Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 15-30 recibido 18-11-2023 revisado 10-01-2024 aceptado 30-01-2024

# Revisiting Challenges and Hazards in Large Language Model Evaluation

## *Análisis de los Desafíos y Riesgos en la Evaluación de Grandes Modelos del Lenguaje*

**Inigo Lopez-Gazpio**
HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
inigo.lopez@ehu.eus

**Abstract:** In the age of large language models, artificial intelligence's goal has evolved to assist humans in unprecedented ways. As LLMs integrate into society, the need for comprehensive evaluations increases. These systems' real-world acceptance depends on their knowledge, reasoning, and argumentation abilities. However, inconsistent standards across domains complicate evaluations, making it hard to compare models and understand their pros and cons. Our study focuses on illuminating the evaluation processes for these models. We examine recent research, tracking current trends to ensure evaluation methods match the field's rapid progress requirements. We analyze key evaluation dimensions, aiming to deeply understand factors affecting models performance. A key aspect of our work is identifying and compiling major performance challenges and hazards in evaluation, an area not extensively explored yet. This approach is necessary for recognizing the potential and limitations of these AI systems in various domains of the evaluation.
**Keywords:** Large language models, evaluation, evaluation challenges and hazards, evaluation dimensions.

**Resumen:** En la era de los modelos de lenguaje de gran escala, el objetivo de la inteligencia artificial ha evolucionado para asistir a personas de maneras sin precedentes conocidos. A medida que los modelos se integran en la sociedad, aumenta la necesidad de evaluaciones exhaustivas. La aceptación de estos sistemas en el mundo real depende de sus habilidades de conocimiento, razonamiento y argumentación. Sin embargo, estándares inconsistentes entre dominios complican la evaluación, dificultando la comparación de modelos y la comprensión de su funcionamiento. Nuestro estudio se enfoca en organizar y aclarar los procesos de evaluación de estos modelos. Examinamos investigaciones recientes para analizar las tendencias actuales e investigar si los métodos de evaluación se ajustan a los requisitos del progreso. Finalmente, identificamos y detallamos los principales desafíos y riesgos que afectan la evaluación, un área que aún no ha sido explorada extensamente. Este enfoque es necesario para reconocer las limitaciones actuales, el potencial y las particularidades de la evaluación de estos sistemas.
**Palabras clave:** Modelos de lenguaje de gran escala, evaluación, desafíos y riesgos de evaluación, dimensiones de la evaluación.

## 1 Introduction

Since the early days of expert systems, it has been recognized that for these systems to be accepted in real-world domains, they must not only demonstrate their knowledge (Khalfa, 1994) but also be able to reason and argue about it (Buchanan and Shortliffe, 1984; Lacave and Díez, 2002; Korb and Nicholson, 2010). In the era of large language models (LLM), the aim of artificial intelligence has shifted from merely imitating natural intelligence to supporting humans in novel unprecedented ways (Deng and Lin, 2022). The acceptance of AI by users hinges on the quality of the evaluations performed. The advent of pre-trained language models has marked a significant advancement. These models, developed by training Transformer models (Vaswani et al., 2017) on extensive corpora, have exhibited exceptional capabilities in various

ISSN 1135-5948 DOI 10.26342/2024-72-1

natural language processing (NLP) tasks, sometimes presumably surpassing human performance (Orrù et al., 2023; Hadi et al., 2023a; Zhao et al., 2023; Chang et al., 2023). The recent surge in LLM performance evaluation reflects the complexity and necessity of tailored evaluation approaches.

Recently, the evaluation of LLMs has continuously evolved, placing greater focus on evaluation beyond fixed knowledge traditional datasets and focusing on innovative aspects, such as: comprehensive assessments (Xu et al., 2023a), ethical considerations (Head et al., 2023), and sustainability (Khowaja, Khuwaja, and Dev, 2023). These aspects are now considered alongside the traditional evaluations of knowledge and generalization capabilities. As LLMs increasingly become a part of our societal frameworks, the need for multidimensional and thorough evaluations becomes more pronounced. This diverse range of evaluation approaches not only improves the quality of LLMs but also ensures their responsible and advantageous application in real-world scenarios. Consequently, the process of evaluating LLMs has become a crucial component closely tied to the development and refinement of these models.

The evaluation criteria for LLMs, including BERT (Aftan and Shah, 2023), GPT-3 (Floridi and Chiriatti, 2020), InstructGPT (Ouyang et al., 2022), PaLM (Chowdhery et al., 2022), GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023) and their successors (Lehman et al., 2023), is turning into a complex and multifaceted process crucial for understanding their capabilities, limitations, and impacts. Traditional metrics like perplexity (Gamallo, Campos, and Alegria, 2017) and BLEU score (Reiter, 2018), focusing on linguistic accuracy and fluency, are no longer sufficient (Tang, Chuang, and Hu, 2023). As LLMs become more advanced, their evaluation also needs to evolve, encompassing a broader range of criteria to ensure their robustness, effectiveness, fairness, interpretability, environmental impact and safety in task-specific settings. Recent evaluations have concentrated on several key aspects:

**1. Robustness and Generalization:** testing across diverse topics and contexts to ensure consistent performance even in unfamiliar scenarios. Generalization tests are essential as they assess a model's ability to effectively apply its acquired knowledge to new and unfamiliar domains (Dong et al., 2023).

**2. Fairness and Bias Testing:** LLMs can inadvertently perpetuate societal biases present in their training data. Rigorous testing is required to identify and mitigate biases to prevent discrimination over race, gender, or other sensitive attributes (Li et al., 2023; Huang et al., 2023).

**3. Interpretability and Explainability:** understanding the decision-making process of LLMs is vital. Interpretability tools and methods are being developed to provide insights into model's knowledge. Transparency is crucial for trust and reliability in sensitive applications (Saha et al., 2023).

**4. Environmental Impact:** computational demands of training and operating large models have brought attention to their environmental effects. Assessing these models for energy efficiency and carbon footprint is now crucial, guiding the field towards sustainable practices (Rillig et al., 2023).

**5. Task-Specific Evaluations:** task-specific evaluations are vital beyond just general metrics. For instance, in a translation task, fluency and cultural appropriateness are key, while in a medical diagnosis application, accuracy and reliability are paramount (Chang et al., 2023).

**6. Human-Centric Evaluations:** including human judgment in evaluation processes is becoming more popular. Human evaluators offer detailed feedback on elements such as usefulness, coherence, empathy, and the suitability of responses, areas where automated metrics may fall short (Ouyang et al., 2022; Zhong et al., 2023).

**7. Adversarial Testing:** Exposing LLMs to adversarial examples, where inputs are deliberately modified to test the model's resilience, is another emerging evaluation strategy. This helps in understanding the limits of a model's understanding and reasoning capabilities (Xu et al., 2023b).

Recent advancements in LLM evaluations have led to diverse, non-standardized approaches. A comprehensive evaluation approach is crucial for developing robust, fair, and efficient models, but it introduces challenges such as complexity, consistency, resource demands, and adaptability. The wide array of evaluation areas requires unique methodologies, tools, and expertise, making the process complex and resource-intensive. With varying standards and benchmarks across domains, consistency in evaluations is difficult, complicating model comparisons and full understanding of their strengths and weaknesses. Furthermore, finding a balance among different evaluation criteria is challenging. Enhancing a model's performance in one task could compromise its effectiveness in another area. Under these circumstances, there's also a risk of overemphasizing certain domains, like reading comprehension, machine translation or generability, at the expense of others, such as truthfulness, fairness or interpretability. This imbalance can lead to models excelling in certain tasks but falling short in vital areas. Additionally, domains involving human-centric criteria, such as ethics or user satisfaction, bring subjectivity into evaluations, causing inconsistent outcomes and interpretations. For newcomers or smaller institutions, the broad spectrum of evaluation areas poses a challenge. The need for extensive resources and expertise to perform thorough evaluations may limit innovation and diversity in the research community

In the current landscape, where even the challenges of LLM evaluation are not clearly defined, this study aims to clarify the well-known evaluation domains of LLMs. By reviewing the latest in LLM research, we aim to highlight recent trends and keep evaluation methods aligned with the rapid developments in this field. An ongoing challenge is to ensure these evaluation domains and methodologies remain updated. Additionally, this research attempts to link the main hazards associated with key LLM evaluation dimensions, an area that has not yet been thoroughly explored. Understanding these hazards is crucial for creating more effective evaluation scenarios for LLMs. However, tackling these hazards demands a multi-disciplinary approach that goes beyond technical solutions, incorporating considerations of ethics, user experience, and societal impact. As LLMs continue to advance, methods for evaluating and addressing these hazards must also evolve.

This study is organized as follows: Section 1, "Introduction", sets the stage and context for our work. Section 2, "Review on LLM evaluation", reviews the dimensions of LLM evaluation based on current research and analyzes the performance of state-of-the-art LLMs. Section 3, "Discussion on LLM evaluation", delves into a detailed discussion on LLM evaluation, highlighting the primary hazards associated with these evaluation dimensions. Section 4, "Description of main hazards", specifically focuses on identifying and detailing the main hazards in LLM evaluation. Finally, Section 5, "Conclusions", summarizes our findings, outlines future research directions, and discusses the limitations.

## 2 *Review on LLM evaluation*

LLMs are increasingly popular in both academic and industrial settings due to their remarkable performance across various applications (Devlin et al., 2018; Gao and Lin, 2004; Kasneci et al., 2023; Zhao et al., 2023). As LLMs become more integral to research and everyday use, understanding their potential risks at both task and societal levels is essential. Recent years have seen considerable efforts in evaluating and assessing LLMs from multiple angles. Typically, LLMs are defined as language models with hundreds of billions of parameters, trained on vast text datasets (Shanahan, 2022). Most LLMs share similar model architectures, based in the Transformer architecture, and pre-training objectives, such as language modeling, with size variable training parameters. The key distinction of LLMs lies in their significantly larger scale in terms of model size, data used for training, and computational power. This scaling enables them to better comprehend natural language and generate high-quality text based on given contexts or prompts. The improvement in capability with model size is partially explained by the scaling law, where performance increases substantially with model size (Kaplan et al., 2020). However, certain abilities, as noted in (Zhao et al., 2023), only become apparent when the model size reaches a specific threshold, deviating from what the scaling law predicts.

LLMs have recently received substantial interest in both academic and industrial sectors (Bommasani et al., 2021; Wei et al., 2022; Zhao et al., 2023). As indicated by recent research (Bubeck et al., 2023), the impressive performance of LLMs has sparked optimism about their potential as a form of Artificial General Intelligence (AGI). Unlike previous models limited to specific tasks, LLMs are adept at a wide range of tasks, from general language tasks to domain-specific applications. This versatility makes them increasingly popular among users with critical information needs.

Furthermore, these billion-parameter models, despite being resource-intensive, are surprisingly user-friendly. They don't demand access to specialized hardware or software, nor a deep understanding of machine learning or natural language processing. Instead, LLMs are accessible through APIs and are capable –or at least claimed to be– of handling complex tasks with minimal (few-shot) or no (zero-shot) prior information. This accessibility offers a more intuitive and natural way of interacting with computers (de Wynter et al., 2023). The complexity inherent in the linguistic interactions of a LLM makes it challenging to establish a concise, standardized method for assessing its quality or gaining a deeper understanding of how to evaluate its composed representations. Consequently, a diverse array of evaluation methods for LLMs is emerging to address these multiple challenges.

This section provides a detailed review of the principal methods used to evaluate LLMs in the state-of-the-art, highlighting several critical dimensions. In line with recent trends, these evaluations focus on various aspects: (i) robustness and generalization reliability of the models, (ii) fairness and the presence of bias in model outputs, (iii) interpretability and explainability of the models, (iv) environmental impact of the models, (v) task-specific evaluation such as translation or summarization, (vi) human-centric evaluation including user trust and confidence, and (vii) resilience against adversarial testing.

Current consensus in the field of LLM evaluation suggests that it should be structured around three key dimensions, each encompassing distinct aspects and challenges: (dimension 1) the scope of the evaluation, (dimension 2) the extent of the evaluation, and (dimension 3) the procedure of the evaluation.

The studies conducted by (Orrù et al., 2023; Hadi et al., 2023a; Chang et al., 2023; Zhao et al., 2023) are among the first comprehensive surveys in this area. They concur on the importance of these three dimensions for LLM evaluation. The first dimension covers the range of evaluation tasks applicable to LLMs. The second dimension focuses on selecting suitable scenarios for the evaluation (i.e. benchmarks). The third dimension deals with the actual evaluation process, employing the chosen tasks, datasets or benchmarks. These dimensions collectively form the cornerstone of effective evaluation. We will now describe each of these dimensions in detail.

## 2.1 Fixed-knowledge evaluation

Evaluating fixed-knowledge in LLMs is complex, with no universal solution fitting all scenarios. The primary aim of such evaluations is to compare different systems that generate varied representations for a specific task. Although the ultimate objective is to apply these models in high-level tasks or market applications, evaluating them on manually annotated, more detailed tasks often provides deeper insights and facilitates error analysis in controlled environments. In fact, focusing on intermediate tasks has been instrumental in advancing fixed-knowledge evaluation, thanks to widely-used datasets in key natural language processing (NLP) categories.

Main categories in NLP encompass Natural Language Understanding (Bates, 1995) and Natural Language Generation (McDonald, 2010) tasks. Examples include text classification (Song et al., 2014), reading comprehension (Baradaran, Ghiasi, and Amirkhani, 2022), machine translation (Baltrušaitis, Ahuja, and Morency, 2018), language modeling (Min et al., 2023), grammar analysis (Wang et al., 2020), code generation (Shin and Nam, 2021), question answering (Bouziane et al., 2015), dialogue (Motger, Franch, and Marco, 2022), logic reasoning (Costantini, 2002), language inference (Storks, Gao, and Chai, 2019), truthfulness (Oshikawa, Qian, and Wang, 2018), fact checking (Lazarski, Al-Khassaweneh, and Howard, 2021), toxicity detection (Garg et al., 2023), bias detection (Garg et al., 2023), multimodality (Erdem et al., 2022), summarization (Awasthi et al., 2021), negation (Mahany et al., 2022), sentiment analysis (Zhang, Wang, and Liu, 2018), semantic understanding (Salloum, Khan, and Shaalan, 2020), and more.

## 2.2 Evaluation of versatility

Evaluating how well foundational models handle tasks at a human level is crucial in their development towards AGI. Traditional fixed-knowledge datasets, often based on single tasks might not fully capture human-like abilities, as the latter ones potentially combine multiple objectives.Thus, the approach of fixed-knowledge evaluation for LLMs is becoming recognized as inadequate for a thorough assessment. This method, which uses a static set of datasets, falls short due to the dynamic and complex nature of language and knowledge, as well as the continuous evolution of

LLMs. Such evaluations don't always reflect the real-world versatility and adaptability required of these advanced systems.

The shortcomings of fixed-knowledge evaluation have prompted the creation of large-scale, dynamic benchmarks. These benchmarks are tailored to encompass a wider range of language understanding and generation tasks, striving to be more inclusive and reflective of real-world language usage. They typically involve diverse and complex tasks, extensive enough to capture broad linguistic trends. Additionally, these benchmarks often incorporate considerations of fairness, bias detection, and ethics, acknowledging the increasing importance of social responsibility in LLMs. By assessing models against these expanded criteria, we can better ensure their linguistic proficiency as well as their ethical and social integrity (Zhong et al., 2023).

Recently, a variety of benchmarks have been developed to evaluate LLMs across a range of tasks. We now enumerate and briefly describe some of the most notables:

GLUE (Wang et al., 2018) (General Language Understanding Evaluation) and SUPERGLUE (Wang et al., 2018) consist of sota benchmarks designed to mimic real-world language processing scenarios. They encompass a variety of tasks such as text classification, machine translation, reading comprehension, and dialogue generation, offering a comprehensive assessment of capabilities.

PromptBench (Zhu et al., 2023) highlights the sensitivity of current LLMs to adversarial prompts, underscoring the need for meticulous prompt engineering.

WinoGrande (Sakaguchi et al., 2021) is a benchmark designed to test AI systems' common sense reasoning and natural language understanding. It features a series of nearly identical sentence pairs, each with a subtle variation that alters the meaning of a crucial word. The test for AI systems is to accurately interpret these sentences and resolve the ambiguities.

AGIEVAL (Zhong et al., 2023) stands out as a human-centric benchmark based on standardized exams. It encompasses a diverse array of tests, including college entrance exams, law school admission tests, math competitions, and lawyer qualification exams. This benchmark is designed to evaluate AI systems in contexts that require a high level of academic and professional understanding.

Another significant benchmark is MMLU (Hendrycks et al., 2020) (Massive Multitask Language Understanding). MMLU offers a comprehensive evaluation framework to test AI models' language understanding across various subjects and disciplines. It includes tasks from humanities and social sciences to STEM fields, aiming to gauge the models' depth and breadth of knowledge. MMLU is distinctive for its focus on complex comprehension and reasoning, challenging language models to demonstrate their understanding and processing abilities across diverse areas of expertise.

BigBench (Ghazal et al., 2013) is recognized as an industry-standard benchmark for big data analytics. BIG-bench benchmark serves as a thorough and varied tool for evaluating LLMs. It covers a broad spectrum of tasks, testing different aspects of NLU and NLG, and extends beyond the scope of traditional benchmarks. BIG-bench is specifically designed to challenge LLMs in areas like advanced reasoning, creativity, and comprehension of complex and subtle language nuances.

HELM (Liang et al., 2022) offers a comprehensive evaluation framework for LLMs. It assesses language models on multiple fronts, including NLU, NLG, coherence, context sensitivity, common-sense reasoning, and domain-specific knowledge. The goal of HELM is to provide a holistic evaluation of language models, gauging their performance across a variety of tasks and domains.

HellaSwag (Zellers et al., 2019) is a benchmark specifically designed to assess common sense reasoning and contextual understanding in LLMs. It provides context-rich scenarios, each accompanied by multiple-choice endings, and the model's task is to select the most plausible conclusion for each scenario. The scenarios in HellaSwag are intentionally diverse and challenging, often demanding a nuanced comprehension of everyday activities and situations. This benchmark aims to advance AI capabilities in complex, real-world common sense reasoning.

The HumanEval benchmark (Chen et al., 2021) is designed to test the code generation abilities of LLMs. It presents a series of programming challenges, each consisting of a function signature, a body with a TODO comment, and several unit tests. The model's task is to complete the function body so that it successfully passes all the tests. HumanEval specifically focuses on models' capacity for understanding and generating functional programming code. It evaluates the algorithmic thinking, problem-solving, and coding skills, making it an important tool for gauging software development skills.

The GSM benchmark (Cobbe et al., 2021) is tailored to test LLMs' mathematical reasoning skills. It comprises grade-school level math problems that span a range of mathematical skills, from basic arithmetic to advanced problem-solving. This benchmark challenges AI models to comprehend and manipulate numerical information, execute calculations, and utilize mathematical concepts to solve problems. GSM is particularly valuable for evaluating capabilities in logical reasoning and numerical understanding.

## 2.3 Methodology of the evaluation

The third dimension of evaluation revolves around the evaluation methodology and particularly whether human judgment is incorporated into the process. Incorporating human feedback into the evaluation of LLMs is becoming increasingly essential, complementing automated scoring metrics like BLEU or perplexity. While automated metrics offer valuable quantitative data, they often miss the nuanced, qualitative elements of language crucial for a comprehensive understanding and enhancement of knowledge-based systems (Qin et al., 2023; Bang et al., 2023).

Automated metrics are typically designed to assess specific linguistic aspects, such as grammatical accuracy or lexical similarity to a reference text. However, effective language use involves more than just grammatical correctness. It encompasses context, cultural nuances, pragmatics, and the conveyance of subtle meanings, which automated metrics may not fully grasp. Human evaluators bring a crucial perspective to these qualitative elements, providing a more complete evaluation of performance. Furthermore, human evaluation is key in determining the relevance and coherence of LLM-generated content (Novikova et al., 2017). A model may generate text that scores highly on automated metrics like BLEU or perplexity, but this doesn't guarantee that the content is contextually appropriate or coherent. Human reviewers are able to assess if the text is useful, logical and consistent within its context, factually accurate, and maintains overall coherence.

Another crucial aspect of LLM evaluation is assessing creativity and novelty in language use. As LLMs are increasingly employed for creative tasks the limitations of automated metrics become evident (Bubeck et al., 2023). These metrics typically rely on comparisons with existing data and are not equipped to judge originality. Human evaluators, on the other hand, can appreciate and assess creativity, offering insights vital for fostering innovation in model development. Moreover, human input is indispensable in detecting and addressing biases in LLM outputs. Automated metrics fall short in identifying biases or ethical concerns in generated content. Human evaluators, with their understanding of societal and cultural nuances, are better positioned to spot when a model outputs biased or potentially harmful content. This human oversight is crucial for the development of responsible and ethical AI systems. Additionally, human evaluators play a pivotal role in user experience testing, particularly for LLM applications designed for human interaction (Demetriadis and Dimitriadis, 2023). Human feedback on the engagement, usefulness, and enjoyment level of these interactions is invaluable, as it provides insights that automated metrics cannot capture. This human-in-the-loop approach ensures that the models are not only technically proficient but also effective and satisfying in real-world interactions.

## 2.4 Qualitative performance

Much of the leading research on LLM evaluation involves empirical assessments using many well-known models (Xu et al., 2022; Lai et al., 2023; de Wynter et al., 2023; Zhao et al., 2023; Zhang et al., 2023; Koh, Salakhutdinov, and Fried, 2023; Liu et al., 2021). This includes GPT-3, GPT-3.5, InstructGPT, LlaMa, PaLM, and their variants. This subsection synthesizes findings from readily available off-the-shelf models and public research or leaderboard results. The goal is to summarize the overall qualitative performance of LLMs as reflected in current state-of-the-art. Notable evaluations of LLMs are detailed in studies like  (Hadi et al., 2023a; Zhao et al., 2023). These investigations assess the effectiveness and superiority of LLMs across a broad range of tasks and benchmarks, particularly in relation to the first and second dimensions of evaluation defined in Section 2.

Regarding the first dimension of evaluation, (Zhao et al., 2023) primarily focused on language generation tasks, including language modeling, conditional text generation, and code synthesis. They also concentrated on knowledge utilization and complex reasoning tasks. The authors aimed to cover the most widely discussed or studied tasks in LLM evaluation, rather than encompassing all specific tasks in the NLU and NLG fields. The findings from this investigation align with those of (Brown et al., 2020; Costa-jussà et al., 2022), showing that LLMs significantly outperform previous state-of-the-art methods on fixed-knowledge evaluation datasets. This is evident in public leaderboards (e.g., SNLI, MNLI matched, MNLI mismatched, X-NLI), where LLMs with billions of parameters demonstrate clear superiority over smaller models in considerable sized fixed-knowledge datasets. (Kaplan et al., 2020) noted that performance in language modeling tasks tends to adhere to the scaling law. This suggests that increasing the size of language models leads to improved accuracy and lower perplexity, further underscoring the advantages of scaling up LLMs.

Conditional text generation, a key task in NLG, focuses on creating text that meets specific requirements based on given conditions. Studies by (Li et al., 2022; Zhao et al., 2023) identify conditional generation as a complex task, requiring at least an understanding of machine translation, text summarization, and question answering. While evaluation for these tasks often intersects with the second dimension of evaluation, involving the use of segments from various fixed-knowledge datasets to create more intricate benchmarks, LLMs have shown exceptional performance. They excel not only on existing datasets but also on these comprehensive benchmarks, in some cases even presumably outperforming human abilities due to their advanced language generation skills. In line with these developments, (OpenAI, 2023) reported significant progress with GPT-4. This model has presumably already surpassed state-of-the-art methods, including those with benchmark-specific training, across a broad array of tasks like NLU, commonsense reasoning, and mathematical reasoning. Yet, the true nature of the model is not known, nor the evaluation procedures employed in the validation of the model. There might be several factors that affect the cited surpass, such as data contamination among others. As a consequence, until the evaluation process is clarified it must be doubted that the nature of that surpass is due to the model generalization capabilities and not to contamination (Sainz et al., 2023).

The study by (Bubeck et al., 2023) goes a step further by likening GPT-4 to an early form of AGI. They highlight GPT-4's human-like performance in real-world exams such as Advanced Placement tests and the Graduate Record Examination, covering areas like mathematics, computer vision, and programming. However, they also note significant limitations in GPT-4's performance. Consistent with the scaling law observed in fixed-knowledge evaluation, GPT-4 shows marked improvements over GPT-3.5, which itself surpassed earlier GPT versions. (Bang et al., 2023) provide a detailed analysis in which they demonstrate (in 9 out of 13 NLP datasets) the superiority of modern GPT over earlier LLMs using zero-shot learning. Their work also reveals that recent GPT versions outdo fully fine-tuned task-specific language models in 4 different tasks on the MMLU benchmark. For the rest of the scenarios, GPT's performance is comparable to, or slightly below, that of fully fine-tuned models, though statistical significance in these comparisons is not always clear. (Srivastava et al., 2022) corroborate these findings. They show that GPT-3 with context can surpass a fine-tuned BERT-Large on SuperGLUE score with only 32 example inputs. This further substantiates the scaling law's impact on LLM performance, which is still in need of further investigation. In their analysis of MMLU, (Hoffmann et al., 2022) demonstrate that LLMs nearly double the average accuracy of human raters. Notably, GPT-4 exhibits state-of-the-art performance in 5-shot settings, achieving an average accuracy improvement of over 10 % compared to the previously best-performing model.

Regarding the third dimension of evaluation, comparisons and investigations involving LLMs are less common, partly due to the high costs and complexities involved. However, recent studies, including (Creswell, Shanahan, and Higgins, 2022), indicate that automatic metrics might underestimate the quality of LLM-generated content, while human judgment tends to offer more favorable assessments. This finding outlines the increasing necessity of incorporating human evaluation into the loop, highlighting its crucial role in providing a more accurate measure of LLMs' generation capability and quality.

As efforts continue to focus on the development of new metrics that better align with human judgment, human-in-the-loop LLM evaluation is increasingly incorporating tasks that mimic pseudo-human judgments, like code synthesis. In this task, LLMs are required to do more than just generate high-quality natural language as they also need to demonstrate proficiency in creating formal language that meets specific human-defined conditions (Wang et al., 2022). This shift not only tests LLMs' natural language abilities but also their capability to adhere to structured coding requirements, offering a more comprehensive evaluation framework.

Unlike in NLG, the quality of the generated code can be directly verified through execution with appropriate compilers or interpreters. Current research performed in this domain often evaluates the effectiveness of LLMs by measuring the pass rate of the generated code against human-designed test cases, lending this method a pseudo-human evaluation character. Recent developments have seen the introduction of several code benchmarks focused on functional correctness to assess LLMs' code synthesis capabilities. As these tasks increase in complexity, smaller models often perform almost as random baselines (Perez et al., 2022; Bradbury et al., 2018; Nijkamp et al., 2023).

## 3 Discussion on LLM evaluation

Overall, state-of-the-art results in LLM evaluation reveal that increasing model size seems to continuously improve performance. (Chowdhery et al., 2022) report that the most advanced LLMs can surpass average human performance in many scenarios under a few-shot setting, particularly in well-known benchmarks assessing the models' generalizing capabilities across various fixed knowledge settings. It is important to recognize that human performance by itself is not universally defined within the state of the art. This variability underscores the complexity of directly comparing LLM capabilities with human benchmarks. However, understanding when and how LLMs develop these abilities is crucial, as highlighted by (Fu, Peng, and Khot, 2022). The fact that LLMs are primarily developed by industry players, who often don't disclose critical training details like data collection and cleaning, complicates efforts to replicate and conduct detailed analyses.

(Zhao et al., 2023) argue that, despite their progress and impact, the fundamental mechanisms underlying LLMs remain largely unexplored. Also, there is a notable uncertainty on why highly advanced abilities emerge in LLMs, while the very same abilities are absent in smaller models. This lack of understanding calls for a more in-depth examination of the key factors contributing to the superior capabilities of billion-parameter LLMs.

The study by (Bang et al., 2023) highlights certain drawbacks and limitations of LLMs, particularly in how they generalize. They identify areas where LLMs, specifically GPT variants, struggle. For example, GPT models show weaknesses in inductive reasoning, as opposed to deductive or abductive reasoning. They also lack spatial reasoning capabilities, although they perform better in temporal reasoning. Another significant limitation noted is in mathematical reasoning, a concern also echoed by (Frieder et al., 2023). Furthermore, (Bang et al., 2023) claims that GPT-like models demonstrate acceptable performance in causal and analogical reasoning. They also note that these models are relatively more proficient in commonsense reasoning compared to non-textual semantic reasoning.

All in all, there seems to be an unknown number of hazards affecting the performance of LLMs across different evaluation dimensions, but there has been limited analysis identifying these hazards. (Ji et al., 2023) conducted a thorough investigation into the hallucination hazard, which is one of the most common one. Hallucinations in LLMs refer to factual statements generated by the model that cannot be verified based on the information contained within its parametric memory, spanning all the model's knowledge. While hallucination is perhaps the most recognized hazard associated with LLMs, there exists a range of other, less-known hazards that impact evaluation. These hazards raise critical questions about the effectiveness of existing benchmarks in properly evaluating and reflecting LLMs' capabilities. Acknowledging this challenge, the next section of our study aims to highlight what we consider the most significant performance affecting hazards in LLM evaluation. This analysis spans across the three main dimensions of evaluation, aiming to provide a comprehensive understanding of the factors that influence LLM performance.

## 4 Description of main hazards

This section enumerates a comprehensive list of hazards in LLM evaluation, each linked to a specific area or dimension of evaluation they are associated with. With the understanding of the factors that influence poor performance we aim to clarify the current challenges in each evaluation dimension.

**The Reversal Curse.** This Natural Language Inference hazard refers to the phenomenon where models incorrectly assign higher probability to the reverse of a true statement. For instance, if a model recognizes "A implies B", it might also incorrectly assess "B implies A" as true, showcasing a fundamental misunderstanding of logical inference (Ma et al., 2023; Berglund et al., 2023).

**Lack of Common Sense Reasoning.** As a generalization of the previous hazard, LLMs sometimes fail in tasks requiring common sense reasoning, generating outputs that are logically absurd or factually incorrect (Kejriwal et al., 2023).

**Hallucination.** A content generation domain hazard in which LLMs produce plausible but entirely fabricated information, known as hallucinations. This is particularly hazardous in domains where factual accuracy is critical, such as for fixed-knowledge evaluation. This hazard has been very well documented in the state-of-the-art (Ji et al., 2023; Puchert et al., 2023; Bang et al., 2023).

**Interpretability and Explainability Issues.** As LLMs grow in complexity, understanding the reasoning behind their decisions becomes more challenging. This lack of transparency is a hazard in applications where understanding model decision-making is crucial for trust and reliability (Saha et al., 2023; Saha et al., 2023).

**Catastrophic Forgetting.** In the domain of the learning stability, this refers to a model's tendency to forget previously learned information upon learning new data (Zhai et al., 2023; Sun et al., 2020).

**Bias and Stereotyping.** Linked with fairness and ethics, this hazard states that LLMs can inherit and amplify biases present in their training data. This includes gender, racial, and cultural biases, leading to unfair or stereotypical outputs. This is a significant hazard where fairness and ethical considerations are paramount, such as in the third dimension (Kotek, Dockum, and Sun, 2023).

**Model Overfitting and memorization.** Generalization hazard that occurs when a model is too closely tailored to the training data and fails to perform well on unseen data (Peng, Wang, and Deng, 2023). This is a critical hazard in evaluating the model's ability to generalize beyond its training set affecting all dimensions. Serious concerns regarding memorization are raised by the authors in (Sainz et al., 2023) where they expose test data from benchmarks being present as training for LLMs in different conditions.

**Adversarial Attacks.** LLMs can be vulnerable to adversarial attacks affecting the robustness domain, where slight, often imperceptible, alterations to input data can lead to drastically different outputs. This hazard challenges the robustness and security of models (Sainz et al., 2023; Sakaguchi et al., 2021; Xu et al., 2023b).

**Vulnerability to Misinformation** When trained on data containing misinformation, LLMs can inadvertently propagate false or misleading information (Saha et al., 2023).

**Inconsistency in Long-Term Interactions.** Similar to the previous hazard, in applications involving long-term interactions, LLMs may exhibit inconsistency in personality or knowledge over time, affecting user experience and trust. Also, LLMs may struggle with understanding and maintaining context over longer conversations or texts, leading to responses that are out of context or irrelevant (Chen, Arunasalam, and Celik, 2023).

**Output Toxicity.** Affecting content safety domain, LLMs can generate harmful or offensive content, especially if they are exposed to such content. This is a significant hazard in public-facing applications (Chetnani, 2023).

**Echo Chamber Effect.** In the domain of content diversity LLMs can reinforce the same ideas or perspectives, especially if trained on homogeneous data, leading to a lack of diversity in generated content and potentially reinforcing biases (Demarco, de Zarate, and Feuerstein, 2023).

**Language and Cultural Limitations.** For cross-lingual domains, LLMs often struggle with languages with low digital resources or with cultural nuances, leading to poor performance in multilingual or multicultural contexts (Hadi et al., 2023b).

**Misalignment with Human Values.** Concerning the third dimension of evaluation and the ethical alignment, LLMs might generate outputs that are technically correct but misaligned with human ethical standards, especially in sensitive areas like medical, legal, or moral advice. Also, interactions with users can create feedback loops where the model increasingly reinforces user biases or undesirable behaviors (Chiang and Lee, 2023).

**Difficulty with Nuanced or Subtle Language.** Related to the previous hazard, LLMs may struggle with understanding and generating nuanced or subtle language, such as sarcasm, irony, or metaphor (Băroiu and Trăuşan-Matu, 2023).

**Environmental issues.** This hazard relates to addressing not only global sustainability goals, but also the long-term viability and ethical development of AI technologies (Rillig et al., 2023). The environmental impact of LLMs emerges as a critical hazard, characterized by the significant energy consumption and carbon footprint associated with their training, validation and operation. This domain transversal hazard underscores the need for sustainability in AI practices, advocating for the development and adoption of energy-efficient algorithms.

**Privacy and Copyright.** Privacy and copyright issues present a significant hazard in the context of LLMs, reflecting concerns around the unauthorized use of proprietary data and the potential for privacy breaches. Aligning with the OECD's principles on artificial intelligence fairness and ethics, it's crucial to ensure that LLMs operate within frameworks that respect copyright laws and protect personal data.

**Benchmark over-reliance.** Linked with human-centric evaluation weaknesses, benchmark over-reliance emphasizes the transversal risk of overvaluing benchmark results when assessing NLU capabilities of LLMs. This hazard challenges the notion of superhuman performance, arguing that benchmarks may not fully capture the nuances of human language comprehension and often lack transparency and fairness in comparisons. This hazard calls for the development of more comprehensive and equitable benchmarks to accurately measure and understand the capabilities of language models in relation to human performance (Tedeschi et al., 2023).

## *5 Conclusions*

This work provides a comprehensive understanding of the challenges in evaluating LLMs, focusing on the identification of key performance hazards. It emphasizes the need for continuous evolution of evaluation methods to keep up with the advancements in LLM technology and ensure responsible development and deployment. As LLMs become integral to societal frameworks, there is a growing need to emphasize the importance of multidimensional and comprehensive evaluations. The acceptance of these systems in real-world applications is tied not only to their knowledge demonstration but also to their reasoning and argumentation abilities.

Our study reviews recent research in LLMs, tracking current trends to ensure that evaluation methods keep pace with rapid advancements in the field. We analyze key evaluation dimensions with the aim of understanding factors that affect the performance of LLMs. A significant aspect of this investigation is identifying major performance hazards in LLM evaluation, an area not extensively explored previously. This approach is crucial for recognizing the potential and limitations of these AI systems in various evaluation domains. Evaluating LLMs is crucial for several reasons. First, it allows us to understand their strengths and weaknesses more clearly, and, second, enhanced evaluations offer better guidance for human-LLM interactions, informing future interaction designs and implementations.

### 5.1 Limitations on LLM evaluation

As LLMs grow in size and develop more emergent abilities, current evaluation protocols may no longer suffice to accurately assess their capabilities and potential risks. Therefore, our goal is to heighten awareness within the community about the significance of LLM evaluation. We achieve this by reviewing existing evaluation protocols and, more importantly, by highlighting the need for future research focused on developing new LLM evaluation protocols that take into account the underlying hazards that affect each dimension. This approach is crucial for keeping pace with the rapid advancements in LLM technology and ensuring their responsible development and deployment.

## *References*

Aftan, S. and H. Shah. 2023. A survey on bert and its applications. In *2023 20th Learning and Technology Conference (L&T)*, pages 161–166. IEEE.

Aiyappa, R., J. An, H. Kwak, and Y.-Y. Ahn. 2023. Can we trust the evaluation on chatgpt? *arXiv preprint arXiv:2303.12767.*

Awasthi, I., K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni. 2021. Natural language processing (nlp) based text summarization-a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1310–1317. IEEE.

Baltrušaitis, T., C. Ahuja, and L.-P. Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bang, Y., S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023.*

Baradaran, R., R. Ghiasi, and H. Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Băroiu, A.-C. and Ş. Trăuşan-Matu. 2023. How capable are state-of-the-art language models to cope with sarcasm? In *2023 24th International Conference on Control Systems and Computer Science (CSCS)*, pages 399–402. IEEE.

Bates, M. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Berglund, L., M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. 2023. The reversal curse: Llms trained on.$^{a}$ is b"fail to learn"b is a". *arXiv preprint arXiv:2309.12288.*

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

Bouziane, A., D. Bouchiha, N. Doumi, and M. Malki. 2015. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375.

Bradbury, J., R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, et al. 2018. Jax: Composable transformations of python+ numpy programs (v0. 2.5). *Software available from https://github. com/google/jax.*

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712.*

Buchanan, B. G. and E. H. Shortliffe. 1984. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence).* Addison-Wesley Longman Publishing Co., Inc.

Chang, Y., X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109.*

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374.*

Chen, Y., A. Arunasalam, and Z. B. Celik. 2023. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. *arXiv preprint arXiv:2310.02431.*

Chetnani, Y. P. 2023. *Evaluating the Impact of Model Size on Toxicity and Stereotyping in Generative LLM.* Ph.D. thesis, State University of New York at Buffalo.

Chiang, C.-H. and H.-y. Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Clark, E., S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das, and A. P. Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *arXiv preprint arXiv:2305.13194*.

Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Costantini, S. 2002. Meta-reasoning: A survey. In *Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski Part II*. Springer, pages 253–288.

Creswell, A., M. Shanahan, and I. Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

de Wynter, A., X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*.

Demarco, F., J. M. O. de Zarate, and E. Feuerstein. 2023. Measuring ideological spectrum through nlp. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023)*.

Demetriadis, S. and Y. Dimitriadis. 2023. Conversational agents and language models that learn from human dialogues to support design thinking. In *International Conference on Intelligent Tutoring Systems*, pages 691–700. Springer.

Deng, J. and Y. Lin. 2022. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, G., J. Zhao, T. Hui, D. Guo, W. Wang, B. Feng, Y. Qiu, Z. Gongque, K. He, Z. Wang, et al. 2023. Revisit input perturbation problems for llms: A unified robustness evaluation framework for noisy slot filling task. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 682–694. Springer.

Erdem, E., M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Floridi, L. and M. Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Frieder, S., L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. 2023. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Fu, Y., H. Peng, and T. Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Gamallo, P., J. R. P. Campos, and I. Alegria. 2017. A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 109–114.

Gao, J. and C.-Y. Lin. 2004. Introduction to the special issue on statistical language modeling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):87–93.

Garg, T., S. Masud, T. Suresh, and T. Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Ghazal, A., T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. 2023a. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al. 2023b. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Head, C. B., P. Jasper, M. McConnachie, L. Raftree, and G. Higdon. 2023. Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, 2023(178-179):33–46.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Huang, D., Q. Bu, J. Zhang, X. Xie, J. Chen, and H. Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*.

Jain, N., K. Saifullah, Y. Wen, J. Kirchenbauer, M. Shu, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. 2023. Bring your own data! self-supervised evaluation for large language models. *arXiv preprint arXiv:2306.13651*.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jin, Z., J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kejriwal, M., H. Santos, K. Shen, A. M. Mulvehill, and D. L. McGuinness. 2023. Context-rich evaluation of machine common sense. In *International Conference on Artificial General Intelligence*, pages 167–176. Springer.

Khalfa, J. 1994. *What is intelligence?* Cambridge University Press.

Khowaja, S. A., P. Khuwaja, and K. Dev. 2023. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *arXiv preprint arXiv:2305.03123*.

Koh, J. Y., R. Salakhutdinov, and D. Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Korb, K. B. and A. E. Nicholson. 2010. *Bayesian artificial intelligence*. CRC press.

Kotek, H., R. Dockum, and D. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Lacave, C. and F. J. Díez. 2002. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Lai, V. D., N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613.*

Lazarski, E., M. Al-Khassaweneh, and C. Howard. 2021. Using nlp for fact checking: A survey. *Designs*, 5(3):42.

Lehman, J., J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. 2023. Evolution through large models. In *Handbook of Evolutionary Machine Learning*. Springer, pages 331–366.

Li, J., T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273.*

Li, Y., M. Du, R. Song, X. Wang, and Y. Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149.*

Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110.*

Lin, S., J. Hilton, and O. Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958.*

Liu, F., E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238.*

Ma, J.-Y., J.-C. Gu, Z.-H. Ling, Q. Liu, and C. Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322.*

Mahany, A., H. Khaled, N. S. Elmitwally, N. Aljohani, and S. Ghoniemy. 2022. Negation and speculation in nlp: A survey, corpora, methods, and applications. *Applied Sciences*, 12(10):5209.

McDonald, D. D. 2010. Natural language generation. *Handbook of natural language processing*, 2:121–144.

Min, B., H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Motger, Q., X. Franch, and J. Marco. 2022. Software-based dialogue systems: survey, taxonomy, and challenges. *ACM Computing Surveys*, 55(5):1–42.

Nijkamp, E., H. Hayashi, C. Xiong, S. Savarese, and Y. Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309.*

Novikova, J., O. Dušek, A. C. Curry, and V. Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875.*

OpenAI, R. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2.

Orrù, G., A. Piarulli, C. Conversano, and A. Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in Artificial Intelligence*, 6:1199350.

Oshikawa, R., J. Qian, and W. Y. Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770.*

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Peng, Z., Z. Wang, and D. Deng. 2023. Near-duplicate sequence search at scale for large language model memorization evaluation. *Proceedings of the ACM on Management of Data*, 1(2):1–18.

Perez, E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286.*

Puchert, P., P. Poonam, C. van Onzenoodt, and T. Ropinski. 2023. Llmmaps–a visual metaphor for stratified evaluation of large language models. *arXiv preprint arXiv:2304.00457.*

Qin, C., A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476.*

Reiter, E. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Rillig, M. C., M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Ruder, S., J. H. Clark, A. Gutkin, M. Kale, M. Ma, M. Nicosia, S. Rijhwani, P. Riley, J.-M. A. Sarr, X. Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938.*

Saha, T., D. Ganguly, S. Saha, and P. Mitra. 2023. Workshop on large language models' interpretability and trustworthiness (llmit). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5290–5293.

Sainz, O., J. A. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018.*

Sakaguchi, K., R. L. Bras, C. Bhagavatula, and Y. Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Salloum, S. A., R. Khan, and K. Shaalan. 2020. A survey of semantic analysis approaches. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 61–70. Springer.

Shanahan, M. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551.*

Shin, J. and J. Nam. 2021. A survey of automatic code generation from natural language. *Journal of Information Processing Systems*, 17(3):537–555.

Song, G., Y. Ye, X. Du, X. Huang, and S. Bie. 2014. Short text classification: a survey. *Journal of multimedia*, 9(5).

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615.*

Storks, S., Q. Gao, and J. Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172.*

Sun, J., S. Wang, J. Zhang, and C. Zong. 2020. Distill and replay for continual language learning. In *Proceedings of the 28th international conference on computational linguistics*, pages 3569–3579.

Tang, R., Y.-N. Chuang, and X. Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205.*

Tedeschi, S., J. Bos, T. Declerck, J. Hajic, D. Hershcovich, E. H. Hovy, A. Koller, S. Krek, S. Schockaert, R. Sennrich, et al. 2023. What's the meaning of superhuman performance in today's nlu? *arXiv preprint arXiv:2305.08414.*

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461.*

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y., Y. Wang, J. Liu, and Z. Liu. 2020. A comprehensive survey of grammar error correction. *arXiv preprint arXiv:2005.06600*.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xu, F. F., U. Alon, G. Neubig, and V. J. Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

Xu, P., W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. 2023a. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Xu, X., K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. 2023b. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*.

Zellers, R., A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhai, Y., S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Zhang, L., S. Wang, and B. Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Zhang, R., J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. 2023. A survey of large language models. *arXiv e-prints*, pages arXiv–2303.

Zhong, W., R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Zhu, K., J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, Y. Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.