Global Journal of Engineering and Technology [GJET].
ISSN: 2583-3359 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journal-gjet-home/

RESEARCH ON LANGUAGE MODELS TOWARDS INCREASED TRANSPARENCY

BY

Ayse Kok Arslan
Researcher Silicon Valley Oxford Northern California Alumni San Francisco, Northern California, United States

Article History

Received: 15/12/2022
Accepted: 19/12/2022
Published: 21/12/2022

Vol – 1 Issue – 4

PP: - 01-06

Abstract:
One of the mostly advanced artificial intelligence (AI) technologies in recent year has been language models (LM) which is the object of the study with the aim to necessitate a comparison or benchmark among many LM to enhance transparency of these models. The essence of the results aim to provide a fuller characterization of LMs rather than to focus on a specific aspect in order to increase societal impact.

As a result of this study, a framework for designing a LM benchmark has been developed with a focus on metrics, model, and scenarios so that one can taxonomize the vast design space of language model evaluation into scenarios and metrics.

Based on core scenarios one can comprehensively measure major metrics (accuracy, calibration, robustness, fairness, bias, toxicity, efficiency). One can also evaluate existing LMs under the standardized conditions of the benchmark, ensuring models can now be directly compared across many scenarios and metrics.

The results can be explained based on the suggested model – scenario, adaptation, metric-required to provide a roadmap for how to evaluate language models. The results can be used under the condition that rather than assuming it as a complete model, it is a step towards the design of more sophisticated models and aims to raise awareness of the importance of developing benchmarks for AI models.

Keywords: *language models, artificial intelligence, transformers, benchmark, neural language programming, transparency, neural networks.*

UDC 006.659.2

1. Introduction

When it comes to topics such as understanding, reasoning, planning, and common sense, scientists are divided about how to assess LM (language models).

At its core, a LM is a box that takes in text and generates text (Fig. 1). LMs are general purposes text interfaces that could be applied across a vast expanse of scenarios. For each scenario, there may be a broad set of desiderata such as accuracy, fairness, efficiency, etc. among many others.

The problem to be solved in this paper is the design of a benchmark model to evaluate and compare LM models.

As a simplified example, we consider chess as a complicated intelligence challenge because, on their way to mastering chess, human beings must acquire a set of cognitive skills through hard work and talent. Yet, from a computational perspective, there can be a shortcut for finding good chess moves through a good algorithm and the right inductive biases.

As this example demonstrates, even some of the most carefully crafted benchmarks can be prone to computational shortcuts. In other words, while benchmarks are a good tool to compare machine learning models against one another, they are not definite and only measures of cognitive skills in machines.

This rapid proliferation of LMs necessitates a comparison or benchmark among many language models.
The research aim consists of two aspects:

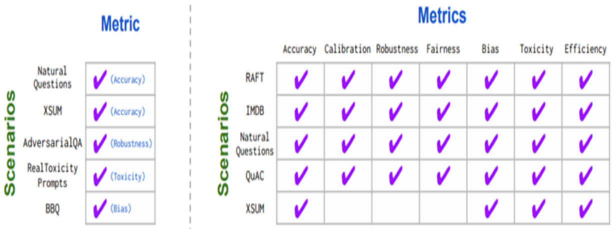
– The scientific aspect includes the design of a benchmark model for comparing language models.
– The practical aspect includes increased awareness on the challenge of comparing LM models as authors of [1] claim that benchmarks given their encoded values-specify directions for the AI community to be improved upon.

2. Materials and Methods

The object of this research is a benchmark model to evaluate and compare LM models.

When implemented and interpreted appropriately, benchmarks enable the broader community to better understand AI technology and influence its trajectory. In general, a benchmark involves three elements (Fig. 1):

1. *Broad coverage and recognition of incompleteness.*
   As it is not possible to consider all the scenarios and the desiderata that (could) pertain to LMs, a benchmark should provide a top-down taxonomy and make explicit all the major scenarios and metrics that are missing.
2. *Multi-metric measurement.*
   Societally beneficial systems reflect many values, not just accuracy. A benchmark should represent these plural desiderata, evaluating every desideratum for each scenario considered.
3. *Standardization*.
   As the object of evaluation is the LM, not a scenario-specific system, the strategy for adapting an LM to a scenario should be controlled for.



Natural Questions	✓ (Accuracy)
XSUM	✓ (Accuracy)
AdversarialQA	✓ (Robustness)
RealToxicity Prompts	✓ (Toxicity)
BBQ	✓ (Bias)

	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✓	✓	✓	✓	✓	✓	✓
IMDB	✓	✓	✓	✓	✓	✓	✓
Natural Questions	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓
XSUM	✓				✓	✓	✓

Fig. 1. Overview of an LM Benchmark Example

Overall, a benchmark builds transparency by assessing LMs in their totality. Rather than focusing on a specific aspect, the aim is to strive for a fuller characterization of LMs to improve scientific understanding and increase societal impact.

A benchmark of LM has two levels:

1) an abstract taxonomy of scenarios and metrics to define the design space for LM evaluation; and
2) a concrete set of implemented scenarios and metrics that were selected to prioritize coverage (e. g. different English varieties), value (e. g. user-facing applications), and feasibility (e. g. limited engineering resources).

When doing a benchmark some key considerations should be taken into account. To begin with, while standardizing a model evaluation, in particular by evaluating all models for the same scenarios, same metrics, models themselves may be more suitable for particular scenarios, particular metrics, and particular prompts/adaptation methods.

Moreover, while the evaluation itself may be standardized, the computational resources required to train these models may be very different (e. g. resource-intensive models generally fare better in our evaluation).

Furthermore, models may also differ significantly in their exposure to the particular data distribution or evaluation instances in use, with the potential for train-test contamination.

Even for the same scenario, the adaptation method that maximizes accuracy can differ across models which poses a fundamental challenge for what it means to standardize LM evaluation in a fair way across models.

Given the myriad scenarios where LMs could provide value, it would be appealing for many reasons if upstream perplexity on LM objectives reliably predicted downstream accuracy. Unfortunately, when making these comparisons across model families, even when using bits-per-byte (BPB), which could provide more comparison than perplexity-, this type of prediction might not always work well.

LMs are a sub-category of NLP (neural language programming) within the field of AI. As in any other field of AI, the challenge of transparency in AI models and datasets continues to receive increasing attention from academia and industry.

When it comes to developing an AI model, ***producers*** are upstream creators of dataset and documentation, responsible for dataset collection, ownership, launch, and maintenance. ***Agents*** are stakeholders who read transparency reports and possess the agency to use or determine how themselves or others might use the described datasets or AI systems.

Agents are distinct from ***users***, who are individuals and representatives who interact with products that rely on models trained on dataset. Users may consent to providing their data as a part of the product experience and require a significantly different set of explanations and controls grounded within product experiences.

Dataset design also plays a crucial role for LM development. All data is processed in a common markdown format to blend knowledge between sources. For the interface, one can use task-specific tokens to support different types of knowledge. Uncurated data also means more tokens with limited transfer value for the target use case; wasting compute budget.

Transparency refers to *a clear, easily understandable, and plain language explanation of what something is, what it does, and why it does that*. Table 1 includes core aspects of transparency.

Table 1

Core traits of transparency

Transparency Characteristic	*Description*
Balance opposites	Disclosing information without leaving actors vulnerable or reporting fairness analyses without legitimizing unfair systems are some examples.

*Corresponding Author: AyseKok Arslan

Increase in expectations	Any information included in an artefact can be expected to receive more scrutiny.
Constant availability	Transparency information should be made available at multiple levels.
Check and balances	Transparency artifacts should be subject to 3rd party evaluation as excessive transparency can make an AI system vulnerable for adversarial actors.
Trust enabler	Accessible and relevant information abut AI systems increases the willingness of a user to take a risk based on the expectation of benefits from data, algorithms, and products they use.
Reduce knowledge asymmetries	Cross-disciplinary collaboration is more effective were there is a shared mental model and vocabulary to describe the aspects of AI systems.
Reflects human values	It comes from both technical and non-technical disclosure about assumptions, facts, and alternatives.

Yet, attempts to introduce standardized and sustainable mechanisms for transparency is hindered by real-world constraints of the diversity of goals, workflows, and backgrounds of individual stakeholders participating in the life cycles of datasets and AI systems.

In order to increase the transparency of NLPs, it might be useful to gain an understanding of the different tasks that they accomplish.

To begin with, Question answering (QA) is a fundamental task in NLP that underpins many real-world applications including web search, chatbots, and personal assistants. QA is very broad in terms of the questions that can be asked and the skills that are required to arrive at the answer, covering general language understanding, integration of knowledge, and reasoning [2, 3].

Information retrieval (IR) refers to the class of tasks concerned with searching large unstructured collections (often text collections), is central to numerous user-facing applications. IR has a long tradition of study as mentioned by authors of [4]and is one of the most widely deployed language technologies.

According to authors of [5, 6, 7], text summarization is an established research direction in NLP with growing practical importance given the ever-increasing volume of text that would benefit from summarization.

One can formulate text summarization as an unstructured sequence-to-sequence problem, where a document (e. g. a CNN news article) is the input and the LM is tasked with generating a summary that resembles the reference summary (e. g. the bullet point summary provided by CNN with their article).

To evaluate model performance, the model-generated summary is compared against a human-authored reference summary using automated metrics for overall quality as asserted by authors of [8, 9, 10, 11]. Extractiveness refers to the extent to which model summaries involve copying from the input document.

Consequently, it is important to measure and improve the faithfulness of these systems since unfaithful systems may be harmful by potentially spreading misinformation, including dangerous, yet hard-to-detect errors, when deployed in real-world settings.

Sentiment analysis has blossomed into its own subarea in the field with many works broadening and deepening the study of sentiment from its initial binary text-classification framing according to authors of [12].

Text classification has a long history in NLP – as claimed by Authors of [13, 14]with tasks such as language identification, sentiment analysis, topic classification, and toxicity detection being some of the most prominent tasks within this family.

Focusing on fairness of models is essential to ensuring technology plays a positive role in social change according to authors of [15, 16]. Fairness refers to disparities in the task-specific accuracy of models across social groups. One way to operationalize fairness is by means of counterfactual fairness which refers to model behavior on counterfactual data that is generated by perturbing existing test examples as mentioned by authors of [17].

In contrast, bias refers to properties of model generations, i.e. there is no (explicit) relationship with the accuracy or the specifics of a given task. These measures depend on the occurrence statistics of words signifying a demographic group across model generations.

Toxicity detection (and the related tasks of hate speech and abusive language detection) is the task of identifying when input data contains toxic content, which originated due to the need for content moderation on the Internet as mentioned by Authors of [18, 19].

Critiques of the task have noted that:

1) the study of toxicity is overly reductive and divorced from use cases;
2) standard datasets often lack sufficient context to make reliable judgments; and
3) the construct of toxicity depends on the annotator as mentioned by authors of [20, 21].

Another crucial concept for ML models is toxicity used as an umbrella term for related concepts like hate speech, violent speech, and abusive language as claimed by Authors of [22]. To operationalize toxicity measurement, one can use the Perspective API to detect toxic content in model generations.
Given these features of LM, the next section explores a conceptual framework for designing a LM benchmark.

3. Results and Discussion

To carry out the research, the study suggests implementing the following aspects for designing a LM benchmark (Fig. 2):

1. *Taxonomy.*One can taxonomize the vast design space of language model evaluation into scenarios and metrics. By stating this taxonomy, one can select systematically from this space, which makes explicit both priorities in benchmark design and the limitations in the benchmark at present.
2. *Broad coverage.* Given the taxonomy, one select and implement core scenarios, for which one can comprehensively measure major metrics (accuracy, calibration, robustness, fairness, bias, toxicity, efficiency).
3. *Evaluation of existing models.* One can evaluate existing LMS under the standardized conditions of the benchmark, ensuring models can now be directly compared across many scenarios and metrics. These models might vary in terms of their public accessibility: while some of them are open, others are limited-access, and a few might even be closed.

4. *Empirical findings*.
The extensive evaluation will offer guidance for future language model development and ample opportunities for further analysis.

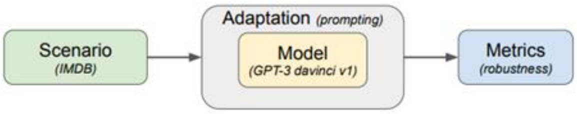


Fig. 2 . Suggested LM Benchmark Process

As seen in Fig.2, the following aspects (scenario, adaptation, metric) are required to evaluate a LMto provide a roadmap for how to evaluate language models:

– *Scenarios*. A scenario instantiates a desired use case for a LM. Scenarios are what we want models to do. Each instance consists of:
1) an input (a string); and
2) a list of references.

Each reference is a string annotated with properties relevant for evaluation (e. g. is it correct or acceptable?).
– *Adaptation.* Adaptation is the procedure that transforms a LM, along with training instances, into a system that can make predictions on new instances. Examples of adaptation procedures include prompting, lightweight-finetuning, and finetuning.

We define a language model to be a black box that takes as input a prompt (string), along with decoding parameters (e. g. temperature). The model outputs a completion (string), along with log probabilities of the prompt and completion. Viewing language models as text-to-text abstractions is important for two reasons:

First, while the prototypical LM is usually a dense Transformer trained on raw text, LMs could also use an external document store, issue search queries on the web or be trained on human preferences as claimed by Authors of [21]. An ideal model should be agnostic with regard to these implementation details.

Second, the text-to-text abstraction is a convenient general interface that can capture all the (text-only) tasks of interest, an idea that was pioneered by Authors of [24].

– *Metrics.* To determine how well the model performs, one can compute metrics over these completions and probabilities. Metrics concretely operationalize the abstract desiderata required for useful systems.

To evaluate a LM, a series of runs must be implemented, where each run is defined by a scenario, adaptation method, and metric. Each of these scenarios, adaptation, and metrics define a complicated and structured space, which one implicitly navigates to make decisions in evaluating a LM.

One can taxonomize scenarios based on the following:
1) a task (e. g. question answering, summarization), which characterizes what we want a system to do;
2) a domain (e. g. a Wikipedia 2018 dump), which characterizes the type of data we want the system to do well on; and
3) the language or language variety (e. g. English).

Tasks, domains, and languages are not atomic or unambiguous constructs: they can be made coarser and finer. Given this structure, one can deliberately select scenarios based on main overarching principles:

1) coverage of the space,
2) minimality of the set of selected scenarios, and
3) prioritizing scenarios that correspond to user-facing tasks.

Given the ubiquity of natural language, the field of natural language processing (NLP) considers myriad tasks that correspond to language's many functions. To generate this set, one can take the tracks at a major NLP conference [21], and for each track, one can map the associated subarea of NLP to canonical tasks for that track.

Moreover, domains are a familiar construct in NLP, yet their imprecision complicates systematic coverage of domains. One can further decompose domains according to 3 W's:

1) *What (genre).* The type of text, which captures subject and register differences. Examples: Wikipedia, social media, news, scientific papers, fiction.
2) *When (time period)*. When the text was created.
Examples: 1980s, pre-Internet, present-day (e.g. does it cover very recent data?)
3) *Who (demographic group)*. Who generated the data or who the data is about. Examples: Black/White, men/women, children/elderly.

3.1. Models

When deployed in practice, models are confronted with the complexities of the open world (e. g. typos) that cause most current systems to significantly degrade as mentioned by authors of [23].

One suggestion is to measure the robustness of different models by evaluating them on transformations of an instance. That is, given a set of transformations for a given instance, one can measure the worst-case performance of a model across these transformations.

On the one hand, measuring robustness to distribution or subpopulation shift requires scenarios with special structure (i. e., explicit domain/subpopulation annotations) as well as information about the training data of the models.

On the other hand, measuring adversarial robustness requires many adaptive queries to the model in order to approximate worst-case perturbations, which might not always be feasible [21].

Moreover, the transformation/perturbation-based paradigm has been widely explored to study model robustness in order to understand whether corruptions that arise in real use cases (e. g. typos) affect the performance of the model significantly. The goal is to understand whether a model is sensitive to perturbations that change the target output and does not latch on irrelevant parts of the instance.

3.2. Metrics

To taxonomize the space of desiderata, one can begin by enumerating criteria that are necessary for developing useful systems. Yet, what does it mean for a system to be useful?

Too often in AI, this has come to mean the system should be accurate in an average sense. While (average) accuracy is an important, and often necessary, property for a system, accuracy is often not sufficient for a system to be useful/ desirable.

Unfortunately, while many of the desiderata are well-studied by the NLP community, some are not codified in specific tracks/areas (e. g. uncertainty and calibration). Therefore, it is suggested to expand the scope to all AI conferences, drawing from a list of AI conference deadlines.

3.3. Discussion

For the reproducibility of the results, one can use explicit rule induction and implicit function regression, which corresponds to making and applying claims about the likely causal structure for observations.

For rule induction, one can design and implement rule_induct inspired by the LIME induction tasks, where we provide two examples generated from the same rule string, and task the model with inferring the underlying rule.

For function regression, one can design and implement numeracy_prediction, which requires the model to perform symbolic regression given a few examples and apply the number relationship (e. g. linear) to a new input.

The study is limited to some extent as in order to distinguish reasoning from language and knowledge as much as possible, one can focus on relatively abstract capacities necessary for sophisticated text-based or symbolic reasoning.

For future studies, one can also evaluate language models on more complex and realistic reasoning tasks that require multiple primitive reasoning skills to bridge the gap between understanding reasoning in very controlled and synthetic conditions and the type of reasoning required in practical contexts.

4. Conclusions

The study helped to solve the thorny problem of benchmark development for LM by designing the main features of benchmarks – scenario, adaptation, metric- required to provide a roadmap for how to evaluate LMs. It also made recommendations of how to use model and metrics for fairness and transparency when it comes to developing LM.

From the quantitative perspective, the benchmark model includes the following aspects:

1) metrics to define the design space for LM evaluation;
2) metrics that were selected to prioritize coverage (e. g. different English varieties);
3) value (e. g. user-facing applications); and
4) feasibility (e. g. limited engineering resources).

Given the lack of studies in the field, it is a step towards the design of more sophisticated models and thus, right now, far from perfect. Nevertheless, it aims to raise awareness of the importance of developing benchmarks for AI models.

Conflict of interest

The author declares that she has no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

References

1. Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020.
2. Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres (2019) Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.
3. Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher (2016) Ask me anything: Dynamic memory networks for natural language processing. In International conference on machine learning, pages 1378–1387. PMLR.
4. AtoosaKasirzadeh and Iason Gabriel (2022). In conversation with artificial intelligence: aligning language models with human values.
5. Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster (2021) Reduced, reused and recycled: The life of a dataset in machine learning research. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

6. Christopher Potts, and MateiZaharia (2021) A moderate proposal for radically better AI-powered Web search. Stanford HAI Blog.
7. Daniel Khashabi, Yeganeh Kordi, and HannanehHajishirzi (2022) Unifiedqa-v2: Stronger generalization via broader cross-format training. ArXiv, abs/2202.12359. Omar Khattab,

8. Davis, D., Seaton, D., Hauff, C., &Houben, G. J. (2018, June). Toward large-scale learning design: Categorizing course designs in service of supporting learning outcomes. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1-10). https://doi.org/10.1145/3231644.3231663

9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

10. Divyansh Kaushik, Eduard Hovy, and Zachary Lipton (2019) Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations (ICLR).
11. FereshteKhani and Percy Liang. (2020) Feature noise induces loss discrepancy across groups. In International Conference on Machine Learning (ICML).

12. González-Carvajal, S., & Garrido-Merchán, E. C. (2020). *Comparing BERT against traditional machine learning text classification*. arXiv preprint arXiv:2005.13012. https://arxiv.org/abs/2005.13012

13. Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: An overview*. arXiv preprint arXiv:2008.05756. https://arxiv.org/abs/2008.05756

14. Hannah Rose Kirk, AbebaBirhane, Bertie Vidgen, and Leon Derczynski (2022) Handling and presenting harmful text in nlp research. Andy Kirkpatrick. 2020. The Routledge handbook of world Englishes. Routledge.
15. Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. ArXiv, abs/2001.08361.

16. Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980

17. Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva (2017) Counterfactual fairness. In Advances in Neural Information Processing Systems (NeurIPS), pages 4069–4079.
18. Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst (2022) SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. Transactions of the Association for Computational Linguistics, 10:163–177.
19. Tom Kwiatkowski, JennimariaPalomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, IlliaPolosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. In Association for Computational Linguistics (ACL).
20. TomšKočisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2017. The NarrativeQA reading comprehension challenge. arXiv preprint arXiv:1712.07040.

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., &Polosukhin, i. (2017). Attention is all you need. *31st Conference on Neural information Processing Systems* (NiPS 2017; pp. 5998-6008). Long Beach, USA. https://arxiv.org/abs/1706.03762