the box. We are quite open-minded and excited to see the possibility of clever, innovative projects.
- Digging into *how* the model makes mistakes rather than just *whether* it makes mistakes. That is, including follow-up tests that examine the mechanisms underlying any errors.

We will give you feedback about this when you submit your prospectus.

Are there examples of others' benchmarking attempts we can see? On Learn, we have posted several examples of academic articles that benchmark various cognitive abilities. You can use these as inspiration, but we don't expect your project to rise to the level of the published academic literature. The papers posted to Learn include a few papers examining different abilities (analogy, theory-of-mind, explanation, and metacognition) as well as a series of several papers examining the same ability (in this case, analogy) as an example of how different methods can lead to different conclusions.

How will these be marked? The report will be marked on 4 criteria:

- ***Explanation of concepts (20%).*** Since you are examining a cognitive ability, you should include an explanation of what is known about that cognitive ability in humans (and, if applicable, prior work examining that ability in LLMs). Be sure to explain what the ability is and what processes are thought to support this ability in humans. In some cases, this might mean explaining one or more published journal articles because the ability has been studied in detail already. In other cases, you may have thought of an ability that hasn't been studied in the academic literature; in that case, it would be sufficient to thoughtfully analyze what mental processes that ability requires without extensive references.
- ***Testing approach and execution (40%).*** You should include an explanation of your approach to testing the LLM (possibly including an explanation of the methods used in prior studies testing humans or LLMs, if you adapt such methods). Report enough examples (and, if applicable, descriptive statistics) to illustrate the central results you have discovered.

- ***Interpretation (25%).*** Interpret what these results mean in the context of your broader explanation of the cognitive capacity. What conclusions should we draw? What are the limitations of your method (and how might future research address those limitations)?
- ***Exposition (15%).*** Write clearly and concisely, linking each paragraph of your report together into a coherent narrative. It should go without saying that your report should be free of spelling or grammatical errors.

Although you may use an LLM to help you write your report, (i) you must include all prompts and outputs in your appendix (including those used to generate text for the report itself) and (ii) you are responsible for any errors—including factual or logical errors—made by the model.

Fine Print (from syllabus)

Groups. The project may optionally be done in groups of up to four students; all group members will receive the same mark except in the event of highly unequal contributions where this is convincingly documented. The standards for ambition are somewhat higher for a group project than for an individual project (and higher in larger groups than in smaller groups).

Length. There is no firm length requirement, but we expect that about 4 double-spaced pages per group member would be required to demonstrate an appropriately detailed and ambitious project (e.g., about 4 pages for an individual, ranging up to about 16 pages for a group of four).

Use of GenAI. This project is unusual in that you may use LLMs in whatever way your group would like, including the generation of text if your group deems this desirable. However, you will be required to cite sources and to provide a transcript of all prompts and their outputs as an appendix to your report (this is separate from the page count and may be quite lengthy).

Deliverables. To help you pace your project, you must complete the milestones listed by the deadlines in the syllabus.

By the first deadline (Sep 19, 2025), one member of your group must email the TA with either (i) a list of your group members, (ii) a statement that you would like to be assigned to a group, or (iii) a statement that you will work on the project individually. If you do not email your TA by this time, we will assume you want to be assigned to a group and will do so automatically.

By the second deadline (Oct 24, 2025), one member of your group must submit a brief prospectus (1 page) to Learn which (i) summarizes your idea and (ii) identifies a plan for carrying it out (e.g., what tasks are necessary and who will do them). Your TA will provide brief feedback on this document to ensure you are on the right track. The prospectus is worth 2.5% of your course grade.

By the third deadline (Dec 15, 2025), one member of your group must submit the report itself to Learn, as a single PDF document with your write-up of your findings and your appendix containing all prompts and outputs. The report is worth 22.5% of your course grade.