Post Turing:
Mapping the landscape of LLM Evaluation

Alexey Tikhonov
Inworld.AI
Berlin, Germany
altsoph@gmail.com

Ivan P. Yamshchikov
CAIRO, THWS
Würzburg, Germany
CEMAPRE, ISEG,
University of Lisbon, Portugal
ivan@yamshchikov.info

Abstract

In the rapidly evolving landscape of Large Language Models (LLMs), introduction of well-defined and standardized evaluation methodologies remains a crucial challenge. This paper traces the historical trajectory of LLM evaluations, from the foundational questions posed by Alan Turing to the modern era of AI research. We categorize the evolution of LLMs into distinct periods, each characterized by its unique benchmarks and evaluation criteria. As LLMs increasingly mimic human-like behaviors, traditional evaluation proxies, such as the Turing test, have become less reliable. We emphasize the pressing need for a unified evaluation system, given the broader societal implications of these models. Through an analysis of common evaluation methodologies, we advocate for a qualitative shift in assessment approaches, underscoring the importance of standardization and objective criteria. This work serves as a call for the AI community to collaboratively address the challenges of LLM evaluation, ensuring their reliability, fairness, and societal benefit.

1 Introduction

Alan Turing began his famous article "Computing Machinery and Intelligence" (Turing, 1950) by stating that it is extremely difficult to formulate objective definitions of the terms "machine" and "think" in the context of the question: *Can machines think?* Instead, he proposed looking for an answer to another question: *Can machines reliably imitate human dialogue?*

Back then, in 1950, the answers to both questions were so far apart from us that the difference between them was insignificant, and this substitution helped to set the "north star metric" for a long time, the direction of development for the entire field of research, including dialog systems, human-machine interfaces, and various kinds of AI. A possible reason for this success is that a practical solution to this imitation task implies the need to fulfill (to some extent) several complex conditions simultaneously, including natural language proficiency, interactivity, and effective grasp on the context of the conversation. Moreover, since the initial setup does not specify the fixed protocol, other strong requirements may be implied, such as common knowledge of the world, reasoning, abstract or creative thinking, concept of causality, and so on, depending on the particular interviewer's questions.

Now, 73 years after Turing's paper, modern systems have greatly evolved, successfully mimicking human-like behaviors and interactions. The first officially documented machine passed the Turing test in 2014 (Warwick and Shah, 2016), long before the era of Large Language Models. Since then, the quality of dialog simulation and text generation in general has increased even more, so the Turing test has long since ceased to serve as a reliable proxy for evaluation of modern systems. Instead, a wide variety of approaches are used in practice, aimed to assess different individual abilities and properties of a system. However, we have neither a unified system of criteria nor clear formulation of the evaluation goals. In the meantime this new evaluation methodology will not only influence the trajectory of AI research but will also have broader implications. Thus, it is paramount to ensure that LLMs are reliable, unbiased, and beneficial for society.

This paper does not set a general goal for further development of LLMs but tries to provide a comprehensive overview of the evaluation methodologies for Large Language Models and dialog agents. In Section 3, we present a chronological overview of the recent history of LLM development and their evaluation methods. Specifically, we explore benchmarks, human assessments, and model assessments, among others, that are prevalent in both academic research and practical applications. In Section 4, we propose a primary taxonomy of these approaches and discuss their strengths and internal

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 398–412
December 6, 2023 ©2023 Association for Computational Linguistics

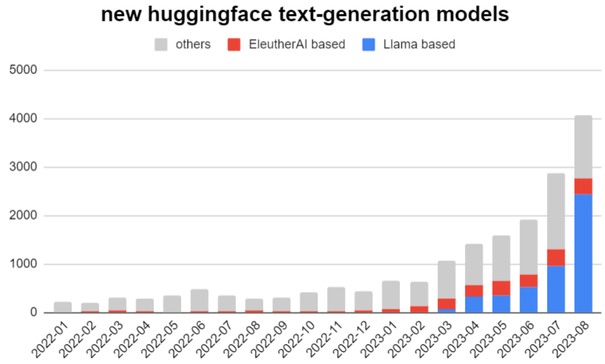

Figure 1: Cambrian explosion of large language models: the number of monthly created text-generation model repositories on huggingface, based on statistics by HF-Community.[1]

issues, including noticeable errors, problems, and contradictions. Section 5 examines which specific aspects of LLMs are commonly evaluated in contemporary studies. Finally, in Section 6 we use the proposed taxonomy to discuss current challenges and possible directions for further progress in the field.

One has to state that the current evaluation approaches have are not effective and do not meet modern requirements. Moreover, further extensive development of the existing approaches (for example, increasing the number of benchmarks and creating new tasks within existing benchmarks) cannot address these issues. We drastically need a qualitative rather than quantitative leap in evaluation. In our opinion, the first step towards a solution should be the survey of the existing evaluation taxonomy, and a detailed discussions of the weaknesses of the available methods that we try to provide in this paper.

2 "Cambrian explosion" of large language models

Lately the landscape of language models has expanded remarkably (Figure 1). As of October 2023, the number of generative text models on Hugging Face (HF) has reached a remarkable 25 000+ and 86 59 models are based explicitly on the LLaMA model (Touvron et al., 2023). This explosion can also be observed in real time[2].

With such an abundance of models, it becomes essential to evaluate and compare their quality. A state-of-the-art survey by Yang et al., 2023 provides valuable insights into the diverse applications and capabilities of language models beyond ChatGPT. However, the various works in this field employ different methodologies for assessing quality. Expansion at such a rate brings inevitable confusion[3] within the field. So, common evaluation methodologies are not only far from consistent but are also contradictory sometime.

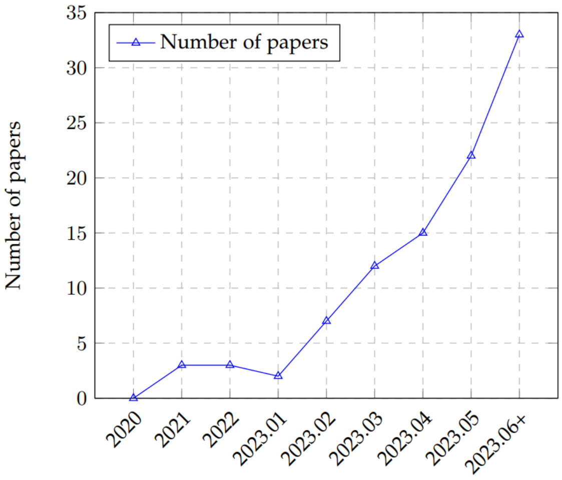

Figure 2: Trend of LLMs evaluation papers over time from Chang et al., 2023

This paper has no intention to provide a complete and comprehensive survey of the field. We suggest focusing on one aspect of LLM development that we personally see as the most crucial for the future progress of the field, namely, evaluation. However, even in this narrowed context, it is hardly possible to guarantee any form of a complete review due to the number of relevant papers on the topic (Figure 2). We address the reader to Chang et al., 2023 for an example of such a survey. In this paper, we instead discuss selected examples to illustrate the trends and challenges we are facing. We believe those examples are relevant to the field and had a high impact at the moment of their release. We do not claim we can provide a full review of all evaluation techniques used for LLMs, but to the best of our knowledge, this paper lists all significant conceptual approaches.

[1] https://som-research.github.io/HFCommunity/index.html

[2] For example, visit https://github.com/hollobit/GenAI_LLM_timeline

[3] For example, the very term "large language models" is constantly used, but there is no universally accepted threshold for the number of parameters after which the model is considered large.

3 Evolution of LLM Evaluation

Let us review the trends in LLM Evaluation. Subjectively, we split LLM development into three core periods with specific properties. We list some of the models for every period and briefly describe the methods used for performance evaluation. We do not imply that the list of the models is complete. We also list only some of the evaluation methods used for every model since they are numerous and tend to overlap. Nevertheless, we enumerate the primary evaluation methodologies so the reader can have a fair and complete representation of the spectrum of evaluation methods available today. Let us briefly discuss each period and highlight some of the methods that were used for evaluation.

3.1 "Prehistoric" LLM Evaluations

In this subsection, we discuss evaluations of models that emerged before the appearance of GPT-3[4], which was initially released in beta on June 11, 2020. We have mentioned above that there is no consensus on the threshold for the "large" language model. Thus, we suggest discussing models with more than one billion parameters[5].

During this period, the models are mainly assessed on relatively simple and common NLU benchmarks such as LAMBADA (Paperno et al., 2016), GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), SQuAD (Rajpurkar et al., 2016), MNLI (Williams et al., 2018), QQP (Wang et al., 2017), SQuAD, Winograd Schema Challenge (Levesque et al., 2012), RACE (Lai et al., 2017), or similiar. Since LLMs from this period achieved at most 50%-80% of human-level performance on these tasks, the progress across various models was clearly visible. In some papers, the authors try to devise additional metrics for model performance comparison. For example, several papers compare the perplexities using the same WikiText dataset, which is questionable since models often have different tokenization vocabularies. Hence, comparing such perplexities could only be fair with some additional tricks (see, for example, Mosin et al., 2023).

3.2 From GPT-3 to ChatGPT

During this period, before the end of 2022, the number of new LLMs has increased[6], since several major developers joined the race. These new models consistently achieved scores of 90% or higher on some of the old benchmarks (e.g., SuperGLUE, LAMBADA, SQuAD, GLUE), so they became less informative because of limitations of their sensitivity.

Consequently, researchers tend to use more complex and/or specific benchmarks, such as StoryCloze (Mostafazadeh et al., 2017), HellaSwag (Zellers et al., 2019), TriviaQA (Joshi et al., 2017), ARC (Clark et al., 2018), CoQA (Reddy et al., 2019), DROP (Dua et al., 2019), QuAC (Choi et al., 2018), SQuADv2 (Rajpurkar et al., 2018), hoping to capture nuances of different models' quality.

Moreover, new complex benchmarks (such as PIQA (Bisk et al., 2020) and Closed Book Question Answering (Wang et al., 2021)) were introduced. Notably, benchmarks such as MMLU (Hendrycks et al.), BIG-Bench (Srivastava et al., 2022) as well as HELM meta benchmark (Liang et al., 2022), often covering multiple disciplines akin to a human exam, have emerged as evaluation tools.

However, there is no universally agreed-upon system of benchmarks, leading to arbitrary comparisons across various evaluation criteria. At the same time, such an abundance of comparison scales leads to the absence of Pareto superiority for any given model[7]. Instead, authors now commonly state, "*our model outperforms the prior state-of-the-art on X out of Y tasks.*"

Another essential trend of this period is the wide usage of human labeling primarily used to deal with specific or subjective aspects of evaluation. Since the costs of high-quality human labeling are high, using an analog of the chess Elo rating, known as ELO (Arpad, 1978), established itself as a potential solution for sparse pairwise comparisons.

During this period, researchers attempt to assess the toxicity, biases, and harmful behavior of LLMs, using dedicated benchmarks together with human evaluation. In this paper, we deliberately do not discuss toxicity assessment or alignment issues, as this is a separate significant topic for which we refer to Sorensen et al., 2023.

3.3 Modern Era

Finally, we would like to highlight notable language models released in 2023 (Table 1C) and provide details about their evaluations.

[4] https://openai.com/blog/gpt-3-apps/

[5] Appendix contains Table 1 with comprehensive overview of all core models discussed in the paper

[6] See Table 1B

[7] Pareto superiority is as a situation when a new model outperforms the previous ones on all evaluation tasks.

The introduction of open models such as LLaMA and Pythia (Biderman et al., 2023), among others, has significantly increased the number of researchers conducting experiments with LLMs. Since the number of models is rising exponentially, see Figure 1, probably, a couple of new models appeared just while you read this paper. We have no intent to enumerate all available LLMs; instead, we try to capture the main trends and patterns here:

- the development and heavy usage of various complex benchmarks continues,
- many new evaluations are based on human school exams or other tests initially designed for humans, such as GMAT, SAT, LSAT, etc.
- toxicity/bias/hate speech assessments (as well as helpfulness, honesty, and harmlessness) become a mandatory attribute of the overall model evaluation,
- the complexity of the evaluation criteria motivates researchers to use pairwise evaluation when possible,
- high costs of pairwise labeling lead to the extensive use of other, superior models (mainly ChatGPT or GPT-4) for evaluation,
- these sparse pairwise or side-by-side evaluations, combined with an Elo rating system, enable the creation of leaderboards for model comparison.

Another trend worth mentioning is the rise of code-generation LLMs since they have significant specifics in application and evaluation approaches. We mention just some of them, including StarCoder (Li et al., 2023), CodeGeeX (Zheng et al., 2023b), and WizardCoder (Luo et al., 2023). Such models usually utilize special benchmarks with auto-tests for generated code (including HumanEval (Chen et al., 2021), HumanEval+ (Liu et al., 2023), DS-1000 (Lai et al., 2022), or MBPP (Austin et al., 2021)).

4 Prevalent Evaluation Methodologies

As the field evolved, several generalized approaches to evaluation established themselves. These include comparing the models on a set of benchmarks, assessment by humans, and modeling human evaluation (either using heuristics, dedicated models, or a superior LLM model). Each of these approaches has its advantages, limitations, and potential drawbacks. Let us analyze them sequentially to understand their specifics.

4.1 Comparison on benchmarks

Benchmarks may provide a fast and reliable evaluation of models. In some sense, benchmark evaluation resembles commonly used tests for human performance evaluation. The critical requirements here are the standardization of test sets and the controlled environment of evaluation. There are several interesting developments towards standardization such as HELM[8], BIG-Bench[9] or Gao et al., 2021. The last one makes an interesting step to provide a unified benchmarking framework that includes 200+ tasks for evaluation and supports a variety of available LLMs.

At the same time, similarly to human tests, LLM benchmarks have disadvantages:

- While we are in the active phase of LLM quality improvement, old benchmarks become obsolete quickly; however, they are often still included in the evaluation procedures.
- Since new benchmarks are not fully standardized yet, they often overlap or contradict, which may lead to some inconsistency.
- Taking into account the low number of tasks per topic (for example, MMLU consists of 57 types of questions on mathematics, history, psychology, etc., with an average of 280 questions per topic), the randomness may affect the outcome for each topic a lot. For example, it was shown that minor changes in the multiple-choice formatting can cause a performance jump of 6-10 points on MMLU[10]. The standard way to deal with noise is to measure confidence intervals; however, the limited data available does not enable the use of bucket test statistics.
- A tempting idea for noise control is averaging results across several different independent benchmarks and publishing the resulting ratings[11]. However, the resulting rating often

[8] https://crfm.stanford.edu/helm

[9] https://github.com/google/BIG-bench

[10] https://twitter.com/ArmenAgha/status/1669084129261162497

[11] See some examples: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, https://shorturl.at/DGPW3, https://github.

fails to account for possible methodological flaws or deliver a tangible value to a larger NLP community (Ethayarajh and Jurafsky, 2020).

- The known problem of standardized benchmark evaluations is leakage or so-called *test set pollution* since some of the benchmarks have been available on the internet for years (e.g., MMLU since 2021) and can easily occur in pre-training or fine-tuning datasets. A couple of such recent high-profile cases have sparked heated discussion in the community[12], and led to criticism in satirical papers like Schaeffer, 2023.

- Another known issue of modern benchmarking is its massive computational costs: benchmarks typically have the order of 10^5 validation examples, with 10^3 - 10^4 per task, extending the load up to hundreds of GPU hours per model evaluation. Some recent works, like Vivek et al., 2023 and Perlitz et al., 2023, try to reduce these computational costs, but it is still hard to keep the reasonable stability of results simultaneously.

- Also, as we mentioned before, reducing the number of test topics or tasks may be dangerous in terms of intended or unintended cherry-picking, making it easy to choose the ones where a particular model performs well.

Summing up, using benchmarks is a good starting point for rough evaluation. However, benchmarks have several significant drawbacks, including insufficient standardization, high computational costs, poor robustness to noise, and frequent cases of test set leakage. Moreover, benchmark assessments often do not agree with the human assessment of the model performance[13], making, potentially, the whole evaluation inconsistent. Let us now discuss the human evaluation more thoroughly.

4.2 Evaluation by Human Assessors

Evaluation by human assessors is an expensive yet widely used approach. While it may be possible to train and use a dedicated model for almost any well-formulated aspect of evaluation, the core problem is precisely in formulating a detailed definition of the evaluation criteria. The typical way to evade this is by asking about assessors' overall preference in a pairwise (side-by-side) setup and then building a rating between available models or configurations based on these pairwise scores. However, this workaround comes with its own set of challenges and drawbacks.

First, the complete pairwise evaluation is too expensive and time-consuming to compare a significant number of models since the complexity of the procedure grows like $O(n^2)$ with the number of compared models.

Second, pairwise comparisons can yield non-transitive results, making it challenging to establish a consistent global ranking. In other words, without clearly articulated criteria, human assessors may prefer system A to system B, system B to system C, and system C to system A. Researchers use different methods to deal with such situations. One alternative could be Elo rating[14] or relative comparison of evaluated models with one clearly weaker LLM. For an example of a more advanced ranking method, see Lou et al., 2022.

On the other hand, numerous co-existing leaderboards[15] may provide different rankings for the same models since they are based on different sets of noisy human pairwise labels, while the noise measurements and confidence intervals are usually absent due to the low amount of data.

Another significant issue is the quality of human labels, which can be relatively low for different reasons. Human assessors' motivation is sometimes insufficient to provide high-quality answers; moreover, some assessors secretly use LLMs as to speed up the labelling (Veselovsky et al., 2023). This might introduce unexpected shifts in the obtained assessments. Furthermore, the absence of global criteria may lead to situations when human assessors prefer more good-looking and stylish responses to correct and factual ones (Gudibande

com/FranxYao/chain-of-thought-hub, https://cevalbenchmark.com/static/leaderboard.html, https://bellard.org/ts_server/, https://huggingface.co/spaces/toloka/open-llm-leaderboard

[12] Check, for instance, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard/discussions/213

[13] Some examples of such inconsistency are available at https://llm-leaderboard.streamlit.app/ or https://github.com/LudwigStumpp/llm-leaderboard

[14] Elo ratings have their own limitations discussed in (Szczecinski and Djebbi, 2020).

[15] Examples include https://chat.lmsys.org/?leaderboard, https://github.com/LudwigStumpp/llm-leaderboard, https://aviary.anyscale.com/, and https://llm-leaderboard.streamlit.app/

et al., 2023).

Since the research community tend to treat human assessment as an expensive ground truth, researchers often try to model human evaluation with heuristics or some dedicated algorithm to reduce the evaluation's complexity and cost. Let us discuss these methods in the following subsection.

4.3 Modeling Human Evaluation

One of the common ways to obtain a cheaper estimation of human assessment is to train a dedicated model on existing human labels to predict them and then use it as a replacement for human assessors. Dozens of such approaches are proposed; for example, in the domain of dialog agents evaluation there are methods like FED (Mehri and Eskenazi, 2020), USL (Phy et al., 2020), Flowscore (Li et al., 2021), QuestEval (Scialom et al., 2021), Open AI detector[16], CT Score[17], FULL score (De Bruyn et al., 2022), Reranker[18], Cross-Encoder[19] for MS-Macro[20], Quality Adapt (Mendonca et al., 2022), Deam score (Ghazarian et al., 2022), RankGen (Krishna et al., 2022) and many others.

Although successfully implementing a human preferences model is usually necessary for the RLHF to have the so-called *Reward Modeling*, there is still no ultimate solution. However, the situation has changed significantly with the appearance of modern LLMs since one can compare the outputs of to models using a superior one.

As of today, GPT-4 is the most prominent candidate for such a superior model, which can be used (see, for example, Zheng et al., 2023a) to evaluate or compare the candidates instead of humans without additional fine-tuning. Moreover, Thomas et al., 2023 reports that GPT-4 produces better relevance labels than third-party workers. However, even GPT-4 has a couple of known significant issues, including:

- GPT-4 is also known to have a specific vocabulary bias, particularly it prefers its own generations more than humans do (Zhou et al., 2023),
- GPT-4 seems to have specific positional biases[21],
- Some systematic contradictions between GPT-4 and human assessment are reported (Xu et al., 2023),
- GPT-4 biases may be misaligned with human biases, which makes the idea of the blind comparison by a GPT-4 model quite challenging.

Such problems are not specific to GPT-4 but appear in the results of different models in different ways. The recent paper on the CoBBLEr benchmark (Koo et al., 2023) studies these effects across 15 existing LLMs.

Overall, it seems like we cannot avoid a clear definition of what we are evaluating without introducing significant noise or bias into the results.

5 What Are We Evaluating?

With dozens of actively used benchmarks with hundreds of task types, researchers naturally tend to group them into general aspects of the model's performance, so providing several high-level scores becomes standard practice. Often, researchers present them as so-called *radar diagrams* to highlight the advantages and disadvantages of the given model over baselines.

However, an overview of recent papers reveals no structure or system of these aspects, even on the highest level (see Figure 3). Sometimes, they remind the famous fiction animals classification (Borges), mixing different types and principles altogether. Building a proper taxonomy for these aspects is a complex and extensive endeavor, far beyond the scope of this paper. For deeper insights on this topic, we address the reader, for example, to Ziyu et al., 2023 or Xuanfan and Piji, 2023. Here, we just mention some commonly used approaches and group them intuitively, then discuss the results.

- **Text-specific and dialog-specific abilities** are crucial since textual dialogues are the common medium for modern LLMs. They may include:
  - General text comprehension and natural language understanding (for example, LAMBADA benchmark);

[16] https://huggingface.co/roberta-base-openai-detector

[17] https://github.com/tanyuqian/ctc-gen-eval

[18] https://github.com/luyug/Reranker

[19] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

[20] https://github.com/microsoft/MSMARCO-Passage-Ranking

[21] https://twitter.com/nazneenrajani/status/1667224735573487616

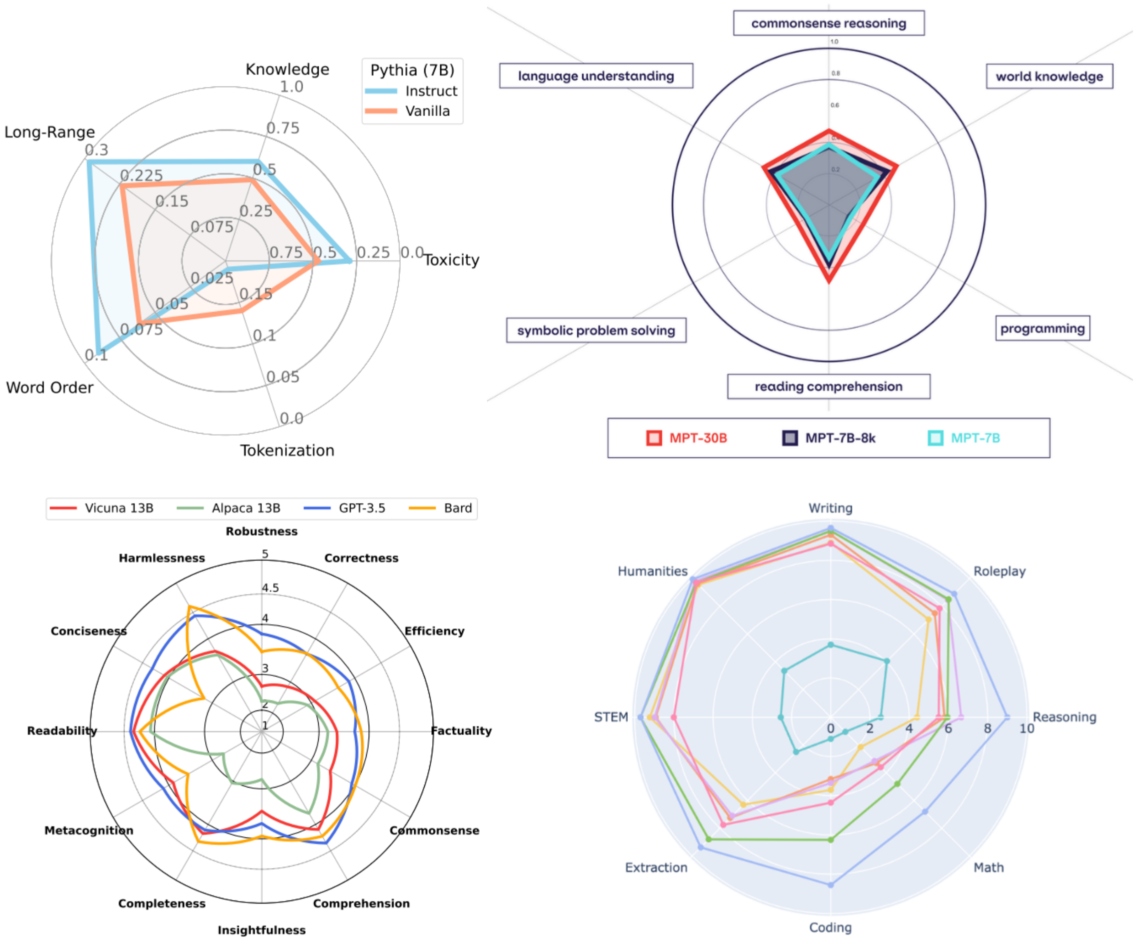

Figure 3: Radar diagrams for several recent models. Top-left is from (Jain et al., 2023), top-right is from Mosaic Eval Gauntlet, bottom-left is from (Ye et al., 2023), bottom-right is from the Giraffe-70b release.

  - Multilingualism (many options, including recently published BELEBELE (Bandarkar et al., 2023));
  - Plausibility of dialog communication;
  - Capability to understand and control the text quality, style, and level of details;

- **Knowledge-specific characteristics** - characteristics of knowledge obtained by the model during training:
  - Common knowledge is essential since human communication is built on the existence of implicitly shared contexts (Clark and Brennan, 1991);
  - Depending on the context or application, we may want to assess models' niche knowledge, such as Humanities or STEM; benchmarks here are usually compiled based on human exams or manually crafted tests like BIG-Bench;

- **Skill-specific abilities** - abilities to solve problems that require some skills besides knowledge:
  - Commonsense reasoning[22];
  - Abstract reasoning and ability to generalize[23];
  - Specific skills (Code generation, Roleplay, Math reasoning, Image manipulation, Chess problem solving, etc.)[24];

- **Personality and CogSci features** - since the general modern models' UI is a dialog via chat, users and researchers tend to treat them

[22] See a survey on Commonsense Reasoning benchmarks in (Davis, 2023)

[23] (Chollet, 2019) proposes to assess reasoning without modulation by prior knowledge and experience

[24] There are many specific skills benchmarks, see, for example, the recent NuclearQA bencmark (Acharya et al., 2023) or the RoleLLM framework (Wang et al., 2023b)

as personalities; this leads to the idea of corresponding attributes measurement:

  - Creativity[25], Empathy, Emotional Intelligence (Wang et al., 2023a), or Social awareness (Zhan et al., 2023);
  - Cognitive Science-related aspects include planning and cognitive mapping abilities (Momennejad et al., 2023), deductive competence (Seals and Shalin, 2023), and complex reasoning skills (Kuo et al., 2023);

- **Alignment, Reliability, and Safety related features**, including
  - Alignment to human values[26];
  - Security, which encompasses various aspects, like privacy, preventing malicious use, and addressing potential biases;
  - H4 attributes[27], namely being Helpful, Honest, Harmless, and Huggy, reflecting positive social qualities;
  - Factuality (Chen et al., 2023), truthfulness, and the ability to acknowledge uncertainty or lack of knowledge;
  - Explainability[28];

- **Technical characteristics** (including Long-range context (Dong et al., 2023), tokenization quality, etc)

These diverse evaluation dimensions highlight the multifaceted nature of assessing language models, each with unique considerations and challenges. For example, the precise definition of text style remains challenging (Tikhonov and Yamshchikov, 2018), while storytelling evaluation needs a deeper understanding of the concept of narrative (Gervás et al., 2019; Yamshchikov and Tikhonov, 2023). Indeed, the evaluation guidelines proposed in (Hämäläinen and Alnajjar, 2021) for creative, generative systems are relevant for the LLM evaluation in general: *"clearly defining the goal of the generative system, asking questions as concrete as possible, testing the evaluation setup, using multiple different evaluation setups, reporting the entire evaluation process and potential biases clearly, and finally analyzing the evaluation results more profoundly than merely reporting the most typical statistics."*

A well-defined and structured list of aspects we want to evaluate LLM on is essential to optimize and prioritize the evaluation of language models. Do we really need them all? How do they interrelate? Without a clear understanding of what aspects we are assessing and why, it becomes difficult to focus on specific areas for improvement or to allocate resources effectively.

6 Discussion

Let us now try to sketch the main trends in evaluation approaches and hypothesize their further development in the context of the multiple challenges we highlighted above.

6.1 Human-like Evaluation

It is worth noting that most of the current approaches to model evaluation listed in this paper are essentially anthropocentric. One reason for this may be that benchmarks are opportunity-driven. Instead of creating new, specifically targeted tests, many researchers adapt existing ones created for humans in the past.

At first glance, this simplifies not only their creation but also the interpretation of results. However, some of these tests are designed specifically for assessing human adults and might not be well suited for evaluating a broader range of signatures of intelligent behavior (Eisenstein, 2023).

Another disadvantage of this approach is that it may limit the assessment scale. Now, when superhuman performance has been achieved in some tasks, this may become a constraint or an extra incentive that distorts goal setting. For example, the need to pass a classical Turing test may encourage a model to deceive the tester and hide part of its abilities (as it may be given away by too high a calculation speed or too deep an encyclopedic knowledge).

Suppose we want to drive and track the development of models' abilities at levels qualitatively higher than the current humans. In that case, we should consider creating fundamentally new approaches, for example, developing particular competitive evaluation environments that assess not built-in knowledge and abilities but the speed and quality of forming new skills in an interactive, unfamiliar environment. We see the ARC benchmark

[25] https://bit.ly/3rKZWLm

[26] See the survey by Yao et al., 2023

[27] https://huggingface.co/HuggingFaceH4

[28] Though, Hsia et al., 2023 recently showed the flaws of available explainability metrics.

from Chollet, 2019 as a good step in this direction.

6.2 Decompose and Conquer

However, there is one thing we might want to use from the experience of human skills testing. Just like human IQ test are split into several subcategories, like Short-Term Memory, Reasoning, and Verbal (Hampshire et al., 2012), we need to divide potential LLM skills into a standardized system and define generic baselines.

There still are debates about whether it is possible to develop a universal measure of intelligence. In the meantime, we clearly see the progress of LLMs across specifically defined tasks. With limited resources and various practical tasks, developers may not want to build universally superior models. Instead, they can focus on the selected skills and abilities. For example, creators of a code assistant should not bother themselves with improving the literature style of their model too much. We believe that this tactic of "decompose and conquer" will further dominate the field, so making the rules, requirements, and systematic baselines global and public should benefit the whole community.

6.3 Nobody's Perfect

Another interesting observation is that we tend to perceive and evaluate modern models as agents in communication with humans. We earnestly expect LLMs to behave in a socially acceptable way – imposing requirements like factuality, harmlessness, helpfulness, etc.

For some parameters, we impose stricter requirements on the evaluated models than we would if we were evaluating ordinary people (e.g., we may allow some sloppiness, inattention, or carelessness from a living person, but we require models to be free of such problems). These strong demands might be rooted in the fact that we already use such models to create mass services in which they act as experts in some narrow field (data processing, science, medicine, law, etc).

Accordingly, we already expect LLMs to have confident and stable expert knowledge and skills in the target domain, implying that requirements like natural language skills and the ability to maintain a conversation are self-evident. This perfectionist bias appears likely to stay with us and potentially intensify, as testing specific skills in models will become increasingly complex and expensive.

6.4 Independent Evaluation Bodies

The evaluation and certification of LLMs could be a separate field in itself. Indeed, various global organizations work on evaluations of various human cognitive skills. There is no reason why a similar pattern could not emerge for LLMs. Creating efficient leak-proof test methodologies will only be more demanding as the models progress. At the same time, for-profit organizations clearly need some form of evaluation to compare their solutions with the competition. This might create a market incentive for the creation of for-profit organizations that could be centered around LLM certification and evaluation.

7 Conclusion

This paper provides an overview of the current state of evaluation techniques used for LLMs and analyzes them. We trace the progress of LLMs in the last few years and create a taxonomy of the methods used to evaluate LLM performance. One by one, we analyze significant approaches and highlight challenges that arise with them, including insufficient standardization, poor robustness to noise, and test set leakage of benchmarks; frequent cases of disagreement between benchmark-based evaluations, humans' and superior models' preferences; humans' and superior models' biases; dead ends of Pareto optimization and non-transitive results in the absence of global criteria; no structure or system of aspects of evaluations, even on the highest level.

Based on these observations, the current evaluation approaches have lost their effectiveness and do not meet modern requirements, and there is no clear way to patch them. In our opinion, the first step towards a solution should be the standardization of tasks and evaluation methods, including a precise formulation of the assessed aspects. We still do not know whether there is a new single "Turing question" that can set the main direction of the industry for the following decades. What is certain is that to figure out how to move forward, we need to precisely articulate what we want to measure and for what reason.

References

Anurag Acharya, Sai Munikoti, Aaron Hellinger, Sara Smith, Sridevi Wagle, and Sameera Horawalavithana. 2023. Nuclearqa: A human-made benchmark for language models for the nuclear domain.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Elo Arpad. 1978. The rating of chessplayers, past and present. *Arco Pub*, 216.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Jorge Luis Borges. John wilkins' analytical language. in weinberger e et al., ed. and trans. page 229–232.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

François Chollet. 2019. On the measure of intelligence.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2022. Open-domain dialog evaluation using follow-ups likelihood. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 496–504.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Michael Eisenstein. 2023. A test of artificial intelligence. *Nature*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Pablo Gervás, Eugenio Concepción, Carlos León, Gonzalo Méndez, and Pablo Delatorre. 2019. The long path to narrative generation. *IBM Journal of Research and Development*, 63(1):8–1.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. Deam: Dialogue coherence evaluation using amr-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human evaluation of creative nlg systems: An interdisciplinary survey on recent papers. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 84–95.

Adam Hampshire, Roger R. Highfield, Beth L. Parkin, and Adrian M. Owen. 2012. Fractionating human intelligence. *Neuron*, 76(6):1225–1237.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022a. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022b. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C Lipton. 2023. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*.

Neel Jain, Khalid Saifullah, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Bring your own data! self-supervised evaluation for large language models.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *arXiv preprint arXiv:2205.09726*.

Mu-Tien Kuo, Chih-Chung Hsueh, and Richard Tzong-Han Tsai. 2023. Large language models on the chessboard: A study on chatgpt's formal language comprehension and complex reasoning skills.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2022. Ds-1000: A natural and reliable benchmark for data science code generation. *arXiv preprint arXiv:2211.11501*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*.

Hao Lou, Tao Jin, Yue Wu, Pan Xu, Quanquan Gu, and Farzad Farnoud. 2022. Active ranking without strong stochastic transitivity. In *Advances in Neural Information Processing Systems*, volume 35, pages 297–309. Curran Associates, Inc.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 225.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval.

Vladislav Mosin, Igor Samenko, Borislav Kozlovskii, Alexey Tikhonov, and Ivan P Yamshchikov. 2023. Fine-tuning transformers: Vocabulary transfer. *Artificial Intelligence*, page 103860.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. Efficient benchmarking (of language models).

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL 2018*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rylan Schaeffer. 2023. Pretraining on the test set is all you need.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

S. M. Seals and Valerie L. Shalin. 2023. Evaluating the deductive competence of large language models.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Leszek Szczecinski and Aymen Djebbi. 2020. Understanding draws in elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 16(3):211–220.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Alexey Tikhonov and Ivan P Yamshchikov. 2018. What is wrong with style transfer for texts? *arXiv preprint arXiv:1808.04365*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2023. Anchor points: Benchmarking models with much fewer examples.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book qa? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023a. Emotional intelligence of large language models.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Kevin Warwick and Huma Shah. 2016. Can machines think? a report on turing test experiments at the royal society. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(6):989–1007.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Ni Xuanfan and Li Piji. 2023. A systematic evaluation of large language models for natural. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 40–56.

Ivan Yamshchikov and Alexey Tikhonov. 2023. What is wrong with language models that can not tell a story? In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 58–64, Toronto, Canada. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values – a survey of alignment goals for big models.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023b. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109.

Appendix

A. The "prehistoric" era of LLM	
2019, GPT-2[a]	LAMBADA, WSC, QA, summarization, translation tasks, etc.
2019, T5 (Raffel et al., 2020)	GLUE, SuperGLUE, SQuAD, QA, summarization, translation tasks, etc.
2019, CTRL (Keskar et al., 2019)	no include explicit quality measurements.
2019, Megatron-LM (Shoeybi et al., 2019)	LAMBADA, MNLI, QQP, SQuAD, RACE, etc.
2020, Turing-NLG[b]	LAMBADA, summarization, etc.
B. From GPT-3 to ChatGPT	
2020, GPT-3(Brown et al., 2020)	LAMBADA, StoryCloze, HellaSwag, Closed Book Question Answering, TriviaQA, PIQA, ARC, CoQA, DROP, QuAC, SQuADv2, RACE, SuperGLUE, NLI, OpenBookQA, some other tasks inspired by human school exams, and human side-by-side evaluation.
2021, Blenderbot (Shuster et al., 2021)	human side-by-side evaluation.
2021, Gopher (Rae et al., 2021)	152 diverse tasks from different benchmarks, including LAMBADA, MMLU, BIG-bench, TriviaQA, NaturalQuestions, TruthfulQA, PIQA, WinoGrande, SocialIQA, HellaSwag, plus some tasks inspired by human school exams, plus some toxicity, bias and hate speech evaluation.
2021, GLaM (Du et al., 2022)	compared to GPT-3 and Gopher across 29 benchmarks.
2022, OPT (Zhang et al., 2022)	compared to GPT-3 across 16 tasks, plus some toxicity, bias and hate speech evaluation.
2022, LaMDA (Thoppilan et al., 2022)	human assessments on specific aspects, including sensibleness, specificity, interestingness, safety, and factual grounding.
2022, PaLM (Chowdhery et al., 2022)	evaluated on 29 benchmarks, which were similar to the set of tasks used for GPT-3 + MMLU and BIG-Bench.
2022, Chinchilla (Hoffmann et al., 2022a,b)	benchmarks included MMLU, BIG-bench, and other.
2022, BLOOM (Scao et al., 2022)	20 benchmarks, which were a subset of those used for GPT-3.
2022, InstructGPT[c]	human assessments of specific aspects, used Elo rating.
2022, ChatGPT[d]	evaluations were conducted based on InstructGPT.
C. The "modern" era	
2023, GPT-4[e]	benchmarks including MMLU, HellaSwag, WinoGrande, and others + academic and professional examinations.
2023, LLaMA (Touvron et al., 2023)	MMLU, HellaSwag, WinoGrande, ARC, and more.
2023, Alpaca (Taori et al., 2023)	minimal evaluation.
2023, Claude[f]	minimal evaluation.
2023, Vicuna (Chiang et al., 2023)	side-by-side compared to Alpaca and LLaMa by GPT-4 as a judge.
2023, WizardLM (Xu et al., 2023)	side-by-side assessment by human evaluators and GPT-4.
2023, MPT family of models[g]	several standard benchmarks + code specific tasks, like HumanEval.
2023, Palm-2 (Anil et al., 2023)	similar to GPT-4 - a lot of standard benchmarks (including, for example, BIG-Bench and Winogrande) + language proficiency exams.
2023, Claude-2[h]	benchmarks, alignment, lanugages, long context.
2023, Falcon (Almazrouei et al., 2023)	standard benchmarks, including ARC, HellaSwag, MMLU, TruthfulQA.

[a] https://openai.com/research/better-language-models
[b] https://shorturl.at/epK79
[c] https://openai.com/research/instruction-following
[d] https://openai.com/blog/chatgpt/
[e] https://openai.com/gpt-4
[f] https://www.anthropic.com/index/introducing-claude
[g] https://github.com/mosaicml/llm-foundry
[h] https://www.anthropic.com/index/claude-2

Table 1: Selected examples of LLM Evaluation approaches