

1. *Logical Reasoning*. Please provide three example prompts that shows that an LLM is not capable to conduct logical (or counterfactual) reasoning. For each of the three prompts:

- (a) Please provide the prompt
- (b) Explain why the LLM is not able to provide an correct answer
- (c) Provide an output of at least two different LLMs

Example 1

Prompt: Which weighs more, a two pound brick or a pound of feathers?

Reasoning: The LLM may not be able to provide the correct answer because the more common question is "Which weighs more, a pound of bricks or a pound of feathers?". Many LLMs have overfitted to this question, and hence evaluate similar questions incorrectly.

- **MS Edge Copilot:** A two-pound brick and a pound of feathers actually weigh the same! Both are equal to one pound. The difference lies in the material: the brick is denser, while the feathers take up more space.
- **Gemini:** The two-pound brick weighs more! ...or at least that's what you might expect based on common sense. But, this is a trick question! In reality, a pound of feathers weighs the same as a pound of bricks.
- **ChatGPT 3.5:** A two pound brick weighs more than a pound of feathers.

Result: MS Edge Copilot was completely incorrect, Gemini was partially correct and ChatGPT 3.5 answered correctly.

Example 2

Prompt: Your mom is my grandmother. Who is your cousin?

Reasoning: The LLM may not be able to provide the correct answer because it cannot reason about the various possible family relations that the statement implies. Secondly, it has difficulties handling the concept of "you" and "I".

- **MS Edge Copilot:** I apologize, but I don't have the capability to determine familial relationships or identify cousins. If you have any other questions, feel free to ask!
- **Gemini:** If my mom is your grandmother, then you are my parent. Since cousins are the children of your parents' siblings, then my cousins would be your siblings.
- **ChatGPT 3.5:** If your mom is my grandmother, that means you are either my parent or my aunt/uncle. Therefore, your child would be my cousin.

Result: MS Edge Copilot refuses to answer. Gemini seems to understand the first statement, but misjudges its implications. ChatGPT 3.5 does not correctly evaluate the concept of "you" and "I".

Example 3

Prompt: If it rains, the ground is wet. The ground is wet. Did it rain?

Reasoning: The LLM may not be able to provide the correct answer because it is not capable of logical reasoning. $A \implies B$ does not mean that $B \implies A$. However, as it is not unreasonable to assume it rained, the LLM might get confused.

- **MS Edge Copilot:** Yes, it's reasonable to infer that it rained. The wet ground suggests precipitation occurred.
- **Gemini:** It is likely that it rained. The statement "If it rains, the ground is wet" shows a cause-and-effect relationship. (...) However, it's important to consider limitations. Other things besides rain can make the ground wet, like a sprinkler or a spilled drink. (...)
- **ChatGPT 3.5:** (...) In this specific case, since we know from the conditional statement that if it rains (R), then the ground is wet (W), and we observe that the ground is indeed wet (W), it logically follows that: Rain (R) must have occurred. (...) Yes, it did rain.

Result: MS Edge Copilot's answer is reasonable, but not logically sound. Gemini responds nicely, explaining that we cannot say for certain that it rained. ChatGPT 3.5 starts out nicely, but in the end it makes the mistake of confidently stating that it did rain.

2. *Evaluation.* You have build an own LLM, which is able to complete sentences and even documents and may serve as a base for a chatbot. Now you want to know, how well it is working.

- (a) What is your overall evaluation strategy?
- (b) What evaluation measures do you use?

(a) First of all, my evaluation strategy would depend on the specific use case for the LLM. What are we trying to optimize for? In this case I will assume we want the LLM to serve as a chatbot that is replacing a human, for example in customer service at a bank. What is noteworthy here is that we have access to the "correct answer", by asking the human worker. This enables use of a number of evaluation metrics. Hence, my overall evaluation strategy would include a set of evaluation measures matching the LLM's responses with responses from the human worker. Additionally, I would choose an off-the-shelf local LLM to use as a baseline. Finally, I would track how often real users report that the LLM was helpful, and do A/B testing between the LLM and a human worker to assess the "real" difference in performance.

(b) For evaluation measures, I would use BERT for computing similarity between the LLM and human worker, and between the baseline and the human worker. BERT is chosen because it allows to assess "semantic" similarity, not just word similarity. Additionally, as mentioned, I would use human evaluation in the A/B testing.

3. *Deep Learning.* Many attention-based approaches combine as inputs a token embedding, separator chars, and position embedding.

- (a) Why is there a position embedding, what is its purpose?
- (b) Are there tasks, where the position embedding is not useful? If yes, please provide an example.
- (c) Do you expect that some attention heads specialise on the positional embedding?

-

4. *Generated Text.* The operators of a discussion forum on the internet (e.g., DerStandard, Reddit, ...) ask you to develop a tool to automatically detect post written by bots. You get access to the text of the post, and some meta-data (IP address, time and date, browser fingerprint, ...). In addition you get the access to the archive of the forum.

- (a) What approach do you consider?
- (b) What features would you use? (bullet points)
- (c) How well do you expect your method to work? What are the bottlenecks?

(a) If the archive of the forum has plenty of posts labelled as written by bots or human, I would consider using a machine learning model such as an MLP or Random Forest, and train it to classify the difference. If no such labeled dataset exists, I would try an approach of outlier detection.

(b)

- All the meta-data of the post as features.
- Keyword extraction of the text to select a number of prevalent keywords (that are not just common english words).
- Topic, identifies by topic modeling, as I believe bot posts have reoccurring topics.
- Perplexity score, calculated from the text, because I believe bots and humans might score differently.

(c) Since I believe my feature selection should be able to distinguish bot posts from human posts, I believe my method (at least the machine learning model) to work quite well. However, this relies on that the model is trained on enough data. The bottleneck is the size and quality of the training data.

5. *Causality.* You are asked to build a system to extract causal statements from scientific papers in order to populate a knowledge graph. For example the sentence "Improvement in the annotation tool leads to more consistent and reliable annotations" should yield {Improvement in the annotation tool}_{Cause}, {leads to}_{Cue}, {more consistent annotations}_{Effect}. Then both cause and effect would represent nodes, with the causal cue being represented by an edge in the knowledge graph.

- (a) What type of approach do you choose?
- (b) What pre/post processing do you apply?
- (c) What properties do you expect for your approach?

- (a) The two main approaches are the rule-based approach and the machine learning approach. I would choose the rule-based approach because it is easier to understand, predictable and does not require training on large datasets.
- (b) For preprocessing, I would filter sentences that the rule-based approach can identify as including a causal relation. Then I would split the sentence into cause, relation and effect. For postprocessing, I would try to identify nodes that are similar, such that they can be merged.
- (c) I would expect this approach to be predictable and fast to iterate on. I expect it to initially make a lot of mistakes, due to inaccurate and too generalized rules. However, after a lot of rule adjustments, I believe it can perform decently. I don't expect it to be able to generalize to new datasets in the way a machine model would, but I believe it would be a lot simpler to work with.