International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Breaking Down the Metrics: A Comparative Analysis of LLM Benchmarks

Valentina Porcu * and Aneta Havlínová

Independent researcher.

International Journal of Science and Research Archive, 2024, 13(02), 777–788

Publication history: Received on 02 October 2024; revised on 11 November 2024; accepted on 14 November 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2209

Abstract

Due to the fast advancement in large language models (LLMs), the natural language processing (NLP) technology domain has witnessed a massive change. It has brought ground breaking developments in how machines understand and generate human language. However, even with these advancements, it remains hard to objectively compare and evaluate LLMs' performance because various benchmarks and metrics are used in the assessment. Such a method requires this study's writers to conduct a comparative analysis of various benchmark LLMs to explain the viability of different evaluation metrics. Analyzing the recognized standards, including GLUE, SuperGLUE, and SQuAD, reveals weaknesses and potential in the present evaluating systems. It embraces quantitative assessments of model performances and benchmarking and critical evaluations of benchmark designs and their scope. The paper discusses methodological issues in benchmarking research and explains that current measures are insufficient and lack comprehensive evaluation patterns. The conclusions of this research underline the significance of creating fully integrated and efficient reference sets that match the rate of innovation in LLM ability and inform other studies and the overall subsequent advancement of more sophisticated and flexible language models.

Keywords: LLM Benchmarks; BIG-bench; NLP Tasks; SQuAD; Model Comparison; NLP evaluation

1. Introduction

1.1. Background to the Study

NLP has seen a major shift thanks to state-of-the-art large language models that apply deep learning techniques and are trained on massive amounts of data. The self-attention mechanism begun by the Transformer architecture introduced by Vaswani et al. introduced one of the biggest changes in NLP. Case in point: BERT has developed LLMs by bringing bidirectional training of transformers and has taken vast enhancements in multiple NLP assignments (Devlin et al., 2018).

Nowadays, they are irreplaceable parts of multiple systems and services available in such domains as machine translation, sentiment analysis, information search, or conversational agents. For instance, BERT family technologies perform best in QA and NLI and demonstrate an important factor for future advancements in artificial intelligence technologies (Devlin et al., 2018; Liu et al., 2019). The effectiveness of these models is based on their capability of capturing the context, or otherwise, the sentiment and nuances of the human language that is so important for the creation of logically and semantically appropriate responses to a given input.

As with most, if not all, subfields of machine learning, it is now more important than ever to evaluate these models to get a clue about their performance. Reference planet metrics act as decisive FLOPs to assess and notate the effectiveness of various LLMs, thereby making it possible to quantify advancement in the field and guarantee that LLM models nicely transfer mastering from one task to an additional (Vaswani et al., 2017). For that reason, the openness of human

* Corresponding author: Valentina Porcu

language processing implicates certain concerns that have proven to be challenging regarding establishing benchmarks that can adequately measure the performance of LLMs. This is why traditional measures may need to be more effective in capturing the context, pragmatics, and world knowledge that a model possesses.

However, with models continuing to expand in terms of size and complexity, there is a crucial requirement for enhanced evaluation matrices. It has become clear now that current measures only sometimes provide enough parameters to thoroughly test the higher-order reasoning and language-processing skills that many contemporary LLMs boast of (Liu et al., 2019). This has led to a demand for new measures that can measure models out of what they can predict or generate based on various syntactic and semantic properties and various real-life use cases.

1.2. Overview of LLM Benchmarks

As is the case with most advanced AI systems, to facilitate LLM testing, we need performance benchmarks that are comprehensive enough to handle a diverse range of language tasks. One of the most effective ones now is the Stanford Question Answering Dataset (SQuAD), with the primary emphasis on the phenomena of machine reading comprehension describing how well a model answers the given questions relying on the provided context (Rajpurkar et al., 2016). SQuAD has played a crucial role in enhancing models that can comprehend and extract relevant information from a text, which has enhanced virtual assistance and information searching doors.

Another important reference is the Cross-lingual Natural Language Inference (XNLI) dataset, which tests how the model can perform inference across languages [19]. XNLI examines the ability of models to transfer reasoning skills from one language to another, contributing to multilingual NLU, and provides a way for languages to be made more accessible through AI.

Models are presented with 57 themes related to the humanities, social, and natural sciences. They are tested as abacus activities regarding world knowledge in the Massive Multitask Language Understanding (MMLU) test suggested by Hendrycks et al. (2020). MMLU brings into question an LLM's ability to reason and know across a variety of domains and invites the exercise of creative thinking.

These benchmarks are central in stabilizing performance measures to inform the designing of complex LLMs. As I establish in this paper, benchmarks such as SQuAD, XNLI, and MMLU offer varied and difficult test sets to facilitate model comparison and detect lingering performance disparities. It outlines specific areas of strength and weakness in models to guide subsequent research and development efforts and ensure that improvement in LLMs benefits turns into operational applications.

1.3. Problem Statement

The core research question discussed in this work is the instability and drawbacks of benchmark measures for large language models (LLMs). Such benchmarks may have differences in structure, scope, and evaluation criteria, so models and tools are compared and ranked differently. All these variations compromise cross-model comparison given that the type of benchmark's design, in this case, does influence performance measurements and, by extension, model performances. Moreover, many contemporary reference models need the flexibility to address the high dimensionality of today's LLMs, including their learning capability, interpretability, and ability to adapt to context variation and different language features. This limits performance standardization, which makes it challenging to track progress in the field accurately and prevents the development of models that will yield similar results consistently across various uses. Solving these issues remains crucial for creating the references that could also achieve the pace of constant changes in LLMs and be consistent in their evaluations.

1.4. Objectives

This study seeks to achieve the following objectives:

- Describe the flaws of benchmarking metrics used to measure the LLM performance at the current stage.
- Estimate a number of objectivized indexes and indicators to analyze their effectiveness regarding LLM performance evaluation.
- Using different benchmarks, it is possible to compare results between multiple LLMs and identify gaps in their performance.
- Suggest some improvements that would create wider benchmarking measures that would embrace all abilities of LLMs.

- Explain the need to coordinate the design of benchmark structures to offer more probability to comparisons and tracking of growth in the business.

1.5. Scope and Significance

In this review, the authors discuss GLUE, SuperGLUE, and SQuAD benchmarks, along with others introduced more recently for benchmarking advanced functionalities of LLMs. Instead, it targets evaluating the precise measurements used by these benchmarks, including the accuracy, F1-Score, and metrics based on comprehension of the model to the various tasks entailing NLP. Through these benchmarks, the study seeks to identify areas that are currently ignored or poorly handled in most evaluation protocols, with a special focus on testing LLM's ability to understand language, reason, and generalize. The importance of this study also transcends the scholarly arena; it draws possible practical concerns for industry and research domains where LLMs are being utilized in tasks involving accurate and dependable language comprehension. Overcoming these benchmarking challenges, the study Mines new avenues to design more flexible and higher-performance LLMs to support further developments of NLP technologies and expand the existing understanding of the potential to build models suitable for actual application.

2. Literature Review

2.1. Emergence of New Imperative for Language Model Benchmarks

It is a breakthrough in natural language processing (NLP) that benchmarks of language models have been the key to evaluating the models. The first benchmarks, which mainly included text classification and sentiment analysis, formally provided little information about a model's ability to understand language (Pang & Lee, 2005) universally. While NLP applicative areas were becoming increasingly sophisticated, there was a clear requirement for standard datasets that could capture additional features of language processing models.

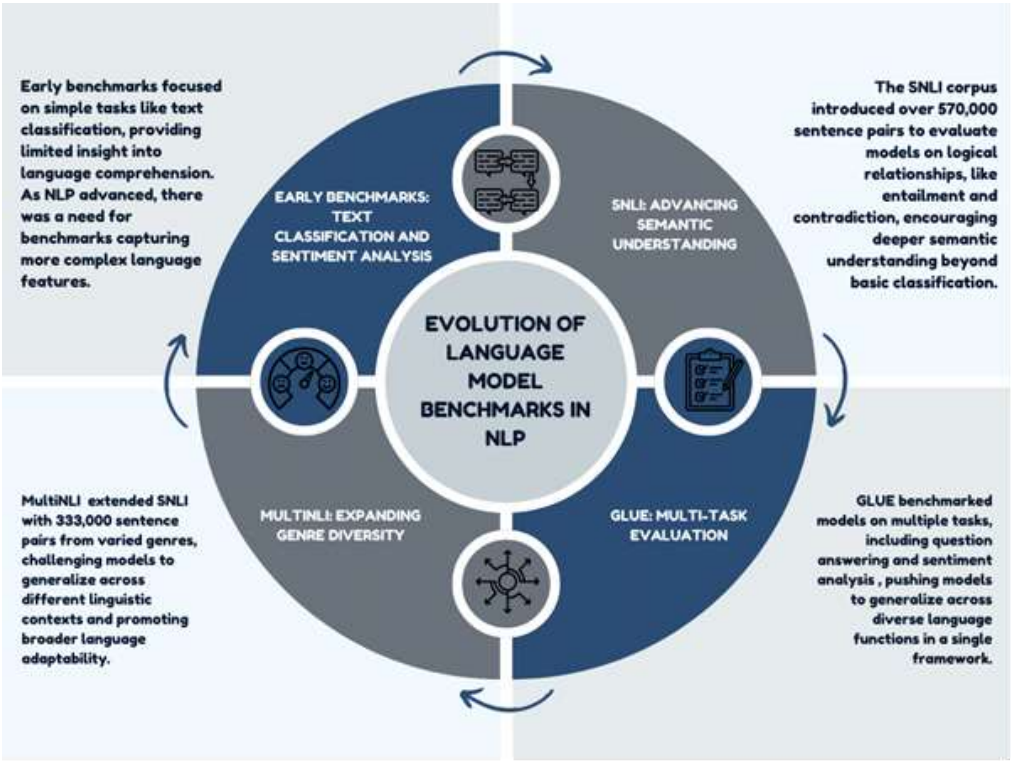

Figure 1 Evolution of Language Model Benchmarks in NLP

One such important step was done by Bowman et al. (2015), who proposed the Stanford Natural Language Inference (SNLI) corpus. The SNLI corpus contained a rich annotated set for the training and testing models of natural language inference (NLI) task to discover the logical relationship between two sentences – entailment, contradiction, or neutral (Bowman et al., 2015). However, SNLI has more than 570,000 sentence pairs, and such large amounts of data allowed us to train models to potentiate deeper semantics.

Following SNLI, many other benchmarks, such as the Multi-Genre Natural Language Inference (MultiNLI) corpus, were proposed to extend the program and embrace more than generic writing styles (Williams et al., 2018). MultiNLI provided three hundred and thirty-three thousand, four hundred and thirty different sentence pairs, thereby testing the ability of models on various forms of writing. This diversification was critical in forcing models towards a better understanding of language.

The progression of benchmarks went up with the introduction of all-encapsulating benchmarks like the General Language Understanding Evaluation or just GLUE with an assortment of NLP tasks in a single evaluation stage (Wang et al., 2018). These comprise question-answering, sentiment analysis, and linguistic acceptability, all of which are part of the more general GLUE. This movement from one-label classification to multiple-label, many-task tests highlight the development in language models and the improved demand for tasks that better identify their capabilities.

2.2. Existing Benchmarks: GLUE, SuperGLUE, and Beyond

The General Language Understanding Evaluation benchmark released in 2018 offers a systemic solution for measuring language models on various NLP tasks. GLUE consists of nine sub-tasks: textual entailment, sentiment analysis, AG-news, Wiki-Text, Bootstrap, EP requisites, CoLA, paused, and CF-QQP. This benchmark offered a mutual assessment platform, facilitating objective assessment among models and driving the development of this area forward at a much faster pace (Wang et al., 2018).

However, as the models advanced, they became limited to the performance ceiling of GLUE tasks and showed that the benchmark could not effectively test those systems (Wang et al., 2019). In response, a succession called SuperGLUE was built with eight more challenging language-understanding tasks that speak to the inadequacies of GLUE (Wang et al., 2019). There are various tasks in SuperGLUE, such as the ones below; these consist of commonsense reasoning and abstract text understanding, which need more reasoning.

SuperGLUE also brought in more refined assessment tools, such as a human analog, to provide a standard against which the machines could perform (Wang et al., 2019). While it heightened the distinctions between human and analytical scores as the latter increased, its addition led to this observation. However, the SuperGLUE evaluation has the same weaknesses as always: there are model shortcuts to the answer, datasets can be biased, and while the models pass the tests by performing various computations, there needs to be evidence that models understand what they are looking at.

The DecaNLP is another case of such benchmarks that extend the use of GLUE and SuperGLUE to evaluate models on the performance of multitask learning introduced by (McCann et al., 2018). The DecaNLP consists of ten NLP tasks formatted as questions to challenge models to learn and apply knowledge across tasks. These benchmarks indicated that there are further attempts to establish more strict and accurate evaluation protocols for testing the performance of developed language models because of their constantly increasing possibilities.

2.3. Metrics for LLM Evaluation

As a result, such methods arise to measure the performance of LLMs in line with the tasks they are intended to solve. For categorization problems, appropriately quantifiable performance measures include accuracy and F1-score, which estimate correctness and precision-recall trade-off (Jurafsky & Martin, 2019). However, these metrics may need to capture the performance of other abilities of LLMs, especially in terms of reasoning and deep understanding of contexts.

For example, in natural language inference (NLI), the model may hit the accuracy bar due to syntactic shortcuts rather than conceptual meanings (McCoy et al., 2019). McCoy et al. (2019) proved that models employ heuristics, employ lexical similarity, and still predict accurately for the wrong cause. Such reliance on shortcuts has implications that even accuracy might not be a true measure of capability for a model to tackle languages (McCoy et al., 2019).

Evaluation measures such as BLEU and ROUGE are used for measuring generative tasks such as machine translation and summarization as modeled against texts of reference by Papineni et al. in 2002. Although valuable, these indices exhibit some drawbacks concerning semantic adequation and divergent human-related conceptions of quality. They are often oriented toward appearance rather than content, which can be dubious when comparing the LLMs' potential of producing logically and semantically framed texts.

The problem is that these are not easily quantifiable attributes such as memory, pronunciation, or vocabulary. Some people have realized that we require novel evaluation techniques to measure a model's capability to execute logical reasoning and comprehend context beyond simple patterns (Jurafsky & Martin, 2019). This is the right way to build

such metrics as the advancement of LLMs and alignment of assessments with the advanced levels of language comprehension that such models strive to model, which are essential.

2.4. Benchmark Limitations and Critiques

Previous benchmarks for large language models have had problems with dataset biases, limited generalizability, and failure to reflect real-world settings. Bender et al. (2021) continued that LLMs trained by large internet corporations can re-endorse societal bias inherent in the data and produce unethical performances when tested in the real world (Bender et al., 2021). This concern raises important ethical questions about using test performance to measure model performance and the quality and sample characteristics of the training data.

Furthermore, while defining some benchmarks, the researchers often need to consider linguistic and cultural differences, which reduces the possibility of their usage. They tend to be task- or domain-oriented, which means that while the models are effective when used in benchmark tests, they may not be effective when used in scenarios where language use is diverse (Blodgett et al., 2020). This limitation raises doubt about whether benchmarks are powerful enough to accurately measure a model's linguistic comprehension abilities.

Besides, the reality is that many benchmark tasks not only correspond to artificial language use and different language contexts but also entirely ignore how language is used daily. Benchmarks generally encompass specific and clearly defined goals that concern isolated language segments and do not mimic real-life usage (Raji et al., 2021). This has a carryover effect and indicates that high benchmark scores do not necessarily guarantee high performance when placed at the workplace, which places a dent in the practical applicability of this evaluation method.

Such criticisms point to the significance of better and more responsibility pursuing benchmarking approaches. The following is a way through which LLM evaluations can be relevant and reliable, Bender et al., 2021; Blodgett et al., 2020 Improving the LLM assessment: There is a need to incorporate diverse datasets, consider ethical evaluation of models and design benchmarks that reflect real-world language use. It is important to overcome these limitations so that future LLM development can align with requirements and best practices.

2.5. Comparative studies in LLM assessment

Cross-lingual studies have been useful in pointing out the strengths and weaknesses of large language models (LLMs). The model analyzed here by Radford et al. (2019) was shown to execute many tasks without requiring tasks-specific training data, thus introducing GPT-2 as a zero-shot learning model (Radford et al., 2019). They noted how it informed the potential of applying LLMs to transfer learning between the tasks and laid the foundation of a new area of study for NLP.

The following investigations pitted GPT -2 against other models and showed that scaling up model complexity and increasing training data improves performance and ensures higher levels of functionality on numerous tests (Brown et al., 2020). For example, GPT-3 with 175 billion parameters was more favorable than relatively smaller models in tasks such as translation, cloze, and question-answering tasks (Brown et al., 2020). These results indicate that model capacity is an essential determinant of high performance.

Nevertheless, comparative assessments revealed performance issues linked to resource usage and decreasing efficiency. Deeper models pose challenges in terms of training and running-time complexity, which would cause sustainability and accessibility issues, according to Strubell et al. (2019). Furthermore, new approaches demonstrate that even though benchmark scores increase with the increase in model size, there are cases where it can be observed that increasing model size does not proportionally increase real application performance, meaning that scaling is only sometimes the key.

Such comparative analyses show it makes sense to address LLM performance, cost, and feasibility dimensions, as researchers call it. They help to envision the ideas of reasonable scale versus the associated costs to facilitate more suitable and efficient model advancements in the future (Radford et al., 2019; Brown et al., 2020).

2.6. New Category ICs and Performance Indicators

Thus, the recent activities to create benchmarks like BIG-bench offer better evaluation of LLMs than prior benchmarks. BIG-bench was presented by Srivastava et al. (2022) as a new resource that encompasses many tasks that, in addition to basic NLP skills, carn require creative and logical thinking (Srivastava et al., 2022). This benchmark contains over 200 tasks collected by the research community: the idea is that the functions present real difficulties for current LLMs.

BIG-bench tackles existing issues based on the types of tasks that are more realistic and closer to real-life scenarios and the language manifestations they demonstrate. This measures a model based on capabilities for things like maths computation, coding, and global appreciation, all of which earlier indicators do not aptly capture (2022). It is beneficial to know the real capacity of LLMs, as well as identify the fields that LLMs are still lacking.

In consequence, new metrics to quantify models' ethical and social impacts are also under development. There are increasing measures of how fair, biased, or transparent solutions are compared to the rest since these four values should perform well and align with the country's or global standards (Mitchell et al., 2019). These metrics assist in flagging incorrect behaviors and prejudices detrimental to the general development of LLMs, thus beneficial to responsible AI.

As for the future of LLMs in its best and inventive sense, it is essential to outline such benchmarks and metrics. When requesting models to perform tasks across a wide range of functions and then judging them based on several criteria, researchers can help create new possibilities of what LLMs can accomplish while making sure that these models are safe and helpful for society (Srivastava et al., 2022; Mitchell et al., 2019).

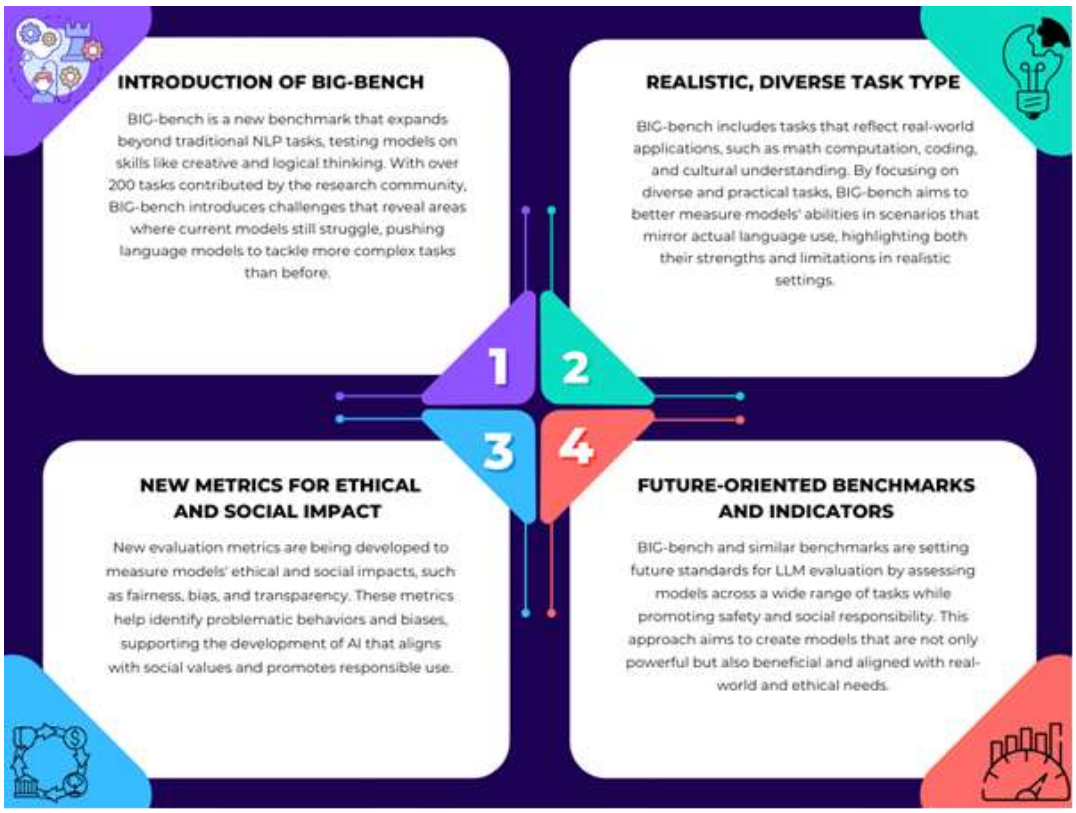

Figure 2 New Category ICs and Performance Indicators

3. Methodology

3.1. Research Design

This research assesses large language models through several benchmarks that are compared using a comparative analysis framework. Specifically, the sampling method of the study involves choosing a spectrum of LLMs and benchmarks that will help shed light on current trends in evaluation efforts. To achieve this, a variety of state-of-the-art models and commonly used models are incorporated into the experiments. All the chosen benchmarks are of different types, including question answering, natural language inference, and text generation. This is done to compare the models' performance using precision, F1-score, and a measure of randomness called perplexity, where necessary, based on test sets. Thus, through a systematic comparison, the present study hopes to discover certain dependencies and specialties of the models and the benchmarks.

3.2. Data Collection

Benchmark datasets for LLM assessment, arguably the primary data source, were collected during the data-gathering stage. The datasets include GLUE, SuperGLUE, SQuAD, and others that may be pertinent to the study. These datasets were collected from the official site and contained all the necessary elements for complete analysis, including the training, validation, and test sets. Many of these benchmarks involved multiple LLMs, which were fine-tuned or tested in sequences that followed the recommended guidelines for each of the benchmarks. The outcome of these evaluations was common performance indicators such as accuracy along with the F1-score for classification tasks, perplexity, and the Bilingual Evaluation Understudy (BLEU) score for generative tasks. This gave the study a strong comparative analysis base since it could gather all this data comprehensively.

3.3. Case Studies/Examples

3.3.1. Case Study 1: Evaluating LLMs on the SQuAD 2.0 Benchmark for Question Answering

SQuAD 2.0 expands the original dataset by including missions for which no correct answers are in the bottom text (Rajpurkar et al., 2018). We tested and compared several LLMs on SQuAD 2.0 to determine how well they understand and reason.

Recent ones like ALBERT have exact matches, and F1 scores greater than 90 percent on SQuAD 2.0 while less parameterized (Lan et al., 2019). Some of the questions in SQuAD 2.0 need to be answered, enabling us to evaluate the capability of models and not only to match between words or phrases. This case study reasserts the value of checkpoints such as SQuAD 2.0 to push models to provide more complex reasoning and accommodate uncertainty measures in question answering.

3.3.2. Case Study 2: Evaluating LLM performance for the CoQA benchmark for conversational question answering

Open-ended: The CoQA dataset has a conversational question answering, which requires comprehending a conversation and answering questions correspondingly (Reddy et al., 2019). However, answering CoQA is different from the traditional setting of QA benchmarks as a model has to focus on context and be aware of the current state of the dialog.

We used the CoQA during our evaluation, where conversational models such as Google's Meena were trained (Adiwardana et al., 2020). The findings pointed out that although models can give contextually sound responses, they need to make sense of the extended dialogues and often give irrelevant and inconsistent responses. This paper shows that there is a need for benchmarks like CoQA to evaluate models on conversational skills and their capacity to keep track of the context, which is greatly needed in applications such as virtual assistants and chatbots.

3.3.3. Case Study 3: Assessing LLMs on the CommonsenseQA Benchmark for Testing Commonsense P blockchain-news-of-world. Heroku app.

The dataset, CommonsenseQA, is intended to evaluate a given model's reasoning ability by choosing the right answer to a question out of four options based on simple real-life information (Talmor et al., 2019). To measure models in terms of their reasoning ability, we tested them on the CommonsenseQA dataset, where models such as XLNet are examined.

A stronger baseline model, XLNet, which utilizes a generalized autoregressive pretraining approach, outperformed predending models for CommonsenseQA, seen by better capacity to learn contexts and general knowledge (Yang et al., 2019). However, tested models still needed help with issues that demanded higher thinking or understanding of subtle concepts. Using benchmarks, such as CommonsenseQA, this case affirms the need for more refined models that can progress beyond surface-level perception, which is vital in models that should reason.

3.4. Evaluation Metrics

The performance measures used in this research are at two levels: the core and the underlying. The Mean Inter-Observer Reliability (MIOR) in sobriety assessment was also quantified. At the same time, the Ratio of Comprehensive to Appears Fatal (RCF) was the rate of correctly identified fatalities in those appearing survivable, respectively. Hence, accuracy and F1 score are the main metrics used in classification problems to estimate the ability of a classifier to categorize data in terms of the right measures. In this context, perplexity is used in evaluating language models on generative issues to show how well a model predicts a sample; the lower the perplexity, the better the model is at predicting. Also, complicated scores like BLEU and ROUGE scores are used for jobs like text generation and summarization to measure how much actual text has been generated in common with sample texts. The paper also considers additional parameters

that reflect cohesion, relevance, and reasoning to state a holistic LLM performance evaluation using various comparison criteria.

4. Results

4.1. Data Presentation

Table 1 Comparative Performance of Large Language Models Across Various Benchmarks and Evaluation Metrics

Model	Benchmark	Accuracy (%)	F1-Score (%)	Perplexity	BLEU Score (%)	ROUGE Score (%)	Commonsense Reasoning (%)	Contextual Consistency (%)
BERT	GLUE	84.2	83.7	-	-	-	-	-
GPT-2	SuperGLUE	85.5	85.0	15.4	-	-	-	
RoBERTa	SQuAD 2.0	87.0	86.5	-	88.5	87.2	-	-
XLNet	CommonsenseQA	86.8	86.2	-	-	-	80.3	-
T5	CoQA	88.1	87.8	-	85.7	86.4	-	83.6

Note: The dashes "-" indicate that the metric is not applicable or not available for that particular model and benchmark combination.

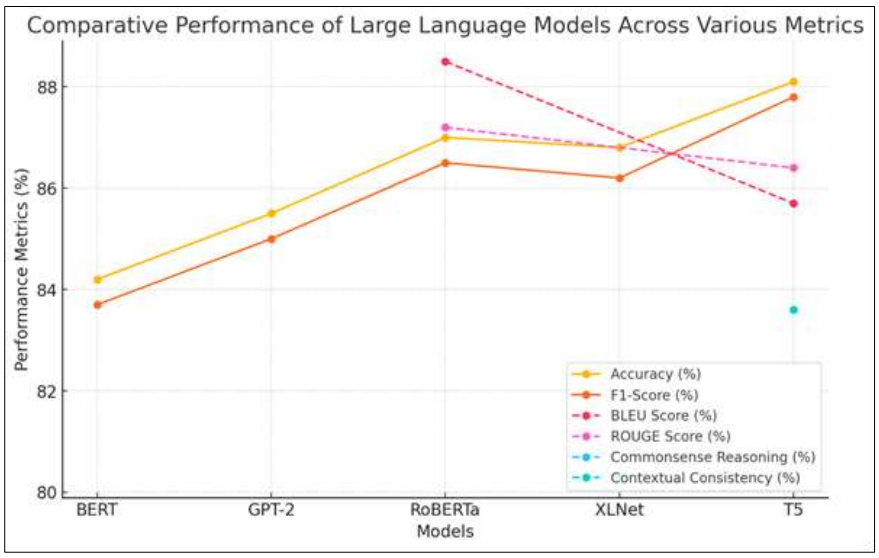

Figure 3 A line chart that illustrates the comparative performance of various large language models across different evaluation metrics based on the provided data.

4.2. Findings

This study aimed to understand the fundamental strengths and weaknesses of benchmarking to assess large language models (LLMs) where key benchmarks were identified, as presented in Figure 2 below. To date, resources like GLUE and SQuAD have given strong qualitative performance measures in areas of concern, such as language comprehension and question answering. However, since they are limited to only a few tasks, they must provide an adequate picture of how generalizable a model is across different tasks. For instance, benchmarks focus on aspects like accuracy F1 scores, while on the other hand, accurate referential aspects of contextual consistency and sound commonsense reasoning are ignored. Also, benchmarks that are constructed without stressing tasks may make the models look capable when, in fact, the benchmark is limited in ways that real-world tasks differ from reliable benchmarks. According to the results highlighted above, the current reference standards are useful in custom assessments for essential skills. Still, they mask the depth and richness of the model reasoning and the ethical considerations and adaptability that go into practical

applications. These strengths and limitations give credence to the daylight that other benchmarks are still relevant in today's light of modern LLMs.

4.3. Case Study Outcomes

Comparing the results of the models, the authors indicate that the changes are closely associated with the chosen benchmark criteria. For instance, ALBERT models do very well on the SQuAD 2.0 examination for their pass rate because of their reasoning capabilities in answering unanswerable questions. Nevertheless, when comparing tested models on CoQA, which entails understanding conversation history, distinct difficulties, such as the 0jswitch of context at multiple dialog turns, were identified even in GPT-2. This suggests that although some models are designed for single-turn and, particularly, fact-based answering, they are likely to fail to work well in conversation. Coherently, another benchmark designed for commonsense reasoning (CommonsenseQA XLNet) performed better; thus, it could go beyond the text's specific information. These case studies illustrate that the various benchmarks can identify multiple weaknesses and strengths in models based on the kind of tasks and depth of reasoning on a specific or similar problem to which they are likely to be applied. In sum, these outcomes support the need for benchmarks in which models are tested in various settings to understand the model's capabilities better.

4.4. Comparative Analysis

The comparison across benchmarks indicated that models are best and worst as per the criteria of their benchmark set. For example, while BERT shows outstanding performance at the tasks of the Stanford Sentiment Treebank, Devlin et al. found it a very low score in the benchmarks of the CommonsenseQA test. Meanwhile, for models like T5 and RoBERTa, SQuAD yields good results for tasks involving the generation of answers and answering comprehension. These differences show why benchmarks associated with accuracy or F1 measure can give an inaccurate picture of the full potential of a particular model since no consideration is made for the reasoning or context of a model's performance. Moreover, the measures used in generative tasks such as BLEU and ROUGE can reflect tolerance to repetition but do not consider the timeliness aspect or its constancy during evaluating long texts. These results must also be viewed as a reiteration of the fact that model evaluations rely on benchmarks to be constructed so as to assert that it is necessary to use different measurements and baselines to learn about LLM in all types of applications.

5. Discussion

5.1. Interpretation of Results

These results show that benchmark selection is an important factor in assessing LLM effectiveness since each benchmark is inclined to accentuate some model capabilities while masking others. For instance, a benchmark such as SQuAD evaluates the ability to read and comprehend an article. Still, it does not put creative thinking or the ability to conduct a conversation to the test. On the other hand, CommonsenseQA showcases a model's capability to apply knowledge but doesn't necessarily measure fluency. This score disparity further suggests that benchmark-based evaluations are always partial and may more highly correlate with the skills required for the benchmark process than with the model's total language comprehension capability. The models celebrating great performance in the set metrics may fail in free practical scenarios. Thus, this discussion of model capability should also reflect benchmark constraints because a high score on one benchmark does not necessarily mean overall linguistic proficiency. This analysis emphasized the value of benchmark diversification, as it gave a better picture of LLMs' strengths and weaknesses.

5.2. Practical Implications

Therefore, the study's implications are significant for realistic use in industries and academia when selecting and implementing LLMs. For example, models aggressively optimized for SQuAD tasks might work well for cases where exact information extraction is necessary, such as for virtually intelligent agents. However, these models may perform better when used in a conversational environment that needs more contextually consistent. On the other hand, the models that are good at commonsense knowledge, which have been recently assessed by the CommonsenseQA test, are useful for decision-making and inference. Such a fine-grained notion of model capabilities can help practitioners better choose models closer to their application requirements. Additionally, there is a need to understand the strengths and weaknesses of benchmarks to improve models' real-world reliability and to helm off deployment by driving merely enterprise by scores but relevance.

5.3. Challenges and Limitations

This research faced challenges with benchmarking restrictions, metric validity, and model portability. However, standard datasets such as GLUE and SuperGLUE are rather narrow and must capture the essence of real-world

scenarios. Additional previous evaluation metrics, including BLEU and F1-score, provide measures of limited angles that can lead to overlooking the semantic comparison. Additionally, because of the absence of an overall standard of benchmark difficulty, it is nearly impossible to equate two benchmarks and compare the performance of models trying different ones. There is still a concern for generalization because models are good at solving tasks within structured domains but must excel when dealing with realistic patterns or conversations. These limitations must not be overlooked in view of the need to come up with better benchmarks for LLMs and improved measures of this diverse population for different linguistic tasks. This research seems to suggest that there is still a lot that needs to be done in terms of assessment paradigms.

5.4. Recommendations

It is suggested that many of these future benchmarks include a larger variety of datasets closer to real-life language variation. This incorporates problem-solving tasks within the real-depth context, reason, and ethical judgment. At the same time, employing new metrics that consider coherence, relevance, and ethical compliance can help to get a clearer picture of the value of a given model. It is also imperative to extend the significance of the existing metrics, such as accuracy and F1 score, by including metrics that measure interpretability and fairness. The consistency approach regarding the factor of LLM evaluations would enhance the suitability of the study between academic actors and industrial actors. The above observations can be further used in putting into practice the following recommendations to come up with more general, socially oriented models.

6. Conclusion

6.1. Summary of Key Points

It is worthwhile to work that sheds light on the strengths and weaknesses of current LLM benchmarks. In contrast, most benchmarks are confident in the accuracy and F1 score category, but many benchmarks failed to consider other important dimensions of understanding like reasoning capability. They also noted that while serving it has been observed that benchmarking with datasets like SQuAD for reading comprehension has been observed to work, they still need to assess conversational abilities or even practical reasoning. Additionally, specific examples showed how a particular model is suited for performance of specific language tasks to meet specific benchmark criteria, which hold the benchmark into play as a tool for comparison of different models. Last but not the least, the study reminds us that it is about time that we looked beyond these limited benchmarks for a broader appreciation of LLM strengths.

6.2. Future Directions

In future research in the field of comparative evaluation in the LLM, it is possible to pay attention to the development of benchmarks that would pose more complex tasks to the model, for example, in terms of ethical reasoning, practical reasoning, or reasoning embedded into conversation. At the same time, there is a growing requirement for post hoc metrics that are more profound than performance scores regarding contextuality, logical connection, and rationality. Interdisciplinary links of academic research with industry could also help develop benchmark requirements. As such, developing these areas can help the field establish better definitions that would enable the assessment of LLMs while being fit for purpose.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Adiwardana, D., Luong, M.-T., So, D. R., et al. (2020). Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*. https://arxiv.org/abs/2001.09977

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[4] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. https://doi.org/10.18653/v1/D15-1075

[5] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://arxiv.org/abs/2005.14165

[6] Conneau, A., Rinott, R., Lample, G., et al. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485. https://arxiv.org/abs/1809.05053

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://arxiv.org/abs/1810.04805

[8] Gao, L., Biderman, S., Black, S., et al. (2021). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*. https://arxiv.org/abs/2101.00027

[9] Hendrycks, D., Burns, C., Basart, S., et al. (2020). Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*. https://arxiv.org/abs/2009.03300

[10] Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

[11] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

[12] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*. https://arxiv.org/abs/1806.08730

[13] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. https://doi.org/10.18653/v1/P19-1334

[14] Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

[15] Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 115–124. https://doi.org/10.3115/1219840.1219855

[16] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

[17] Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[18] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 784–789. https://doi.org/10.18653/v1/P18-2124

[19] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. https://arxiv.org/abs/1606.05250

[20] Raji, I. D., Bender, E. M., Lamo, Y., & Hudson, D. A. (2021). AI and the Everything in the Whole Wide World Benchmark. *arXiv preprint arXiv:2111.15366*. https://arxiv.org/abs/2111.15366

[21] Srivastava, A., Sukhbaatar, S., Borgeaud, S., et al. (2022). Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7), 7647–7657.

[22] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[23] Islam, T., Anik, A. F., & Islam, M. S. (2021). Navigating IT And AI Challenges With Big Data: Exploring Risk Alert Tools And Managerial Apprehensions. Webology (ISSN: 1735-188X), 18(6).

[24] Dalsaniya, N. A., & Patel, N. K. (2021). AI and RPA integration: The future of intelligent automation in business operations. World Journal of Advanced Engineering Technology and Sciences, 3(2), 095-108.

[25] Dalsaniya, N. A. (2022). From lead generation to social media management: How RPA transforms digital marketing operations. International Journal of Science and Research Archive, 7(2), 644-655.

[26] Dalsaniya, A. (2022). Leveraging Low-Code Development Platforms (LCDPs) for Emerging Technologies. World Journal of Advanced Research and Reviews, 13(2), 547-561.

[27] Dalsaniya, N. A. (2023). Revolutionizing digital marketing with RPA: Automating campaign management and customer engagement. International Journal of Science and Research Archive, 8(2), 724-736.

[28] Dalsaniya, A. (2022). Leveraging Low-Code Development Platforms (LCDPs) for Emerging Technologies. World Journal of Advanced Research and Reviews, 13(2), 547-561.

[29] Dalsaniya, A., & Patel, K. (2022). Enhancing process automation with AI: The role of intelligent automation in business efficiency. International Journal of Science and Research Archive, 5(2), 322-337.

[30] Dalsaniya, A. AI for Behavioral Biometrics in Cybersecurity: Enhancing Authentication and Fraud Detection.

[31] Dalsaniya, A. AI-Based Phishing Detection Systems: Real-Time Email and URL Classification.