Evaluation of Large Language Models

Arturo Fredes Cáceres
Universitat de Barcelona
afredeca7@alumnes.ub.edu

Dafni Tziakouri
Universitat de Barcelona
dtziaktz7@alumnes.ub.edu

Abstract

The surge in popularity of Large Language Models (LLMs) across academic and industrial domains is attributed to their unparalleled performance across diverse applications. As LLMs continue to assume a pivotal role in research and everyday applications, their evaluation becomes increasingly crucial, extending beyond the task-specific level to encompass societal implications and potential risks. In this paper we will explore the comprehensive evaluation of LLMs through two distinct methods, and also introduce a new approach. Initially, we provide a brief overview of LLMs, delving into their evolution and highlighting the evolution of evaluation criteria over preceding years. Secondly, we will discuss about what tasks we evaluate LLMs to show their performance. Thirdly, we delineate the two evaluation methods, offering insights into their applicability and significance. Finally, we augment this exploration adding a different point of view, teleological approach, enhancing the relevance and timeliness of our evaluation framework.

1 Introduction

Language models (LMs) [1] constitute computational frameworks endowed with the ability to comprehend and generate human language. These models possess a transformative capacity, enabling them to predict the likelihood of word sequences or create novel text based on a given input. Among the most noteworthy iterations of LMs are the Large Language Models (LLMs), exemplified by GPT-4(OpenAI,2023)[2], and PaLM (Chowdhery et al., 2022) [3]. Distinguished by their large parameter sizes and exceptional learning capabilities, these advanced LLMs receive a text snippet as input and subsequently generate additional text as output.

The evaluation of LLMs is of paramount significance for several compelling reasons. Firstly, it gives us a comprehensive understanding of the strengths and weaknesses inherent in LLMs. Secondly, robust evaluation methodologies serve as invaluable tools for enhancing human-LLMs interaction, thereby inspiring advancements in interaction design and implementation. Thirdly, the widespread applicability of LLMs underscores the critical need to ensure their safety and reliability, particularly in sectors with inherent safety sensitivities, such as financial institutions and healthcare facilities. Consequently, the development of rigorous LLM evaluation protocols assumes a pivotal role in shaping the future trajectory of LLMs.

To elucidate the performance, strengths, and weaknesses of LLMs, a multifaceted evaluation approach is essential. Tasks encompassing natural language processing, robustness, ethics, biases, social sciences, and various other applications collectively contribute to a comprehensive assessment of LLMs. This multifaceted evaluation strategy not only enables a better understanding of the models but also serves as a foundation for refining and advancing their capabilities in diverse domains.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

2 What to evaluate

Within this section, we classify existing tasks into the following categories: natural language processing, robustness, ethics, biases, and trustworthiness, social sciences, natural science and engineering, medical applications, agent applications. For the scope of this document, we have chosen to address the following 2 categories: natural language processing, ethics, biases, and trustworthiness.

2.1 Natural language processing

Comprising a broad range of tasks, natural language understanding endeavors to enhance comprehension of input sequences. This section provides a summary of recent assessments of Language Model Evaluations, examining them from various perspectives.

Analyzing and interpreting text to discern emotional tendencies, sentiment analysis commonly involves binary (positive and negative) or triple (positive, neutral, and negative) class classification. The assessment of sentiment analysis tasks is a prevalent focus. Notably, ChatGPT surpasses traditional sentiment analysis methods in predictive accuracy and closely approaches the performance of GPT-3.5. Additionally, in the realm of fine-grained sentiment and emotion cause analysis, ChatGPT demonstrates exceptional proficiency. Language Models, including ChatGPT, have consistently exhibited commendable performance in sentiment analysis tasks.

Text classification and sentiment analysis, while interconnected, encompass distinct domains. Text classification extends beyond sentiment analysis to incorporate the processing of various texts and tasks. In general, Language Models (LLMs) exhibit proficient performance in text classification, showcasing adaptability in unconventional problem settings as well.

Semantic understanding pertains to grasping the meaning of language and its associated concepts. This encompasses the interpretation and comprehension of words, phrases, sentences, and the relationships between them. Semantic processing delves beyond surface-level comprehension, aiming to understand the underlying meaning and intent. Unfortunately, Language Models (LLMs) exhibit subpar performance in tasks related to semantic understanding.

2.2 Ethic, Bias, and Trustworthiness

The assessment of LLMs covers vital dimensions including ethics, biases, and trustworthiness. These factors have become progressively significant in conducting a comprehensive evaluation of LLM performance.

2.2.1 Ethics and Bias

Large Language Models (LLMs) have been observed to internalize, disseminate, and potentially amplify detrimental information present in their training datasets. This often includes toxic language such as hate speech, and insults, as well as social biases like stereotypes directed towards individuals with specific demographic identities (e.g., gender, race, religion, occupation, and ideology).

Beyond social biases, assessments of LLMs through tools like the Political Compass Test and MBTI test have indicated a tendency towards progressive views and an ENFJ (Extraversion, Intuition, Feeling Judging) Myers-Briggs personality type . Additionally, LLMs like GPT-3 have been identified to exhibit moral biases in alignment with the Moral Foundation theory. These ethical concerns pose significant risks that could hinder the widespread deployment of LLMs and potentially have a profoundly negative impact on society.

2.2.2 Trustworthiness

Beyond ethics and bias, certain research endeavors address additional concerns related to the trustworthiness of Large Language Models (LLMs). While LLMs exhibit the ability to produce coherent and seemingly factual text, there is a potential for the generated information to contain factual inaccuracies or statements lacking a basis in reality, commonly referred to as hallucination. Examining and understanding these issues are pivotal in refining the training methods of LLMs, with the aim of minimizing the occurrence of hallucinations. This research contributes to enhancing the overall reliability and accuracy of the generated content.

3 How to evaluate LLMs

In this section, we present two widely employed evaluation methodologies: automatic evaluation and human evaluation. To distinguish and comprehend the disparity between these two methods, it is essential to clarify that their categorization is contingent upon whether the evaluation criterion can be automatically computed.

3.1 Automatic evaluation

Automated assessment of Language Model Models (LLMs) stands as a prevalent and widely favored evaluation approach, typically employing standard metrics, indicators, and assessment tools to gauge model performance. Metrics like accuracy and BLEU [4] are commonly utilized in this method. For instance, the BLEU score serves as a quantifiable measure for assessing the similarity and quality between the text generated by the model and a reference text in tasks like machine translation. This evaluation protocol has gained widespread adoption in existing evaluation endeavors due to its objectivity, automated computation, and simplicity. Consequently, deterministic tasks like natural language understanding and mathematical problem-solving frequently embrace this evaluation approach. The fundamental principle of automated evaluation aligns with other AI model evaluation processes, wherein standard metrics are employed to calculate specific values under these metrics, serving as indicative measures of model performance.

3.2 Human evaluation

The advancing capabilities of Language Model Models (LLMs) have transcended conventional evaluation metrics in the realm of general natural language tasks. Consequently, in certain non-standard scenarios where automated evaluation falls short, human evaluation emerges as a logical alternative. This is particularly evident in open-generation tasks where embedded similarity metrics prove insufficient, making human evaluation a more dependable option. Human evaluation of LLMs involves assessing the quality and accuracy of model-generated results through active human participation.

In the manual evaluation process for LLMs, evaluators, comprising experts, researchers, or ordinary users, are typically invited to assess the outcomes generated by the model. A notable instance of such human-centric evaluation is the seminal work by Bubeck et al.[5], who conducted a series of human-crafted tests using GPT-4. Their findings revealed that GPT-4 performs closely to, and in some cases even surpasses, human performance across multiple tasks. This form of evaluation necessitates human evaluators to physically test and compare the models' performance, extending beyond the confines of automated evaluation metrics. It is essential to note that even human evaluations can exhibit high variance and instability, often attributed to cultural and individual differences.

3.3 Automatic Vs Human evaluation

While specific generation tasks may adhere to particular automated evaluation protocols, human evaluation holds greater preference in these scenarios due to the inherent potential of generation surpassing standard answers. In contrast to automatic evaluation, manual assessment aligns more closely with real-world application scenarios, offering a more comprehensive and accurate feedback mechanism. Unlike human evaluation, automated evaluation does not necessitate extensive human involvement, resulting in cost and time savings. In practical applications, a balanced consideration of these two evaluation methods is essential, taking into account the specific circumstances and requirements of the task at hand.

4 Sentiment Analysis Experiment

As a practical example, we did an experiment on sentiment analysis, which can be automatically evaluated and highlights the natural language processing capabilities of LLMs. The code can be found in our Github repository [6].

The experiment consisted in classifying 300 Trip Advisor reviews into positive, negative or neutral. We used the ratings as labels and classified as follows:

$$[4,5] \leftrightarrow positive \ ; \ [3] \leftrightarrow neutral \ ; \ [4,5] \leftrightarrow negative$$

There were 100 examples of each class. The model used was GPT-3.5 with temperature set to 1. In the prompting technique we used a few-shot strategy (giving examples of the task), asked the model to reason the classification before giving it (Chain of Thought) and leveraged several iterations (Autoconsistency). The accuracy we obtained was of 68%. In Figure 1 we can observe the classifications produced by GPT.

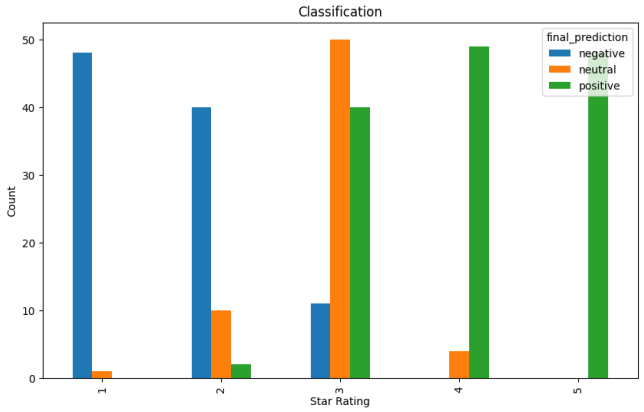


Figure 1: Sentiment analysis classifications produced by GPT-3.5 at temperature 1 for each rating. We divided 300 Trip Advisor reviews into positive, negative or neutral. using ratings as labels: $[4,5] \leftrightarrow positive \ ; \ [3] \leftrightarrow neutral \ ; \ [4,5] \leftrightarrow negative$. There were 100 examples of each class. The model used was GPT-3.5 with temperature set to 1. For the prompting technique we used a few-shot strategy (giving examples of the task), asked the model to reason the classification before giving it (Chain of Thought) and leveraged several iterations (Autoconsistency). In the image we can see the classifications produced by the model for each rating.

5 Teleological approach

On previous sections we have commented on what is being evaluated and how. This evaluations are mainly centered on tasks where the model's and human abilities intersect, carrying the risk of emphasizing the strengths of these models while not adequately exposing their weaknesses.

In [7] the authors argue that in order to understand LLMs, it is necessary to consider the problem that they were trained to solve as well as the architecture and training data used when evaluating their performance. The task in particular is autoregressive next-word prediction of text from large collections of text data, such as books, articles, and web pages. In a second training phase the model's behavior is aligned with human preferences by using instruction tuning. These models are neural networks, specifically, transformers, so the strategies they adopt to solve the previous problem are limited to the ones available in this kind of architecture.

Acknowledging the pressures imposed by this setting enables us to anticipate the tactics that Large Language Models (LLMs) will employ, allowing us to understand the circumstances under which they will be effective or fall short. This approach is called the teleological approach because it focuses on explaining the behavior of a system in terms of its goal (telos in Greek).

As we have commented, LLMs are often used for purposes that differ from next word prediction. This includes a wide range of tasks like text generation, summarization, translation or mathematical reasoning and coding. When a system is is initially tailored for a specific function and is later repurposed for something else, its original design can affect the system's characteristics in ways that might seem illogical when only the new purpose is taken into account. There are plenty of examples of this in biology, for example, our bipedal locomotion which was adapted from quadrupedal organisms, can today be the cause of lower back pains. There are analogous effects due to the mismatch

between the nature of LLMs (stochastic next word predictors) and their actual uses. Three factors that influence their accuracy throughout different tasks are: the probability of the task to be performed, the probability of the target output, and the probability of the provided input. "Probability" relates to the frequency of a task, element or relationship of elements appearing in the training set, and LLMs will perform better when these probabilities are high even in deterministic settings where probability should not matter.

In [7] the authors provide evidence of GPT 3.5 and GPT 4 present these biases. They do so by evaluating the model on 11 different tasks like article swapping, shift ciphers, creating acronyms, sorting lists of words or calculating linear functions. We chose the task of applying linear transformations and observed similar results. In this particular example, we are using GPT, a "statistical next-word prediction system" as a "math problem solver". The code to this experimaent can also be found in our Github repository [6].

5.1 Sensitivity to task frequency

Task frequency is related to how many examples of the task will have been encountered by an LLM during training. When neural networks have access to a greater amount of examples of a task in the training set they usually perform better. Therefore it is expected that LLMs will perform better in tasks that are more common on the Internet and the corpus used for training.

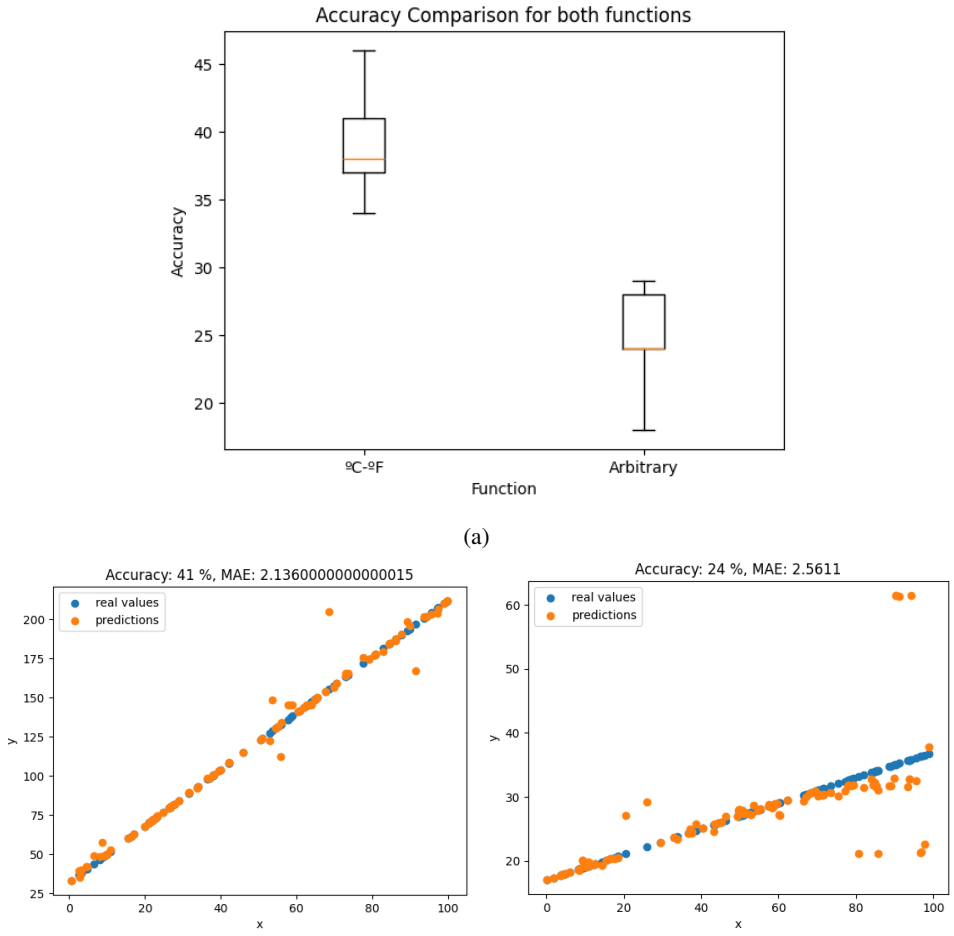
To test this we chose the two following linear functions, which appear to have the same complexity:

$$f_1(x) = \frac{9}{5}x + 32 \tag{1}$$

$$f_2(x) = \frac{1}{5}x + 17 \tag{2}$$

At first, they could look as two arbitrary linear functions, but Equation 1 is the conversion between Celsius and Fahrenheit degrees, so it should appear frequently in the training set, whilst 2 is an arbitrarily chosen function. We asked GPT 3.5 to apply this function to 100 points with one decimal position which were uniformly taken from the interval [0,100]. We used a zero-shot prompting technique, which consists in asking the model to perform a task that it has not explicitly been trained on. It does this based on the knowledge it has already acquired during its initial training. Since this task is deterministic, we set the temperature of the model to zero to get consistent answers.

We repeated this experiment 5 times, obtaining an accuracy of $39.2 \pm 4.1\%$ for Equation 1 and $24.6 \pm 3.9\%$ for Equation 2, showing evidence of sensitivity to task frequency. In Figure 2(a) we can see the accuracy obtained in the experiments for each function and in Figure 2(b)(c) a comparison between the function and the output from GPT-3.5.



(a)

(b)

(c)

Figure 2: We chose two similar linear functions, Equation 1 is the conversion between Celsius and Fahrenheit degrees, so it should appear frequently in the training set while Equation 2 is an arbitrarily chosen linear function. We asked GPT-3.5 to apply these functions to 100 points with one decimal position which were uniformly taken from the interval [0,100]. Experiments were repeated 5 times. There is a clear improvement in model performance on this task when the linear function is common in the training set. (a) Accuracy obtained from querying GPT-3.5 on the calculation of linear functions. (b)Outputs vs true values for one experiment applying the conversion betweenºC and ºF (c) Outputs vs true values for one experiment applying an arbitrary linear function

5.2 Sensitivity to input/output probability

There are indications that LLMs have a tendency toward generating sequences of words with higher probabilities, meaning that their performance will get worse when the correct output is not common.

To motivate this point more formally, we can conceptualize the task of an LLM as identifying the output:

$$max\ P(input \mid output)$$

where the input is the first chunk of a word sequence, and the output is the sequence's continuation. By Bayes' rule, this problem can be reformulated:

$$max\ P(input \mid output)P(output)$$

Therefore, if there are multiple outputs for which $P(input \mid output)$ is nonzero, the LLM's predictions will be influenced by $P(output)$,showing a bias for candidates with higher $P(output)$. If the LLM correctly captured the deterministic cases, this should not have an effect, but it clearly does.

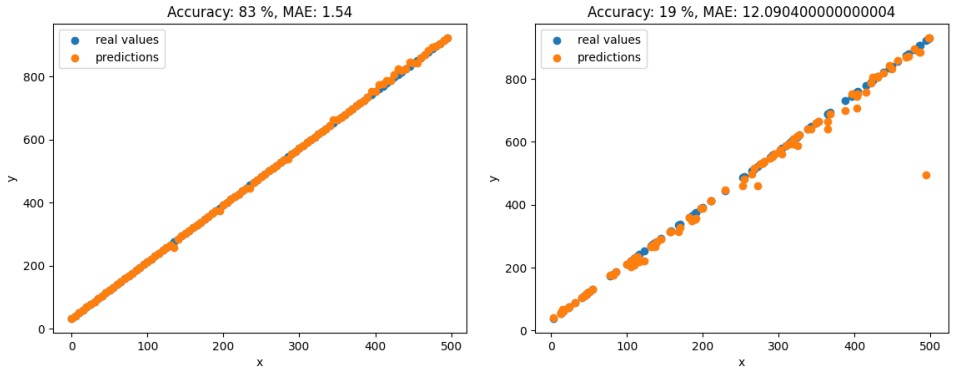
Furthermore, LLMs will have had less experience with low-probability strings than high-probability ones, so input probability is also relevant.

To continue with our experiment, we used Equation 1 but used two different set of points. The first set of points were the first 100 multiples of 5, which produce an integer output (that we suppose more probable). The second set of points consisted in uniformly taking 100 random numbers from [0,500]. Our data sets looked like this:

$$Data_1 = [0.0,\ 5.0,\ 10.0,\ ...,\ 495.0]$$

$$Data_2 = [3.2,\ 12.5,\ 17.6,\ ...,\ 497.8]$$

For the first data set we got an accuracy of 83 %, while for the second we got an accuracy of 19 %, showcasing the effect of input/output probability on accuracy. In Figure 3 we can see how clearly for the high probability case the fit to the data is better.



(a) (b)

Figure 3: For these experiment we used Equation 1, the conversion between Celsius and Fahrenheit degrees, and used two different data sets. The first data set were the first 100 whole multiples of five, which will produce integer outputs. The second data set were arbitrary numbers with one decimal between [0,500]. We considered both the input and output probability of the first data set to be higher. We asked GPT-3.5 to apply this function to both data sets (a) Accuracy obtained from querying GPT-3.5 on the calculation of linear functions. (b)Outputs vs true values for applying the conversion betweenºC and ºF on multiples of 5(c) Outputs vs true values for applying the conversion betweenºC and ºF on randomly chosen numbers.

6 Conclusions

To conclude, a comprehensive assessment of Language Models (LLMs) can be conducted across various tasks, as previously discussed, enabling a thorough examination of their performance in diverse aspects. The two primary methods for such evaluations are automatic and human evaluations, each carrying its unique advantages and challenges. The choice between these methods depends on the specific aspects we aim to evaluate. Moreover, it is important to leverage different points of view when creating evaluation methods that will truly reveal the strengths and weaknesses of LLMs. We should not lose sight of what these models really are, statistical next word predictors, even though they exhibit good performances on other tasks. This perspective helps in maintaining realistic expectations and understanding of their capabilities and limitations.

References

[1] Yupeng Chang et al. *A Survey on Evaluation of Large Language Models*. 2023. arXiv: 2307.03109 [cs.CL].

[2] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[3] Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: 2204.02311 [cs.CL].

[4] Kishore Papineni et al. *Bleu: a Method for Automatic Evaluation of Machine Translation*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA, July 2002. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[5] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712 [cs.CL].

[6] *GitHub repository of the paper*. URL: https://github.com/DaphneDjiakouri/LLMs-Evaluations.

[7] R. Thomas McCoy et al. *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve*. 2023. arXiv: 2309.13638 [cs.CL].