Holistic Evaluation of Large Language Models: Assessing Robustness, Accuracy, and Toxicity for Real-World Applications

David Cecchini[1], Kalyan Chakravarthy[1], Prikshit Sharma[1], Rakshit Khajuria[1], Arshaan Nazir[1], Veysel Kocaman[1], David Talby[1],

[1]John Snow Labs,

Correspondence: cecchini@johnsnowlabs.com

Abstract

Large Language Models (LLMs) have been widely used in real-world applications. However, as LLMs evolve and new datasets are released, it becomes crucial to build processes to evaluate and control the models' performance. In this paper, we describe how to add Robustness, Accuracy, and Toxicity scores to model comparison tables, or leaderboards. We discuss the evaluation metrics, the approaches considered, and present the results of the first evaluation round for model Robustness, Accuracy, and Toxicity scores. Our results show that *GPT 4* achieves top performance on robustness and accuracy test, while *Llama 2* achieves top performance on the toxicity test. We note that newer open-source models such as *open chat 3.5* and *neural chat 7B* can perform well on these three test categories. Finally, domain-specific tests and models are also planned to be added to the leaderboard to allow for a more detailed evaluation of models in specific areas such as healthcare, legal, and finance.

1 Introduction

With the release of Large Language Models (LLM) that demonstrate human-like performance on a variety of natural language understanding tasks, it becomes crucial to build processes to evaluate and control the models' performance on real-world applications. Apart from quantitative metrics such as accuracy, BLEU (Papineni et al., 2002; Lin and Och, 2004), and Rouge scores (Lin, 2004), it is also important to validate other aspects such as Robustness, Bias, Fairness, Toxicity, Representation, among others. In this paper, we describe how to use the open-source toolkit *LangTest* (Nazir et al., 2024) to add scores from those aspects into LLM leaderboards. We discuss the evaluation metrics and approaches used and present the results of the first evaluation round for model Robustness, Accuracy, and Toxicity[1].

LangTest is an open-source Python toolkit for testing and evaluating LLMs and classical Natural Language Processing (NLP) model architectures such as Named Entity Recognition (NER) and Text Classification. Its primary focus is to ensure that these models are robust, unbiased, accurate, non-toxic, fair, efficient, clinically relevant, secure, free from disinformation and political biases, sensitive, factual, legally compliant, and less vulnerable before they are deployed in real-world applications. Other features of the toolkit include the capability to run tests either as Command Line Interface (CLI) or as a Python library in one-liners, tailor made tests for the healthcare domain (to be included in the second round of evaluations), data augmentation for mitigating weaknesses of the models, and support for running tests on dedicated servers or locally.

To illustrate the importance of holistic model evaluation, we designed a new leaderboard to compare not only accuracy, but also other facets that are important to real-world applications such as robustness to perturbations in the text, and toxicity of the generated text. The leaderboard is based on the *LangTest* toolkit, and we present the results of the first evaluation round for model Robustness, Accuracy, and Toxicity. We hope that this toolkit can be a valuable resource for researchers, developers, and practitioners to understand the strengths and weaknesses of the models, and to make informed decisions on which model to use for specific tasks.

The rest of the paper is organized as follows. In Section 2, we discuss the motivation behind the development of the *LangTest* toolkit and the Leaderboard. In Section 3, we describe the tests and metrics present in the *LangTest* Leaderboard. In Section 4, we present the results of the first

[1]Available at https://langtest.org/leaderboard/llm

Proceedings of the 4th Workshop on Trustworthy NLP (TrustNLP 2024), pages 109–117

evaluation round for model Robustness, Accuracy, and Toxicity. Lastly, in Section 5, we conclude the paper and discuss future work.

2 Motivation

Recent research has shown great advances on evaluation metrics for LLM models, such as BLEU, ROUGE, and Word Error Rate (WER) (Jothilakshmi and Gudivada, 2016). Although these accuracy metrics are important to evaluate the model performance on specific tasks such as text classification, information extraction, or summarization, they do not provide a complete picture of the model's performance, especially in domain specific areas such as healthcare (Schwartz et al., 2023; Singhal et al., 2023; Wang et al., 2023), legal (Sun, 2023; Fei et al., 2023), or finance (Xie et al., 2023; Li et al., 2023; Wang et al., 2023).

Our motivation to develop the *LangTest* toolkit comes from the need to provide a more holistic evaluation of LLM models, including aspects such as Robustness, Bias, Fairness, etc., inspired by the previous research by (Ribeiro et al., 2020), (Song and Raghunathan, 2020), (Van Aken et al., 2021), (Dhole et al., 2021), (Liang et al., 2023), (Wang et al., 2023), (Sun et al., 2024) and others, and to address domain-specific needs that needs further consideration for LLM evaluation.

While these studies contain many evaluation approaches and metrics for language models, they are often based on static datasets that represent a good picture of the state of the models at the time of the study or designed to evaluate specific models (e.g., GPT 3.5 or GPT 4). However, as the models evolve and new datasets are released, it is important to have a dynamic evaluation framework that can be updated with new datasets, models, and tests. For example, while (Liang et al., 2023) contributed to a development of holistic evaluation of models using multiple metrics, their approach is based on static datasets and does not provide a dynamic framework to add new tests and metrics. Similarly, (Wang et al., 2023) developed new datasets and standardized prompts and metrics to evaluate models on six categories (truthfulness, safety, fairness, robustness, privacy, and machine ethics) which contributed to a better evaluation framework for LLMs, but researchers and practitioners are not incentivized to make changes the framework to address specific needs and concerns. Another recent development on holistic evaluation of LLMs is the work done by (Sun et al., 2024) which defined a taxonomy of aspects to be evaluated on models with eight categories: truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability, but their approach was designed to evaluate GPT models only.

Other toolkits are available to evaluate models such as the *lm-evaluation-harness* by *EleutherAI*[2], which offers the community a comprehensive and flexible framework. It was primarily designed for assessing the accuracy and performance of models (e.g., through comparisons on the *Open LLM Leaderboard*[3] by *HuggingFace*), yet it still lacks a thorough evaluation of models in other areas such as robustness, bias, fairness, and toxicity.

To address these issues, *LangTest* provides not only benchmark datasets and tests, but also a framework to dynamically add perturbations to the dataset to create new tests for model evaluation. It is a flexible toolkit where researchers and practitioners can define their evaluation criteria based on existing datasets or develop new ones either by modifying existing datasets or designing new ones specific to their use cases. As new techniques are developed to add perturbations and modification in the input data, the toolkit can provide an ever-growing set of tests and procedures to evaluate the models. Apart from evaluating the models, other features of the toolkit are to provide data augmentation techniques to mitigate weaknesses of the models, and to support running tests on dedicated servers or locally. These features empower users to not only have a static evaluation score of models, but also to address the evaluation as a continuous process.

In addition, domain specific evaluation is also critical, as models are often used in specific areas such as healthcare, legal, or finance that have specific requirements for the models. We manually curated datasets for these areas and have a dedicated team to continue researching and curating new datasets and tests that can be used to verify models' performance for healthcare, legal, and finance. Our approach aims to provide base datasets and tests for these areas as a starting point as better curated evaluation datasets are still scarce in the literature.

In illustrating the significance of a holistic model

[2] https://github.com/EleutherAI/lm-evaluation-harness

[3] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

evaluation, we introduce the *LangTest* Leaderboard. This platform facilitates comparisons of various models across specific tasks and benchmark datasets, employing tests and metrics from the *LangTest* toolkit. Leaderboards and benchmark comparisons serve as tools to aid stakeholders in comprehending the strengths and weaknesses of models, enabling informed decisions regarding their suitability for specific tasks. We anticipate that the *LangTest* Leaderboard will emerge as a valuable resource for the community.

3 LangTest Leaderboard

In this section we describe the tests and metrics present in the *LangTest* Leaderboard. For the initial version of the leaderboard, we added three categories of tests: Robustness, Accuracy, and Toxicity. Other categories already supported by *LangTest* will be added in future releases of the leaderboard, including domain specific scores for healthcare.

3.1 Benchmark Datasets and Models

We used a diverse set of benchmark datasets, each with its own characteristics and challenges, to evaluate the models on the Robustness, Accuracy, and Toxicity tests. The datasets used in the first evaluation round are described below.

- **RealToxicityPrompts** (Gehman et al., 2020) - We used the toxic user prompt subset designed by (Wang et al., 2023) containing 1200 examples.
- **MMLU** (Hendrycks et al., 2021) - Curated version of the MMLU dataset which contains the clinical subsets (college biology, college medicine, medical genetics, human aging, professional medicine, and nutrition).
- **BoolQ** (Clark et al., 2019) - Test set containing 3245 unlabeled examples (robustness) and dev set containing 3270 labeled examples (accuracy).
- **TruthfulQA** (Raj et al., 2022) - Test set containing 164 question and answer examples.
- **MedMCQA** (Pal et al., 2022) - We used test (robustness) and validation (accuracy) sets from the dataset with all splits (Anatomy, Dental, Microbiology, etc.).
- **MedQA** (Jin et al., 2020) - Test set containing 1273 question and answers examples.
- **Bigbench** (Ghazal et al., 2013) - We used the test set with the following subsets: abstract narrative understanding, causal judgment, and disambiguation QA.
- **Consumer Contracts** (Kolt, 2022) - Test set from the Consumer-Contracts dataset, containing 396 samples.
- **SocialIQA** (Sap et al., 2019) - Test set containing 1954 question and answer examples.
- **ContractQA** (Guha et al., 2023) - Test set from the Contracts dataset, containing 80 samples.
- **CommonsenseQA** (Talmor et al., 2019) - Test set containing 1140 questions (robustness) and validation set containing 1221 question and answer examples (accuracy).
- **BBQ** (Parrish et al., 2021) - We used the test set containing 1012 question and answers examples.
- **LogiQA** (Liu et al., 2020) - Test set containing 1000 question and answers examples.
- **PIQA** (Bisk et al., 2019) - Test set containing 1500 questions (robustness) and validation set containing 1500 question and answer examples (accuracy).
- **ASDiv** (Miao et al., 2021) - We used the test set containing 2305 question and answers and examples.
- **PubMedQA** (Jin et al., 2019) - We used truncated 500 examples from the *pqaa_artificial* and *pqa_labeled* subsets.
- **OpenBookQA** (Mihaylov et al., 2018) - Test set containing 500 multiple-choice elementary level science questions.

As for the models, we evaluated the most relevant models in the field of LLMs, including *GPT 3.5*, *GPT 4*, *Llama 2 7B*, among others. The selection criteria were made to include models that are widely used in the community, and that have been shown to have good performance on a variety of tasks. We also included models that are quantized, as quantization is an important technique to reduce the memory footprint of the models, and to make them more efficient for deployment in real-world applications. While we understand that there are

other models that could be included in the evaluation, we believe that the models selected provide a good representation of the state-of-the-art in LLMs, and additional result for other models can be added in future releases of the leaderboard.

3.2 Robustness Evaluation

To evaluate robustness, we propose a set of tests that can apply perturbations to the input text and measure if the models' prediction is unchanged. Below we describe the different tests available and their description.

- uppercase - Apply upper casing to the input text.
- lowercase - Apply lower casing to the input text.
- titlecase - Apply title casing to the input text.
- add_type - Add common typo to the input text based on a typo frequency dictionary fo English.
- dyslexia_word_swap - Dyslexia Word Swap dictionary is employed to apply the most common word swap errors found in dyslexic writing to the input data.
- add_abbreviation - Abbreviates words on the input text based on commonly used abbreviations on social media platforms and generic abbreviations for English.
- add_slangs - Substitutes certain words (specifically nouns, adjectives, and adverbs) in the original text with their corresponding slang terms.
- add_speech_to_text_typo - Replaces words in the text by common typos resulting from speech-to-text process.
- add_ocr_typo - Replaces words in the text by common typos resulting from OCR process.
- adjective_synonym_swap - Replaces adjectives in the text by their synonyms.

The robustness tests aim to measure how well the models can perform with small modifications to the input data. We expect that the model prediction does not change when the input data is perturbed, and that the model can generalize well to unseen data. The tests are designed to measure the model's performance on different types of perturbations, and to provide a comprehensive evaluation of the model's robustness. Future work will include additional tests and perturbations to the input data to further evaluate the models' performance, including changes in grammar, punctuation, and sentence structure.

3.3 Accuracy Evaluation

In our leaderboard for LLM performance, we also support common accuracy metrics, allowing the community to compare the performance of different models on these metrics. These metrics are widely used in the NLP community, and provide a good measure of the models' performance on specific tasks. To allow practitioners to fully compare the models, we include the following metrics in the leaderboard:

- min_exact_match_score - Evaluates models based on a minimum threshold value for pass/fail.
- min_rouge1_score - Uses a threshold to test if the model passes or fails based on Rouge score for unigrams.
- min_rouge2_score - Uses a threshold to test if the model passes or fails based on Rouge score for bigrams.
- min_rougeL_score - Uses a threshold to test if the model passes or fails based on Rouge score for longest common sub-sequence (LCS).
- min_rougeLsum_score - Uses a threshold to test if the model passes or fails based on Rouge score for longest common subsequence per sentence (union-LCS).
- min_bleu_score - Uses a threshold to test if the model passes or fails based on smoothBLEU score.
- llm_eval - Evaluates the model based on a set of predefined prompts and expected completions run on a given LLM (e.g., GPT 3.5 turbo).

Depending on the NLP task, other accuracy measures can be used (e.g., f1 score, precision, and recall for classification models). As we continue to develop the toolkit, we plan to add more accuracy metrics to the leaderboard, and to provide a more comprehensive evaluation of the models' performance on specific tasks.

3.4 Toxicity Evaluation

Another important aspect to evaluate in LLM models is the toxicity of the generated text. We measure the generated text for presence of toxic topics such as obscenity, insult, threat, identity attack, sexual orientation attack, and psychiatric or mental illness attack. We aggregated the tests into a single metric that determine a score representing how often the models avoid generating toxic text. The sentences are carefully designed to have an open ending that is leading to harmful/toxic content, and the model is expected to complete the sentence in a way that is not harmful/toxic or denies generating such content.

We use a text classification model *unbiased-toxic-roberta* (Hanu and Unitary team, 2020) to identify toxicity of completions (using label "toxicity"), and for each example the model either passes or fails the test depending upon its toxicity score. Each example is evaluated against a predefined threshold; it passes the test if the toxicity score is below the threshold, and fails if the score exceeds it. The final score is calculated as the percentage of examples that the model passed the test.

4 Results

In this section we present the results of the first evaluation round for model Robustness, Accuracy, and Toxicity. The outputs were obtained by running the experiments on a single machine, and quantization was used when needed (identified with a start on the models' name) using *LM Studio*[4]. We average the scores of individual tests of each category to obtain an overall score for them, but the specific scores can be found on the official leaderboard website. These average scores are presented in tables 1, 2, and 3.

From table 1, we can see that *GPT 4* is the top performer, with *DeciLM 7B*, *Mistral 7B*, *Mixtral 8x7B*, *neural chat 7B*, and *flan t5 xxl* tied with average score of 0.88. The models *Llama 2 7B*, *GPT 3.5*, and *phi 2* have the worst performance on the robustness tests, with *phi 2* having the worst performance on most of the datasets.

It is notable that models with number of parameters from $7B$ to $11B$ can outperform *GPT 3.5* ($175B$) on the robustness tests, which shows that the number of parameters is not the only factor that determines the model's performance.

From table 2, we can see that the models *GPT 4*, *GPT 3.5* and *open chat 3.5* have the best performance on the accuracy tests, with *GPT 4* having the best performance on most of the datasets. The models *phi 2*, *Llama 2 7B*, and *flan t5 xxl* have the worst performance on the accuracy tests, with *flan t5 xxl* having the worst performance on the majority of the datasets but achieving top score in a few ones (*PubMedQA* and *BoolQ*). Although *GPT 4* obtained top performance in the leaderboard, it is important to consider that the size of this model is much larger than the other models, and it is remarkable to achieve fairly good results with smaller models (e.g., *open chat 3.5* with 7B parameters) or mixture of smaller models (e.g., *Mixtral 8x7B*) (Fedus et al., 2022).

Worth mentioning is the difference in the scores from the accuracy table with the ones obtained in the robustness table. The scores for robustness measure the capability of the model to make the same prediction when the input is perturbed, while the accuracy scores measure the capability of the model to make the correct prediction. This means that a model can be inaccurate but robust, or accurate but not robust.

Finally, from table 3, we can see that the model *Llama 2 7B* has the best performance on the toxicity tests, as the outputted text filtered the toxicity present in the prompt or refused to continue the toxic sentences in most of the examples. The models *Mistral 7B*, *Mixtral 8x7B*, and *GPT 3.5* have the worst performance in toxicity tests, meaning that these models generate toxic texts when prompted/suggested to.

Overall, the results show that the *GPT* family of models achieve high performance on robustness and accuracy tests and that the newest version of the family, GPT 4, improved the previous GPT 3.5 on the toxicity generation. In the other hand, *Mixtral 8x7B* can perform well on accuracy and robustness but propagate toxicity in the prompts. *Llama 2* performance on the accuracy and robustness tests was below average, although it was the top performer in the toxicity. These results are consistent with other studies and leaderboards, but it is important to note that the results may vary depending on the dataset and the test used. Furthermore, some applications may be directly impacted by specific tests (e.g., typos coming from OCR or Speech2Text models) while other tests would not be as relevant. To analyze these scenarios, in the official leaderboard website is possible to add filters and select which

[4] https://lmstudio.ai/

Dataset	GPT 3.5	GPT 4	Mixtral 8x7B	flan t5 xxl	Mistral 7B	phi 2*	neural chat 7B*	SOLAR 10.7B*	Llama 2 7B*	open chat 3.5*	DeciLM 7B
ASDiV	0.80	**0.88**	0.78	0.76	0.74	0.65	0.68	0.66	0.68	0.71	0.79
BBQ	0.82	**0.97**	0.88	0.92	0.88	0.77	0.92	0.90	0.89	0.87	
Bigbench	0.83	0.91	0.85	**0.93**	0.86	0.82	0.91	0.87	0.85	0.85	0.90
BoolQ	0.79	**0.96**	0.93	0.94	0.91	0.84	0.93	0.93	0.83	0.91	0.93
CommonsenseQA	0.87	0.90	0.87	**0.91**	0.87	0.71	0.88	0.85	0.83	0.85	0.85
Consumer-Contracts	0.79	**0.98**	0.94	0.96	1.00	0.78	0.92	0.93	0.92	0.85	0.92
Contracts	0.96	0.97	0.98	0.97	**0.99**	0.80	0.90	0.95	0.94	0.97	0.96
LogiQA	0.74	0.87	0.82	**0.96**	0.89	0.72	0.88	0.84	0.85	0.80	
MedMCQA	0.74	**0.90**	0.86	0.85	0.87	0.76	0.86	0.83	0.79	0.82	0.86
MedQA	0.81	**0.93**	0.91	0.88	0.90	0.69	0.90	0.87		0.85	0.89
MMLU	0.87	**0.95**	0.90	0.92	0.89	0.74	0.92	0.90	0.85	0.87	0.92
OpenBookQA	0.87	**0.92**	0.89	0.90	0.89	0.79	0.88	0.86	0.83	0.88	
PIQA	0.93	**0.97**	0.96	0.95	0.96	0.92	0.95	0.94	0.89	0.96	0.96
PubMedQA	0.78	0.96	0.95	0.97	0.92	0.83	0.95	0.93	**0.98**	0.95	0.97
SIQA	0.84	0.87	0.92	**0.93**	0.89	0.84	0.89	0.90	0.89	0.90	0.89
TruthfulQA	0.88	**0.96**	0.89	0.57	0.89						
Average	0.79	**0.91**	0.88	0.88	0.88	0.77	0.88	0.86	0.83	0.84	0.88

Table 1: Robustness results for different models on the benchmark datasets. Models marked with * are quantized.

tests to consider for each category, allowing users to understand the full capabilities of the models for their specific use case.

Notable is the performance of new open-source models such as *Open Chat 3.5* and *Neural Chat 7B* both with seven billion parameters. They achieved good performance on the accuracy and robustness tests, and the toxicity tests showed that they can generate fewer toxic texts than, e.g., *GPT 3.5*. This shows that smaller models can achieve good performance on a variety of tasks, and that the number of parameters is not the only factor that determines the model's performance. These models were released under Apache 2.0 license, allowing for the community to use and modify them for their specific use cases.

5 Conclusion and Future Work

We introduced a holistic evaluation of LLMs toolkit that includes scores for robustness, accuracy, ad toxicity in the generated texts. Our results are available on the *LangTest* Leaderboard, a platform that compare different models on specific tasks and benchmark datasets using the tests and metrics present in the *LangTest* toolkit.

We identified that LLM can achieve remarkable performance when measured by accuracy metrics, but a holistic evaluation is needed when considering robustness and toxicity. The results show that the *GPT* family of models achieve high performance on robustness and accuracy tests, but GPT 3.5 propagates more often the toxicity in the prompts than GPT 4, while in general the models *Mistral 7B* and *Mixtral 8x7B* can perform well on accuracy and robustness but perform worse on toxicity test. The model *Llama 2 7B* has the best performance on the toxicity tests, but its performance on the accuracy and robustness tests was below average. Open-source models such as *Open Chat 3.5* and *Neural Chat 7B* achieved good performance on the accuracy and robustness tests, and the toxicity tests showed that they can generate fewer toxic texts than *GPT 3.5*.

In future works, we aim to keep adding new categories, datasets, tests, and models to the tables, allowing for a more comprehensive evaluation of LLM models. Finally, domain-specific tests and models are also planned to be added to the leaderboard, allowing for a more detailed evaluation of models in specific areas such as healthcare, legal, and finance.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Dataset	DeciLM 7B	flan t5 xxl	GPT 3.5	GPT 4	Llama 2 7B*	Mistral 7B	Mixtral 8x7B	neural chat 7B*	open chat 3.5*	phi 2*	SOLAR 10.7B*
ASDiV	0.28	0.19	0.35	**0.48**	0.23	0.32	0.33	0.23	0.25	0.29	0.26
BBQ		0.08	0.40	**0.58**	0.18	0.13	0.35	0.56	0.50	0.35	0.50
Bigbench	0.54	0.24	0.58	**0.66**	0.39	0.46	0.54	0.59	0.61	0.47	0.46
BoolQ		**0.64**	0.57	0.63	0.55	0.58	0.61	0.61	0.63	0.56	0.62
CommonsenseQA	0.72	0.27	0.77	0.72	0.44	0.67	0.70	**0.74**	0.82	0.54	0.69
Consumer-Contracts		0.66	0.55	**0.67**	0.39		0.46	0.55	0.48	0.54	0.60
Contracts	0.68	0.69	**0.70**	0.67	0.42	0.31	0.65	**0.70**	0.69	0.52	0.68
LogiQA		0.11	**0.52**	0.32	0.24	0.47	0.41	0.43	0.50	0.42	0.41
MedMCQA	0.42	0.14	0.59	**0.72**	0.33	0.44	0.57	0.45	0.51	0.31	0.40
MedQA	0.37	0.11	0.24	0.48		0.31	0.41	0.42	**0.49**	0.23	0.34
MMLU	0.65	0.15	0.77	**0.78**	0.41	0.48	0.67	0.65	0.66	0.48	0.39
OpenBookQA		0.22	0.81	0.81	0.50	0.64	0.80	0.75	**0.88**	0.60	0.74
PIQA	0.90	0.18	0.90	**0.93**	0.65	0.85	0.81	0.79	0.87	0.78	0.34
PubMedQA	0.53	**0.60**	0.36	0.50	0.44	0.48	0.48	0.46	0.58	0.43	0.48
SIQA	**0.79**	0.18	0.73	0.61	0.52	0.68	0.41	0.74	0.78	0.64	0.71
TruthfulQA		0.26	**0.30**	0.26		0.29	0.27				
Average	0.48	0.22	0.57	**0.67**	0.37	0.46	0.55	0.51	0.56	0.39	0.45

Table 2: Accuracy results for different models on the benchmark datasets. Models marked with * are quantized.

Model	Toxicity Score
GPT 3.5	0.54
GPT 4	0.88
Llama 2 7B*	**0.98**
Mistral 7B	0.39
Mixtral 8x7B	0.42
NeuralBeagle 14 7B*	0.83
neural chat 7B*	0.83
open chat 3.5*	0.91
phi 2*	0.73
SOLAR 10.7B*	0.8
zephyr 7B*	0.8

Table 3: Toxicity results for different models. Models marked with * are quantized.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Preprint*, arXiv:2101.03961.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *Preprint*, arXiv:2309.16289.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno 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.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language

understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

S. Jothilakshmi and V.N. Gudivada. 2016. Chapter 10 - large scale data enabled evolution of spoken language research and applications. In Venkat N. Gudivada, Vijay V. Raghavan, Venu Govindaraju, and C.R. Rao, editors, *Cognitive Computing: Theory and Applications*, volume 35 of *Handbook of Statistics*, pages 301–340. Elsevier.

Noam Kolt. 2022. Predicting consumer contracts. *Berkeley Tech. LJ*, 37:71.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA. Association for Computing Machinery.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers. *arXiv preprint arXiv:2106.15772*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Arshaan Nazir, Thadaka Kalyan Chakravarthy, David Cecchini, Rakshit Khajuria, Prikshit Sharma, Ali Tarik Mirik, Veysel Kocaman, and David Talby. 2024. Langtest: A comprehensive evaluation library for custom llm and nlp models. *Software Impacts*, 19:100619.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. *arXiv preprint arXiv:2211.05853*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *Preprint*, arXiv:1904.09728.

Ilan S Schwartz, Katherine E Link, Roxana Daneshjou, and Nicolás Cortés-Penfield. 2023. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. *Clinical Infectious Diseases*, page ciad633.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *Preprint*, arXiv:2303.09136.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Betty Van Aken, Sebastian Herrmann, and Alexander Löser. 2021. What do you see in this patient? behavioral testing of clinical nlp models. *arXiv preprint arXiv:2111.15512*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484. Curran Associates, Inc.