**ML Workshop: Evaluation of LLMs**

**Abstract:** The surge in the 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. This paper explores the comprehensive evaluation of LLMs through three distinct methods, one of which introduces 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 delineate the three evaluation methods, offering insights into their applicability and significance. Finally, we augment this exploration with a contemporary experiment, enhancing the relevance and timeliness of our evaluation framework.

**Introduction:**

Language models (LMs) [add reference] constitute computational frameworks endowed with the remarkable ability to comprehend and generate human language. These models possess a transformative capacity, enabling them to predict the likelihood of word sequences or engender 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), and PaLM (Chowdhery et al., 2022). Distinguished by their expansive 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 affords 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 nuanced understanding of the models but also serves as a foundation for refining and advancing their capabilities in diverse domains.

**How to evaluate LLMs:**

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

**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 [142] 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 understanding and mathematical problem-solving frequently embrace this evaluation approach. The fundamental principle computation, and simplicity. Consequently, deterministic tasks like natural language 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.

**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. [reference], 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.

**Automatic Vs Human Evaluation:**

While specific generation tasks may adhere to particular automated evaluation protocols, human evaluation holds greater favorability 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.

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, the teleological approach adds an intriguing dimension, as it delves into explaining a system's behavior by elucidating its underlying goals. This multifaceted evaluation framework enhances our understanding of LLMs, facilitating a more nuanced analysis of their capabilities and limitations.