Large Language Models: A Succinct Evaluation

Ilaria Curzi
Universitat de Barcelona
icurzicu125@alumnes.ub.edu

Georgia Zavou
Universitat de Barcelona
gzavouza108@alumnes.ub.edu

Theodoros Lambrou
Universitat de Barcelona
thlmbrl67@alumnes.ub.edu

Abstract

The rising popularity of Large Language Models (LLMs) in both academic and industrial spheres can be attributed to their exceptional performance across a wide range of applications. As LLMs become increasingly integral to research and everyday use, the need for thorough evaluation grows, expanding beyond task-specific metrics to include societal impacts and potential risks. This paper aims to comprehensively evaluate LLMs through two established methods while introducing a novel approach.

First, we provide a concise overview of LLMs, tracing their development and examining how evaluation criteria have evolved over the years. Next, we discuss the types of tasks used to assess LLM performance. We then detail two specific evaluation methods, shedding light on their significance and practical applications. Finally, we introduce a new perspective—the teleological approach—offering an enriched framework to enhance the relevance and timeliness of LLM evaluation.

1 Introduction

Language models (LMs) have undergone significant evolution since the 1950s when Shannon applied information theory to model human language. Early statistical approaches, like n-gram models, laid the groundwork for natural language processing (NLP) tasks such as speech recognition and machine translation. Over time, the development of neural language models (NLMs) and pre-trained language models (PLMs) introduced advanced techniques like word embeddings and task-agnostic training. These models act as computational tools that can understand and generate human language, predicting word sequences or creating new text based on input. The most advanced versions of these models are Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023) and PaLM (Chowdhery et al., 2022). LLMs stand out due to their immense parameter sizes and exceptional learning capabilities, allowing them to process text snippets and produce coherent, contextually appropriate outputs. As described in [2] (Yupeng Zhao et al., "A Survey of Large Language Models," 2023, arXiv:2303.18223 [cs.CL]), LLMs are characterized by their pretraining on massive datasets and unique emergent abilities. These include in-context learning, where they can handle new tasks from a few examples; instruction following, which allows them to generalize to new tasks after tuning; and multi-step reasoning, where complex problems are broken into intermediate steps using methods like chain-of-thought prompting. Additionally, LLMs can be enhanced with external tools and knowledge sources, improving their ability to adapt to dynamic environments and practical applications. As LLMs form the backbone of general-purpose AI agents, their evaluation is becoming increasingly important. Rigorous assessments are crucial for identifying their strengths and weaknesses, ensuring safety and reliability, particularly in sensitive areas like healthcare and finance. As noted in [3] (Yupeng Chang et al., "A Survey on Evaluation of Large Language Models," 2023, arXiv:2307.03109v1 [cs.CL]), evaluation methods not only improve human-LLM interaction but also guide advancements in their

design. Comprehensive evaluation strategies that address NLP tasks, robustness, ethics, biases, and societal impact are essential for refining LLMs and unlocking their full potential.

2 What to Evaluate

This section categorizes evaluation tasks for Large Language Models (LLMs) into two critical areas: Natural Language Processing (NLP) and Explainability and Interpretability. These categories are central to assessing LLMs' capabilities, understanding their behavior, and improving their usability.

2.1 Natural Language Processing

Natural Language Processing (NLP) includes a wide range of tasks, with natural language understanding focusing on improving how input sequences are comprehended. This section reviews recent evaluations of Language Models (LLMs) from various angles. One prominent area of study is sentiment analysis, which identifies emotional trends in text using either simple classifications (e.g., positive or negative) or more detailed ones (e.g., positive, neutral, negative). As demonstrated in [1] (Yupeng Zhao et al., "Explainability for Large Language Models: A Survey," 2023, arXiv:2309.01029 [cs.CL]), ChatGPT has shown higher predictive accuracy compared to traditional sentiment analysis methods and performs on par with GPT-3.5. Moreover, ChatGPT excels in detailed sentiment and emotion cause analysis, demonstrating its strength in this domain. Language models like ChatGPT consistently deliver impressive results in sentiment analysis tasks. Although closely related to sentiment analysis, text classification is a broader task that involves categorizing text for different objectives. LLMs have proven highly adaptable, effectively managing diverse and unconventional classification challenges [1]. Semantic understanding, on the other hand, involves grasping the meanings and relationships of words, phrases, and sentences. It aims to go beyond surface-level interpretation to understand deeper meanings and intentions. However, as highlighted in [3], LLMs often struggle with semantic understanding tasks, revealing an area where further improvement is necessary.

2.2 Explainability and Interpretability

Explainability and interpretability are essential for evaluating and understanding how Large Language Models (LLMs) process information and make decisions. As shown in [1] (Yupeng Zhao et al., "Explainability for Large Language Models: A Survey," 2023, arXiv:2309.01029 [cs.CL]) and [3] (Yupeng Chang et al., "A Survey on Evaluation of Large Language Models," 2023, arXiv:2307.03109v1 [cs.CL]), two main types of explanations are commonly used: local explanations, which focus on specific predictions to clarify why a model made a particular decision, and global explanations, which provide an overall understanding of the model's decision-making process and knowledge representation. Techniques such as attention visualization and layer analysis are instrumental in interpreting LLMs. Attention visualization identifies the parts of the input that the model prioritizes during processing, while layer analysis examines how input data is transformed at different stages within the model. The effectiveness of these explainability methods is assessed using key metrics such as fidelity, which measures how accurately the explanation reflects the model's decision-making, and interpretability, which evaluates how easily humans can understand the explanation. These methods are not only valuable for understanding how models work but also play a critical role in debugging, identifying errors or biases, and improving overall model performance [1]. However, explaining LLMs presents unique challenges due to their complexity and size, which surpass that of simpler machine learning models. Despite these difficulties, as noted in [2], there are significant opportunities to develop more intuitive and user-friendly techniques that can keep pace with the rapid advancements in LLM technology. These efforts are essential for ensuring the transparency, reliability, and trustworthiness of LLMs in real-world applications.

3 How to Evaluate LLMs

In this section, we present a comprehensive approach to evaluating Large Language Models (LLMs), which accounts for their wide-ranging applications, emergent capabilities, and potential limitations. Evaluation strategies can be broadly categorized into traditional methods, such as automatic evaluation and human evaluation, as well as newer, more nuanced approaches designed to assess the advanced functionalities and diverse use cases of LLMs.

3.1 Automated Evaluation

Automated evaluation is a key method for assessing LLMs, using metrics like accuracy, BLEU, ROUGE, and F1-score to measure performance in tasks such as translation, summarization, and text classification. As highlighted in [3] (Yupeng Chang et al., "A Survey on Evaluation of Large Language Models," 2023, arXiv:2307.03109v1 [cs.CL]), recent research emphasizes the importance of expanding these metrics with benchmarks like MMLU, which tests generalization and domain-specific expertise across a variety of tasks, and BigBench, which evaluates creative and reasoning abilities beyond traditional measures. Automated evaluations are highly effective for well-defined tasks and large-scale testing because they provide consistent and objective results. However, as described in [2], these methods often fall short in capturing complex behaviors like multi-step reasoning or contextual understanding, highlighting the need for additional evaluation approaches.

3.2 Human Evaluation

Human evaluation is indispensable for open-ended tasks and qualitative assessments, where creativity, coherence, and contextual appropriateness are key. Human evaluators assess attributes like fluency, logical consistency, and ethical alignment, which automated metrics cannot fully capture. Studies from [3] reveal that instruction-following ability and multi-step reasoning are best assessed through human-crafted tasks. For example, GPT-4 has been evaluated in scenarios where human judges rate its performance on tasks like essay writing, ethical reasoning, and code generation. However, human evaluations are costly and subjective, often influenced by individual and cultural differences [3].

3.3 Emerging Evaluation Paradigms

Advancements in LLM research have introduced new ways to evaluate these models. Firstly, Explainability-based metrics, such as attention visualization and counterfactual reasoning, are used to understand how LLMs make predictions, providing valuable insights into their decision-making processes [1]. This is particularly important in critical fields like healthcare and law. Secondly, Multi-modal benchmarks are also emerging, especially with models like GPT-4 Vision, which require testing across text, image, and even video inputs, going beyond traditional text-only evaluations. Additionally, robustness testing evaluates how LLMs perform when faced with challenges such as adversarial inputs, low-resource languages, or biased datasets, helping to ensure their reliability and fairness [3].

3.4 Real-World and Dynamic

Evaluations LLMs are being used more frequently in real-world applications, making it important to evaluate their performance in dynamic and practical environments. This includes interactive evaluations, where LLMs are tested on their ability to manage real-time interactions in tools like chatbots, virtual assistants, or decision-support systems. Another key area is ethical and societal impact assessments, which examine how LLMs handle sensitive topics to ensure they follow ethical guidelines and minimize potential harm to society [2].

3.5 Balancing Evaluation

Approaches While automated evaluations are essential for scalability and consistency, human assessments provide depth and context for complex or subjective tasks. A hybrid approach, combining traditional metrics with emerging techniques like explainability and real-world testing, is critical for capturing the full spectrum of LLM capabilities. Such a balanced strategy ensures that LLMs are evaluated comprehensively, addressing both technical performance and societal impact.

4 Teleological Approach to Evaluating Large Language Models

Evaluation methods for Large Language Models (LLMs) must consider the purposes these models are trained to serve and their broader applicability in tasks beyond their initial design. As noted in [1] (Yupeng Zhao et al., 'Explainability for Large Language Models: A Survey', 2023, arXiv:2309.01029) and [5] (Yue Zhang et al., 'LLMEval: A Preliminary Study on How to Evaluate Large Language Models', 2023, arXiv:2312.07398), LLMs such as GPT-4 or PaLM are pre-trained on large datasets with the objective of next-word prediction, but their emergent abilities allow them to tackle more complex tasks like reasoning, summarization, and coding. This approach to evaluation is referred to as teleological, as it analyzes models based on their design goals ("telos" in Greek) and the extent to which they fulfill their intended objectives.

LLMs often excel in tasks that align closely with their training data but struggle when applied to less frequent or unanticipated tasks. For example, as noted in [1] and [5] , their performance on mathematical problems demonstrates significant variance based on task familiarity. These findings underscore the importance of evaluating LLMs within the context of both their original design and their ability to generalize to novel scenarios.

4.1 Sensitivity to Task Frequency

LLMs demonstrate varying levels of success depending on the frequency with which tasks appear in their training datasets. For instance, tasks like temperature conversion (e.g., Celsius to Fahrenheit) are more effectively handled than arbitrarily selected linear functions. This discrepancy is rooted in the models' exposure to frequent patterns in their pretraining phase, which biases their accuracy in deterministic but rare tasks. An experiment highlighted in [1] tested GPT-3.5 on two linear functions of similar complexity, one representing temperature conversion and another arbitrarily chosen. The model's accuracy for the frequent task was significantly higher ((39.2% ± 4.1%) compared to the rare task (24.6% ± 3.9%). This suggests that the training frequency of specific tasks significantly influences model performance.

Figure 1 shows the model accuracy comparison.

Similarly, [5] demonstrates that high-frequency patterns in training datasets consistently lead to better accuracy in evaluation tasks.

4.2 Sensitivity to Input/Output Probability

In addition to task frequency, LLMs are influenced by the probability distribution of inputs and outputs. According to [1] and [5], LLMs prioritize high-probability sequences due to their stochastic training processes. This can result in reduced accurahincy for less common or low-probability inputs, even when deterministic relationships exist. An illustrative example from [1] involved testing GPT-3.5 on two datasets: one with integer outputs derived from multiples of 5 and another with randomly selected decimal points. Accuracy for the high-probability dataset reached 83%, while the low-probability dataset saw a significant drop to 19%.

Figure 2 shows the model accuracy for High vs Low Probability Datasets.

This behavior suggests that probability mismatches between training and evaluation tasks can adversely impact LLM performance.

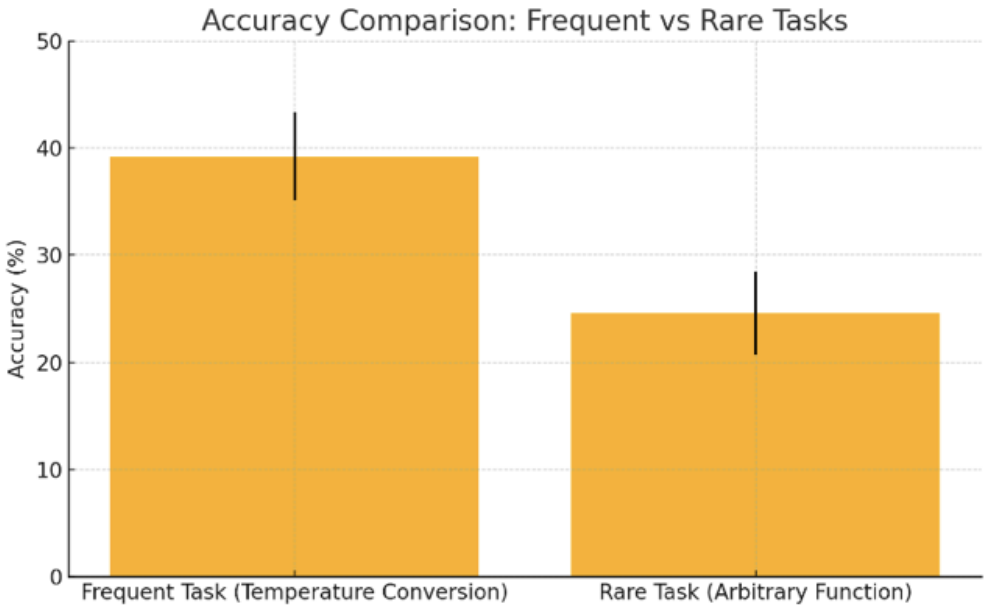

Figure 1: Model accuracy comparison between frequent tasks (temperature conversion) and rare tasks (arbitrary functions).

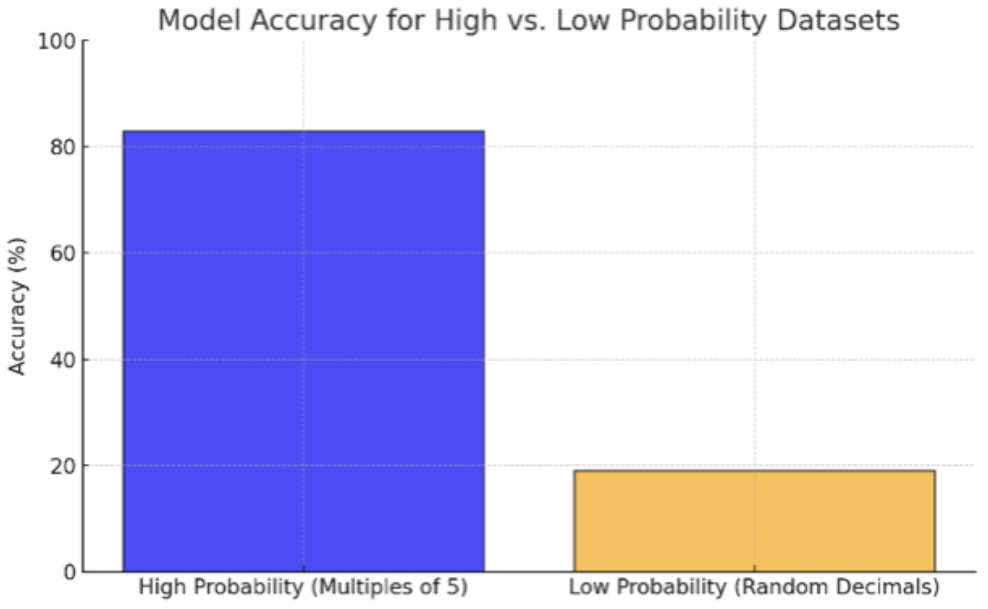

Figure 2: Model accuracy comparison for high-probability (multiples of 5) and low-probability (random decimals) datasets. Data adapted from [5] (Yue Zhang et al., 2023).

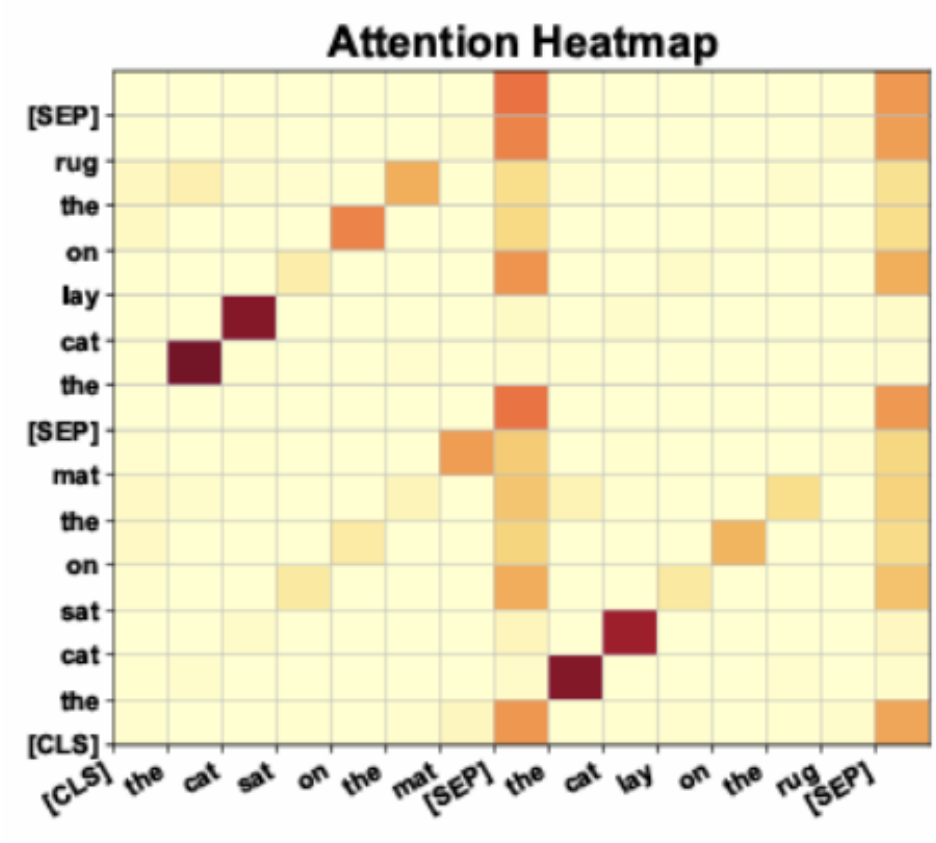

Figure 3: Example of an attention heatmap showing word priority during a text classification task.

5 Enhancing Explainability and Interpretability

Interpreting LLMs is a critical challenge given their complexity and emergent abilities. Local explanation techniques, such as attention visualization, have been proposed to provide insights into specific predictions by highlighting input elements most influential in a model's decision [1] . Meanwhile, global explanation methods, such as layer analysis, examine how data transformations occur across different stages of the model, offering a broader understanding of its internal mechanics. As shown in [1] and [4] (Sean Xie et al., 'Proto-lm: A Prototypical Network-Based Framework for Built-in Interpretability in Large Language Models', 2023, arXiv:2311.01732), attention visualization methods revealed how models prioritize syntactic elements in text classification tasks, while layer analysis uncovered hierarchical structures in language representation. However, existing techniques face limitations in scaling to the size and complexity of modern LLMs. Developing more user-friendly interpretability tools is crucial for ensuring transparency and trustworthiness in real-world applications.

Figure 3 shows an attention heatmap.

Metrics such as fidelity and interpretability are essential for assessing the quality of explainability methods. Fidelity measures how well explanations align with the model's internal logic, while interpretability evaluates their accessibility to human users [1]. High-fidelity explanations are particularly valuable in critical domains like healthcare, where understanding the rationale behind predictions is imperative for ethical and safe deployment.

Emerging frameworks, such as Proto-LM discussed in [4], integrate interpretability directly into the fine-tuning process, enabling LLMs to learn interpretable representations without sacrificing

performance. These innovations hold promise for bridging the gap between model complexity and user comprehensibility.

6 Conclusions

In conclusion, Language Models (LLMs) can be comprehensively assessed across a variety of tasks, facilitating a detailed evaluation of their performance across multiple dimensions. The primary approaches for such evaluations—automatic and human assessments—each offer distinct benefits and pose unique challenges. The choice of method depends on the specific aspects being examined. Additionally, incorporating diverse perspectives when designing evaluation frameworks is crucial to accurately identify the strengths and limitations of LLMs. It is essential to remember that, at their core, these models are statistical predictors of the next word, despite their impressive performance on various tasks. This understanding ensures realistic expectations and a clearer grasp of their true capabilities and constraints.

References

[1] Yupeng Zhao et al. Explainability for Large Language Models: A Survey. 2023. arXiv: 2309.01029 [cs.CL].

[2] Yupeng Zhao et al. A Survey of Large Language Models. 2023. arXiv: 2303.18223 [cs.CL].

[3] Yupeng Chang et al. A Survey on Evaluation of Large Language Models. 2023. arXiv: 2307.03109v1 [cs.CL].

[4] Sean Xie et al., 'Proto-lm: A Prototypical Network-Based Framework for Built-in Interpretability in Large Language Models', 2023, arXiv:2311.01732.

[5] Yue Zhang et al., 'LLMEval: A Preliminary Study on How to Evaluate Large Language Models', 2023, arXiv:2312.07398