Evaluating LLaMA on MCQA: a case study

Tomáš Musil

11 April 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

The Task – Multiple Choice Question Answering

What will most likely result if a high-pressure system remains in an area for a long period of time?

(A) fog

(B) rain

(C) drought

(D) tornado

ARC-Challenge DEV set

The Goal – Replicate LLaMA Results and Compare New Models

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

Baseline: random choice – 25 %

Evaluation Methods

How to evaluate a multiple choice question?

- First idea: just copy the question into prompt and let the LM generate the answer.

Evaluation Methods

How to evaluate a multiple choice question?

- First idea: just copy the question into prompt and let the LM generate the answer.
- Does not work. Why?

Evaluation Methods

How to evaluate a multiple choice question?

- First idea: just copy the question into prompt and let the LM generate the answer.
- Does not work. Why?
  - Answers that are not in the choices.
  - Bias.

Evaluation Methods

- Measuring probability of generating the given choices
    - $P(A_x \mid \text{prompt}(Q))$
        - Compare for all answers
        - Select the most probable
    - With/without normalization

We evaluate LLaMA on free-form generation tasks and multiple choice tasks. In the multiple choice tasks, the objective is to select the most appropriate completion among a set of given options, based on a provided context. We select the completion with the highest likelihood given the provided context. We follow Gao et al. (2021) and use the likelihood normalized by the number of characters in the completion, except for certain datasets (OpenBookQA, BoolQ), for which we follow Brown et al. (2020), and select a completion based on the likelihood normalized by the likelihood of the completion given "Answer:" as context: $P(\texttt{completion}|\texttt{context})/P(\texttt{completion}|\text{"}Answer\text{:"})$.

Prompt Formulation

Context →	Question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? Answer:
Correct Answer →	dry palms
Incorrect Answer →	wet palms
Incorrect Answer →	palms covered with oil
Incorrect Answer →	palms covered with lotion

Figure G.11: Formatted dataset example for ARC (Challenge). When predicting, we normalize by the unconditional probability of each answer as described in 2.

Context →	Answer these questions: Q: Who sang who wants to be a millionaire in high society? A: Frank Sinatra Q: Who wrote the book the origin of species? A:	Context →	Answer these questions: Q: In Scotland a bothy/bothie is a? A: House Q: The ancient city of Troy is located in what modern country? A:
Target →	Charles Darwin	Target →	Turkey

Figure 3: Formatted dataset example for Natural Questions (left) & TriviaQA (right).

LLaMA 7B OBQA – 57.2 %

- "Answer this question:\n" + question + "\nAnswer: "
  - 51.6 %
- "Answer this question: " + question + "\nAnswer: "
  - 52.8 %
- "Question: " + question + "\nAnswer: "
  - 52.2 %
- question
  - 57.4 %

LLaMA 13B OBQA - 56.4 %

- "Answer this question:\n" + question + "\nAnswer: "
  - 57.4 %
- "Answer this question: " + question + "\nAnswer: "
  - 56.6 %
- "Question: " + question + "\nAnswer: "
  - 53.8 %
- question
  - 55.4 %

Summary:

- Task evaluation strategies ≠ end user LLM usage.
- Specific prompt formulation (and tokenization) matters.
- Replicating LLM evaluation results is complicated for the open LLMs and impossible for the proprietary ones.