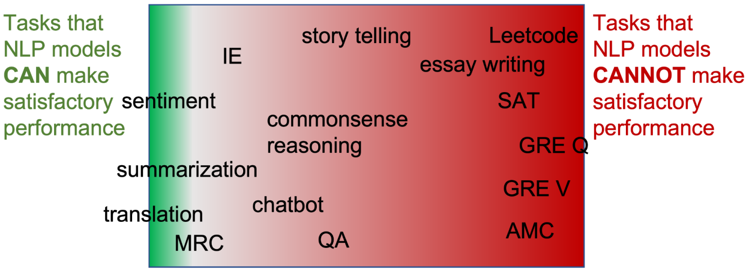
Three Principles to Benchmark Large Language Models

Meng Jiang

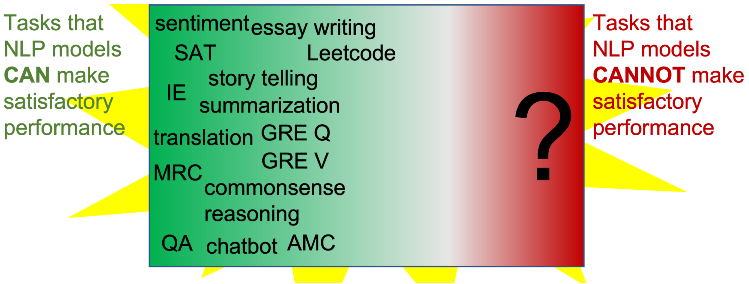
Large language models (LLMs) such as ChatGPT and GPT-4 are amazing. Their intelligent abilities and inabilities are mysteries. Revealing those by creating testbeds or benchmarks on tasks that need intelligence is attracting attention, because it would be an impactful yet not too time consuming effort. What tasks should we create the testbeds for?

We had a great number of NLP benchmarks prior to the emergence of LLMs. By that time, we assume this is what the NLP task space looks like:


Tasks that NLP models **CAN** make satisfactory performance
IE
story telling
Leetcode
essay writing
sentiment
SAT
commonsense reasoning
GRE Q
summarization
GRE V
chatbot
translation
MRC
QA
AMC
Tasks that NLP models **CANNOT** make satisfactory performance

Most of the research was designing, developing, and evaluating supervised learning models to improve the performance of the tasks on the boundaries. Two reasons: (1) working on the pure red zone is meaningless, and (2) the tasks on the boundaries are important for various industrial applications.

LLMs are wonderful, because they change the task space so significantly that we say they have “emerging abilities”. Those abilities emerged without the need of many training examples but just a task instruction and/or a few demonstrations (i.e., zero-shot, few-shot):


Tasks that NLP models **CAN** make satisfactory performance
sentiment
essay writing
SAT
Leetcode
story telling
IE
summarization
translation
GRE Q
GRE V
MRC
commonsense reasoning
QA
chatbot
AMC
?
Tasks that NLP models **CANNOT** make satisfactory performance

The green zone is now too crowded, the red zone is nearly empty, and Microsoft sees “Sparks of Artificial General Intelligence” behind this new task space. Why are the LLMs able to now put the challenging tasks into the green zone? Because for this set of tasks, language is the universal interface that humans and computers use to describe problem samples and their solutions. Because humans use language to do so, there are massive problem and solution samples on the Web that were crawled for the development of LLMs. Because Transformers allow computers to efficiently process language data, the models were able to learn to solve the tasks via pre-training on the crawled data. Though pre-training used no explicit, clean supervisions on specific tasks, it observed and learned implicitly many examples in an autoregressive way. The zero-shot / few-shot in-context learning was not the first time the LLMs saw the task examples.

Therefore, if we really want to find a task in the red zone, the clues are straightforward:

(1) The tasks are so “novel” that there are little problem and solution examples on the Web. Find the tasks that the LLMs have never seen, even implicitly, a problem or solution example during pre-training.

(2) The task inputs are so complex that language, and not even a combination of language and image, could not be an effective interface. Complex supervised learning models are needed to learn sufficiently from the complex task inputs. And humans were able to process the complex task inputs efficiently.

One of the examples is **crossword puzzles**. Web pages may have crossword puzzles, however, those may have been dropped from the pre-training due to the poor quality as “language data.” Also, the data structure of puzzles is very complex. Humans utilize the complex puzzle structure to infer the missing words/letters. So, it is highly possible that this task is in the **red zone**. However, in my opinion, it’s not interesting. And we can find numerous tasks following the aforementioned two clues.

Another example is **word riddles**. This does not follow the two clues. Web pages may have a great number of word riddles. And most of them can be purely in natural language. So it may be interesting to test the abilities of LLMs on word riddles. This task requires the solvers (humans or models) to have commonsense knowledge, world knowledge, language skills, etc.

The public interest in testing LLMs is not on the tricky games. Instead, people are curious whether the LLMs can be immediately used in (new) industrial applications (e.g., mental health, education), and if yes, what the specific applications/tasks are. Also, prior to the deployment, people are concerned about safety challenges. Compared to word riddles, these explorations would create higher societal impacts and economic values.

In summary, here comes the principles I suggest of benchmarking LLMs:

Principle 1: Tasks on the green-red zones boundary. Picking a task in pure green zone (e.g., too similar with the tasks that LLMs have demonstrated on) or pure red zone (e.g., too “special” to have any examples implicitly in pre-training) would be less meaningful.

Principle 2: Tasks that have broader impacts beyond ML/AI/CS. Build a benchmark by experts in a domain relevant to humanity and society.

Principle 3: Tasks whose input information can be completely described in language. LLMs could hardly achieve a better performance than small supervised learning models on the tasks of too complex data structures.

about 750 words
March 29, 2023 (created)
March 30, 2023 (completed)