Thank you for your interest in **CogBench**! This project's goal is to critically examine the cognitive and linguistic capabilities of LLMs through a large collaboration with experts in human cognition and language. CogBench will contain a collection of "experiments" that probe LLM functionalities like cognition, reasoning, and language processing.

 # Project goal

Experiments in fields like cognitive psychology and psycholinguistics have led to strong models of human cognition and language processing. Similarly, we hope that CogBench can contribute to a better understanding of how LLMs "think" as humans do.

We hope to develop a **broad spectrum cognitive benchmark** that evaluates LLMs across areas of cognition that are well-studied in humans – e.g. selective attention, working memory, reading comprehension. Our aim is to understand the ways in which LLM cognition differs from humans, as well as the ways in which it is similar, and we hope the outcome of this benchmark can be interesting to both human-focused and computation-focused researchers.

 # Types of data we'd like to collect

CogBench is a collection of experiments that are designed to probe some aspect of cognition. Ideally, these experiments will be similar to the kinds of experiments that are conducted with human participants (in the following sections, we will show how you can adapt a human experimental paradigm to the LLM setting). Each experiment in CogBench consists of:

- a set of **prompts** to give the LLM, where each prompt contains:
    - instructions about the experimental task and how to respond
    - either a single stimulus or a set of stimuli
- predictions about human responses to those prompts
    - Similar to the predictions you would make in a human study – how would responses for different types of stimuli differ?
    - For example, if implementing a Stroop task with taboo words, a prediction would be that response times for taboo words are longer.
    - These predictions will be used to evaluate LLMs: we will test several state of the art LLMS to see whether their responses support the predictions.
- Information about what field (e.g. cognitive science, linguistics) the experiment relates to, as well as specific sub-areas (e.g. memory, disambiguation) with corresponding publication information if published

The benchmark is a way to measure LLM capabilities. So, we can ask dozens of LLMs to respond to the prompts. Using the predictions, we can evaluate whether the LLM responses match what one would expect from humans.

If you are interested in contributing to CogBench and collaborating with us as a co-author of this paper, please carefully read the following contribution steps.

How to Contribute

I. Step 1: Brainstorm

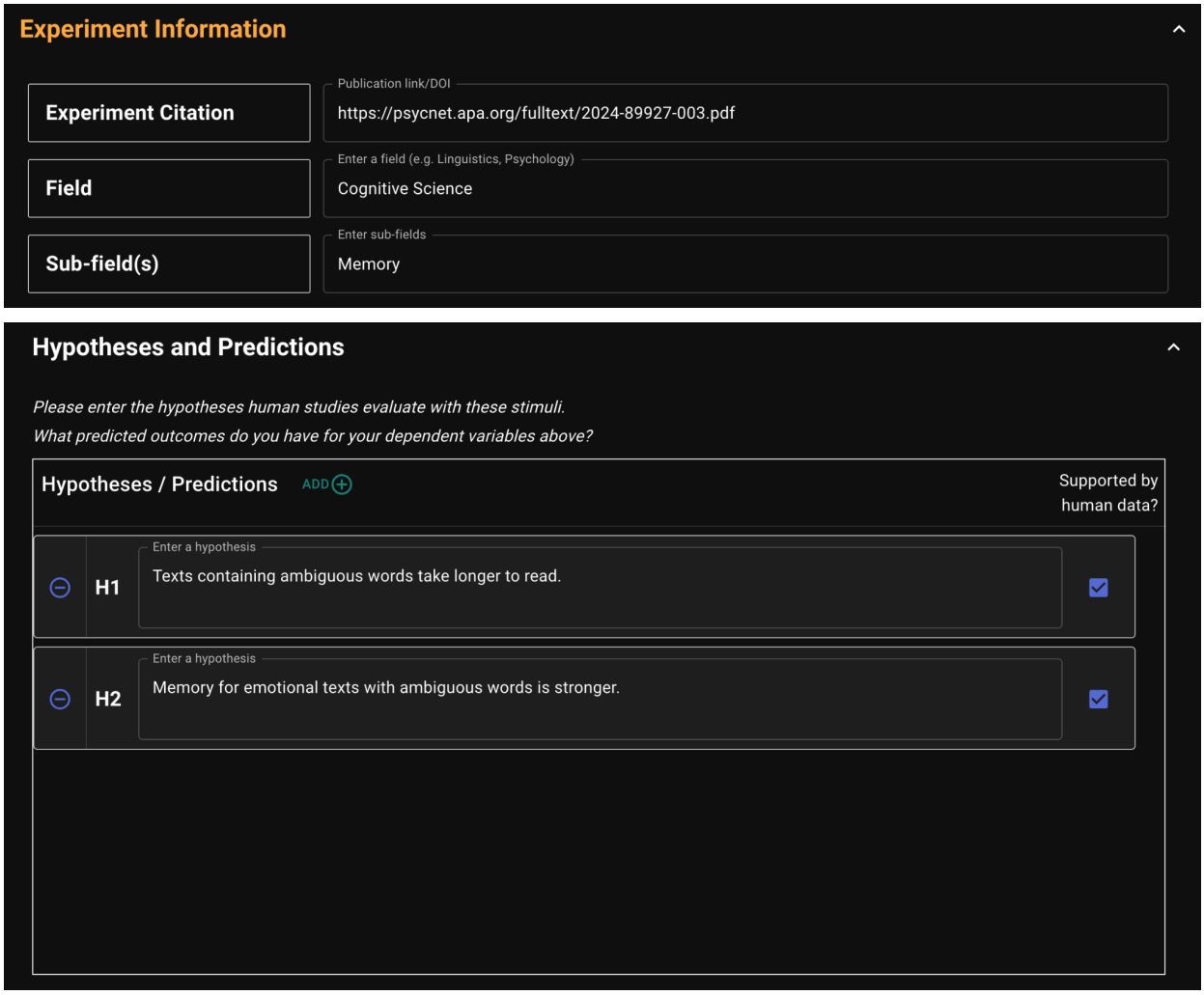
We recommend starting by thinking about what papers/experiments, in your opinion, demonstrate important or interesting properties of human cognition, reasoning, language, language processing, etc. It does not matter whether you are an author on the relevant papers.

These favorite experiments from Step I are all great candidates to "add" to CogBench – if an experiment demonstrates an important property or phenomenon in humans, it will be interesting to do a similar experiment with LLMs.

II. Step 2: Add basic experiment information

For each experiment you would like to "add" to CogBench, you'll fill out a copy of this form. We encourage you to add as many experiments as you like, and it's ok to submit multiple related experiments! But for each experiment, you'll need to submit a new copy of this form.

In the form, you'll (1) specify the experiment, and (2) briefly outline the field and the hypotheses / predictions of the experiment. The predictions are an important component, because we'll test these predictions on a suite of state of the art LLMs. See example inputs below:



III. Step 3: Add data/prompt

Next, you'll suggest a prompt that can be used to translate the human experiment into an LLM experiment.

If you don't have much experience prompting LLMs, you can refer to these best practices: https://huggingface.co/docs/transformers/main/en/tasks/prompting#best-practices-of-llm-prompting

Tips for experiment prompting

In general, when writing your prompt, the same instructions that were given to the human participants is a good starting point. Just like a human, the LLM will need to know what to expect from the stimulus, and what kind of response it should give. However, some experimental components don't translate so easily from human to LLM contexts. Below are some examples:

Human Experimental Paradigm	Similar LLM Strategies

Recall	If you typically manipulate the experiment to test recall, consider the following replacements: – Insert a distracting block of text. You can experiment with a paragraph to several paragraphs. In general, for LLMs, don't hesitate to use more distractors than would be necessary for humans. – Insert a distractor task, then ask the LLM to respond to the target task without actually repeating the target task.
Teaching intervention	If your experiments include a teaching intervention, you can include that within the prompt. The same instructions (and examples) used for humans should work well.
Short stimulus presentation time*	If your experiment briefly presents a stimulus (e.g. in order to minimize conscious processing of the stimulus), consider the following replacements: – Insert a random story, news article, etc before *and* after the stimulus. This is because LLMs pay the least attention to the information in the middle of the prompt. – Insert random words/letters between your target words/letters
Moving stimulus (e.g. animation or movie)	Several still images, presented in order

Some dependent variables from human studies, like Likert ratings, are very straightforward to translate to LLMs. You can simply describe the scale in the prompt, like this:

> Please read the following text and rate how 'complete' or 'finished' the painting is on a scale of 1-7, with 7 being 100% finished. Respond only with the number on the scale of 1-7:
> Parts of the apartment needed work. Pat first painted a bedroom (as promised).

However, for many human study measurements (for example, EEG data), there is no immediate equivalent for LLMs. Below are some suggestions to consider, but you are not limited to following these suggestions. Please let us know if you'd like to add anything to this list!

Human Experiment Measure	**LLM prompting strategies**
Reaction times, psychometric data (heart rate, electroencephalogram, galvanic	1. Ask for confidence rating along with answer

skin response, etc)	2. Ask for the answer only – sometimes items that result in longer response times from humans result in incorrect answers from LLMs. 3. Look at word probability scores from model*
Button press	Try reframing as a multiple choice question. For example, imagine an experiment in which participants have the choice between pressing a button to take a card from Deck A, Deck B, Deck C, or Deck D. Instead, you can write something like: Choose a card from one of the following decks: A, B, C, or D. Respond only with the letter of the deck.
Likert scale/other rating	Describe scale, ask for specific number
Lexical decision task	1. Word edge completion task for target word. Sample instructions: A "word edge" is the first or first and second letter of a word, followed by a number of blanks that you are to fill in exactly to make an English word. You should complete the edge with the first allowable word that comes to mind that fits into the blanks. A word may be a compound (for example, SETUP or LUNCHBOX) or have endings (for example, STICKS or BUSHES). Don't use fewer or more letters than are provided for by the blanks. Also, don't use personal names. If you can't think of a word quickly enough, leave it blank. 2. Ask for a confidence rating for the target word

Your prompt should also contain specific instructions about how the LLM should respond. **LLM responses should be standardized** if it is at all possible for your experiment. For instance, you can require an LLM to respond only with a number, or only with a single word, for each stimulus. This allows us to more robustly evaluate LLM responses.

IV. Step 4: Test the prompt

What you're looking for:

- **Q. Are the responses in the format you specified?**
  For example, if the response contains a rating along with an explanation when you only asked for a rating, this is not good.
- **Q. Do the responses appear to vary depending on the input?**
  If all responses are the same, or if the LLM seems to be randomly oscillating between only two or three possibilities, it's possible that the instructions can be improved so that the LLM understands the task better.
- **Q. Are the responses in line with what you expect from humans?**
  It doesn't matter – we're not sure how much we should expect LLMs to align with humans.

V. Step 5: Add further thoughts

If after looking at the prompt outputs, you have any further thoughts about how to evaluate LLM capabilities for this experiment, please share them here! Our team will try to incorporate this into the evaluation protocol of CogBench. We may also contact you with follow-up questions