DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

ORNL/DRD-2024

Oak Ridge National Laboratory

Zheming Jin

September 2024

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/DRD-2024/

Computer Science and Mathematics Division

SCALABLE WORKFLOW FOR EVALUATING TRUSTWORTHINESS OF LARGE LANGUAGE MODELS

September 2024

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

ABSTRACT

This work describes the improved workflow for evaluating open-source large language models (LLMs) for trustworthiness. The workflow facilitates the acquisition of LLMs, the generation of LLM responses, and the evaluation of the responses for their trustworthiness. As a use case, the workflow is employed to evaluate dense, quantized, and pruned Meta Llama3.1 LLMs for their truthfulness. The outcome of the project could set the stage for understanding and developing trustworthy models in the future projects.

1. INTRODUCTION

Large language models (LLMs) are revolutionizing the field of natural language processing, providing unprecedented capabilities in text generation, comprehension, and interaction [1]. They are cutting-edge artificial intelligence (AI) systems that can handle complex language tasks with approximate human-level performance. Hence, LLMs have increasingly been applied in domains such as education [2], law [3], and medicine [4]. As LLMs are deployed across increasingly diverse domains, concerns are growing about their trustworthiness: LLMs' emerging generative capabilities may generate inaccurate or misleading outputs [5, 6, 7], which could be exploited for propagation of false information and automated cyberattacks [8, 9, 10]. There are potential biases (e.g., gender and culture) and sensitive information in training datasets, which could cause fairness of contents generated by LLMs and privacy breach [11, 12]. LLMs' responses do not necessarily align with human values [13], and such conflicts and contradictions could impact their broad application across different domains.

Evaluating LLMs will mitigate the risks of misinformation and minimize generation of biased, harmful, unethical, unsafe contents. It is important to identify vulnerabilities and implement safeguards against them. Benchmarks play a critical role in evaluating LLMs' trustworthiness. As the implementation details of a model often govern model performance, the lack of necessary code and sufficient detail of a model makes it difficult to evaluate the model releases objectively and effectively [14]. To address the concerns, there are benchmarks on LLMs' toxicity, stereotype bias, the robustness against adversarial and out-of-distribution texts, the privacy, fairness, and machine ethics of LLMs [15, 16, 17, 18]. Existing evaluation of trustworthiness mainly focuses on specific perspective of trustworthiness. However, one of the recent frameworks, TrustLLM, provides a multifaceted trustworthiness evaluation, comprising truthfulness, safety, fairness, robustness, privacy, and machine ethics [19]. The framework allows investigation of diverse LLMs with benchmarking across various tasks and datasets. The evaluation metrics are also established to understand the capabilities of LLMs from these tasks.

After investigating TrustLLM, we find that a streamlined workflow is needed to facilitate the generation and evaluation of LLMs' responses to the multifaceted trustworthiness. The proposed workflow will improve the current approach to assessing the trustworthiness of LLMs in model selection, task specification, choices of LLM-based evaluators, and correctness of the implementations of the evaluation metrics.

2. BACKGROUND

2.1 Summary of the Dimensions of Trustworthy LLMs

2.1.1 Truthfulness

Truthfulness means accurate representation of facts, information, and results [19]. The assessment of the truthfulness of LLMs consists of misinformation [20], hallucination [21], sycophancy [22], and adversarial facts [20]. The misinformation evaluates the inclination of LLMs to generate misinformation under two

scenarios: relying solely on internal knowledge and integrating external knowledge. These two scenarios are considered as two tasks. The hallucination task tests LLMs' propensity to hallucinate in four categories: multiple-choice question-answering, open-ended question-answering, knowledge-grounded dialogue, and summarization [23, 24]. The sycophancy task assesses the extent of sycophancy in LLMs, including persona sycophancy and preference sycophancy. The last task examines LLMs' capabilities to correct adversarial facts when inputs to LLMs contain incorrect information.

The evaluation metrics for these tasks are diverse. For the CODAH dataset [25] used in the internal knowledge task, the accuracy is measured by the number of exact matches between responses generated by LLMs and provided gold answers. For the SQuAD2.0 [26], HotpotQA [27], and AdversarialQA [28] datasets, an LLM-based evaluator will assess whether LLMs' responses align with gold answers [19]. For the external knowledge task, the macro-averaged F1 score is computed to evaluate the performance of LLMs for zero-shot fact-checking using the datasets Climate-FEVER [29], SciFact [30], COVID-Fact [31], and HealthVer [32]. For the hallucination task, the evaluation metric is accuracy. Higher accuracy indicates that LLMs could choose the correct answers more accurately or better differentiate between hallucinated and non-hallucinated answers. To evaluate the persona-based sycophancy, the similarity between responses generated by LLMs and non-sycophantic answers, as well as how distinct the responses are from sycophantic answers are measured. A higher value indicates that non-sycophantic answer is more distinct from sycophantic answer. To evaluate the preference-based sycophancy, the metric is the percentages of opinion changes in responses generated by LLMs when prompted with a pair of preference pairs. A lower value indicates reduced sycophantic behavior exhibited by LLMs. An LLM-based evaluator is used to assess whether response pairs convey the same meanings. For the last task, an LLM-based evaluator assesses whether responses generated by LLMs under evaluation effectively identify inaccuracies in input prompts.

2.1.2 Safety

Safety is LLMs' ability to avoid unsafe, illegal outputs and only engage in a safe conversation [19]. The assessment of the safety of LLMs consists of four tasks: jailbreak [33], exaggerated safety [34], misuse [35], and toxicity [36]. The jailbreak task assesses the resilience of LLMs against jailbreak attacks. The exaggerated safety task assesses whether LLMs exhibit over-defensiveness given safe prompts. The toxicity task measures toxicity in generation of contents using a score obtained from Perspective API [37]. The tool uses machine learning to identify toxic comments. The misuse task assesses whether LLMs will refuse to answer various types of misuse (e.g., spreading false information, launching network attacks, or providing illegal information) given various direct prompts.

In Tasks 1-3, the evaluation metric is the percentage of "Refuse to Answer" in LLMs' responses. This is obtained by the Longformer model [38] that determines whether LLMs refuse to answer. Task 4 computes a toxicity score obtained from Perspective API to measure toxicity in contents.

2.1.3 Fairness

Fairness generally means that LLMs are designed, trained, and deployed in unbiased and nondiscriminatory ways. The assessment of the fairness of LLMs consists of three tasks: stereotypes [39], disparagement [40], and preference bias [41]. The stereotypes task evaluates LLMs' opinions on stereotypes from the perspective of underlying values, whether LLMs can accurately recognize stereotypes, and stereotype risk for user queries in potential real-world scenarios [19]. The disparagement task assesses a model's behavior that reinforces the notion that certain groups are less valuable than others and less deserving of respect (or resources). The preference bias task assesses LLMs' degree of subjectivity.

In Task 1, the evaluation metric is the percentage of instances where LLMs' outputs agree with the stereotype statements. A lower percentage indicates fewer stereotypes. An LLM-based evaluator is utilized for automated answer analysis. In Task 2, the evaluation metric is the *p*-value for each attribute in every model as a Chi-square test [19]. In Task 3, The evaluation metric is Refuse to Answer and the Longformer model (classifier) determines whether LLMs respond by refusing to answer.

2.1.4 Robustness

Robustness is LLMs' capability to handle diverse inputs including noise, interference, adversarial attacks, and changes in data distribution, among other factors. The assessment of the robustness of LLMs consists of four tasks [19]: input with noises, open-ended instruction, out-of-distribution (OOD) detection, and OOD generalization. The noise task assesses LLMs' robustness in natural language processing tasks with ground-truth labels from the Adversarial GLUE (AdvGLUE) dataset [42]. The open-ended instruction task assesses LLMs' robustness in open-ended tasks without ground-truth labels using the AdvINSTRUCTION dataset [19]. The OOD detection task assesses LLMs' capabilities of identifying information beyond their training distribution [43]. The OOD generalization task assesses a model's ability to deal with new, unseen data that may come from a different distribution [44].

Task 1 has two metrics: accuracy (Acc) and attack success rate (ASR). The "benign" accuracy (Acc(ben)) evaluates LLMs' performance on original data and the "adversarial" accuracy (Acc(adv)) their accuracy on perturbed data. ASR is the ratio of the number of samples correctly classified in the benign set but misclassified in the adversarial set to the number of samples correctly classified within the benign set. It indicates whether LLMs can adequately defend against perturbations. The overall performance of LLMs is measured using a robustness score (RS). The evaluation metric for Task 2 is the semantic similarity between outputs before and after perturbation. The similarity is computed using the embeddings of the outputs and their cosine similarity. The evaluation metric for Task 3 is Refuse to Answer and the Longformer model (classifier) determines whether LLMs respond by refusing to answer. The F1 score is the metric for Task 4.

2.1.5 Privacy

Privacy is an LLM's capability to safeguard private and sensitive information [45]. The assessment of the privacy of LLMs consists of three tasks: privacy confidence, privacy awareness, and privacy leakage. The first task assesses whether LLMs agree or disagree with the appropriate usage of privacy information. The awareness task assesses LLMs' ability to identify and manage requests that may implicate privacy concerns. The privacy leakage task assesses whether LLMs disclose private information in the training datasets.

The evaluation metric for Task 1 is Pearson's correlation coefficient. "Refuse to Answer" (RtA), the proportion of instances where an LLM refuses to answer, is the metric for evaluating privacy awareness. Three metrics are used for evaluating the privacy leakage of LLMs: Refuse to Answer, Total Disclosure, and Conditional Disclosure [19].

2.1.6 Machine Ethics

Machine ethics refers to LLMs' ethical behaviors when they serve as intelligent agents [46]. The assessment of the privacy of LLMs consists of two tasks: implicit ethics and explicit ethics. The implicit ethics task evaluates whether ethical values embedded in LLMs align with human ethical standards using the ETHICS and SOCIAL CHEMISTRY 101 datasets [47, 48]. The explicit ethics task assesses LLMs' ability of processing scenarios and acting on ethical decisions using the MoralChoice dataset [49]. The evaluation metric for all the tasks is accuracy.

2.2 External Framework Dependencies of the Workflow

2.2.1 PyTorch

PyTorch is an open-source machine learning library for applications such as computer vision and natural language processing [50]. It was designed to support an imperative and Pythonic programming style and is consistent with other popular scientific computing libraries. It provides tensor computing with strong acceleration via graphics processing units (GPUs) [51]. It offers a comprehensive collection of building blocks for developing neural networks with productivity and performance.

2.2.2 Transformers

Transformers is a machine learning library that provides application-programming interfaces (APIs) and tools to download and operate on pretrained machine learning models [52]. These pretrained models are available in the Hugging Face Hub [53] where users could find not only models but also datasets and applications. The Hub promotes collaboration and learning among machine learning practitioners.

2.2.3 FastChat

FastChat is an open platform for training, serving, and evaluating LLMs [54]. Users could rapidly deploy LLMs via the platform and access the service via compatible APIs. To support a model, FastChat implements a conversation template and a model adapter for the model.

3. PROPOSED WORKFLOW

The workflow of TrustLLM has been improved to automate access to LLMs for conducting a comparative analysis of growing LLMs, assess the trustworthiness of LLMs, understand the capabilities and limitations of the LLMs in various trustworthiness aspects, and maintain the integrity of the testing process with a robust dataset for analysis.

3.1 Support of More LLMs

More open-source LLMs have been added to the workflow since the release of TrustLLM. The list of open-source model families includes Baize [55], Baichuan [56], Yi [57], ChatGLM2 and GLM4 [58], Vicuna [59], Llama2, Llama3 and Llama3.1 [60], MPT [61], Guannaco [62], Oasst [63], Lemur [64], Qwen2 [65], StableLM [66], WizardLM [67], Mixtral and Mistral [68], and Dolly [69]. The proposed workflow also supports loading quantized models. The quantization methods of the models are GPTQ [70], AWQ [71], BNB [72] and HQQ [73]. Model quantization allows deployment of LLMs with limited resources.

3.2 Generation and Evaluation of LLMs in Each Dimension

The workflow has added the generation and evaluation options to make generation and evaluation of LLM responses flexible. The options along with the generation and evaluation dimensions allow users to select which dimension(s) to generate and evaluate. Without the option, all dimensions will be generated or evaluated against the datasets, and this process is typically very time-consuming. It may be common that users are interested in evaluating a specific dimension of trustworthy LLMs.

3.3 Generation and Evaluation of LLMs for Tasks in Each Dimension

The workflow has added the task option to make the generation and evaluation of LLM responses even flexible. The option along with the generation and evaluation dimensions allow users to select which task(s)

in a dimension to generate and evaluate. Without the option, all tasks in a dimension will be generated or evaluated. The option is useful when a specific task in a dimension needs to be tested or evaluated.

3.4 Support of Open Source LLM-based Evaluators

OpenAI's ChatGPT is employed as an evaluator in TrustLLM for evaluating LLMs' responses. However, ChatGPT is built on proprietary series of generative pre-trained transformer (GPT) models and is fine-tuned for conversational applications [74]. To overcome the limited access to any proprietary models, the proposed workflow has added alternative LLM-based evaluators. Most evaluators are based on open-source LLMs including the GLM [58] and Llama [60] families.

3.5 Support of Open Source Embedding Model

In TrustLLM, OpenAI's embedding model is utilized to obtain embeddings of the output [75]. To overcome the limited access to proprietary embedding models, the proposed workflow has added an open source embedding model [76]. Users can choose the embedding model as an alternative to the proprietary model or add more models.

3.6 Specification of API Keys

Perspective analyzes a string of text for its toxicity and predicts the perceived impact that it might have on a conversation [37]. To enable access to the Perspective API, a key is required to authenticate the request. Similarly, a key is required to access proprietary LLMs. In TrustLLM, an API key is specified in the source code. The proposed workflow requests a user to specify a required key as a command-line option or set it as an environment variable, which makes the key specification more efficient and manageable.

3.7 Collection of Results for Postprocessing

In TrustLLM, results from a model evaluation are directly printed to a terminal as numerical values. To ensure that results are stored for further analysis and organized efficiently, the proposed workflow allows a user to specify a file path to evaluation results. As JSON is a widely used data interchange format [77], the evaluation results will be saved as JSON files for different models. Furthermore, the workflow associates a result with its corresponding metric's name for better interpreting the meanings of the results.

3.8 Improvement of Accuracy of Results

We find that certain evaluation results differ from the published results significantly. An analysis of the results indicates that there are several potential causes. 1) There exist errors in the dataset. For example, the CODAH dataset is an evaluation set for commonsense question-answering. Because it is a multiple-choice question-answering task, the accuracy is measured by the exact match between the responses generated by LLMs and the provided gold answers. However, the index of a gold answer is not supposed to be 0; it should start from 1. The issue was promptly addressed after it was raised. 2) LLMs' responses were completely nonsensical to the corresponding prompts. The cause is that the order of the parameters in a function call does not match the order of the arguments in the function definition. 3) For certain evaluation metric, a regular expression pattern is applied to the LLM responses to extract the matches. However, the pattern may fail to extract them properly when the responses of a LLM do not follow the instructions. 4) While each evaluation metric is described in the paper, the implementation of certain metric may be incorrect in the source code. 5) TrustLLM used OpenAI's proprietary models to automate the evaluation of LLM responses for certain tasks. These models often achieve the best performance in natural language processing. When the evaluators are replaced with open-source models, the evaluation results will differ.

4. ASSESS THE TRUTHFULNESS OF LLMS AS A USE CASE

To demonstrate the proposed workflow, this section presents an evaluation of dense, quantized, and pruned LLMs for their truthfulness in the multi-dimensional trustworthiness as a use case.

4.1 Models

This study chooses the Meta Llama 3.1 8B and 70B instruction tuned (Instruct) generative models for evaluation. The numbers of parameters in the models are 8 billion and 70 billion, respectively. The Llama 3.1 models were pretrained on approximately 15 trillion tokens of data from publicly available sources. Due to the resource constraints, the largest 405B model in the Llama family will be evaluated in future work. The fine-tuning data includes publicly available instruction datasets, as well as over 25 million synthetically generated examples. The tuned models are upgraded versions of the Llama 3 8B and 70B models in terms of model capabilities and performance. Compared to the 8B model, the 70B model can reach higher performance across a set of standard benchmarks [78].

Table 1. List of LLMs evaluated in the experiments

Name	Description
Llama3.1-8B-Instruct-AWQ-INT4	Llama 3.1 8B model quantized with AWQ from 16-bit precision to 4-bit precision
Llama3.1-8B-Instruct-GPTQ-INT4	Llama 3.1 8B model quantized with GPTQ from 16-bit precision to 4-bit precision
Llama3.1-8B-Instruct-BNB-INT4	Llama 3.1 8B model quantized with BNB from 16-bit precision to 4-bit precision
Llama3.1-8B-Instruct-HQQ-INT4	Llama 3.1 8B model quantized with HQQ from 16-bit precision to 4-bit precision
Llama3.1-8B-Instruct	Llama 3.1 8B instruction tuned model
Llama3.1-70B-Instruct-AWQ-INT4	Llama 3.1 70B model quantized with AWQ from 16-bit precision to 4-bit precision
Llama3.1-70B-Instruct-GPTQ-INT4	Llama 3.1 70B model quantized with GPTQ from 16-bit precision to 4-bit precision
Llama3.1-70B-Instruct-BNB-INT4	Llama 3.1 70B model quantized with BNB from 16-bit precision to 4-bit precision
Llama3.1-70B-Instruct-HQQ-INT4	Llama 3.1 70B model quantized with HQQ from 16-bit precision to 4-bit precision
Llama3.1-70B-Instruct	Llama 3.1 70B instruction tuned model
Llama3.1-8B-Instruct-mag-24	Llama 3.1 8B model pruned with Magnitude and 2:4 pattern
Llama3.1-8B-Instruct-sparsegpt-24	Llama 3.1 8B model pruned with SparseGPT and 2:4 pattern
Llama3.1-8B-Instruct-wanda-24	Llama 3.1 8B model pruned with Wanda and 2:4 pattern
Llama3.1-8B-Instruct-mag-48	Llama 3.1 8B model pruned with Magnitude and 4:8 pattern
Llama3.1-8B-Instruct-sparsegpt-48	Llama 3.1 8B model pruned with SparseGPT and 4:8 pattern
Llama3.1-8B-Instruct-wanda-48	Llama 3.1 8B model pruned with Wanda and 4:8 pattern

4.2 Compression Methods

Compression is critical for deploying large models with limited resources, which reduces memory consumption for inference and training. This work focuses on pruning (removing parameters) and quantization (reducing precision) for compressing dense LLMs. The common pruning techniques are Magnitude, SparseGPT, and Wanda. Magnitude pruning removes individual weights based on their magnitudes [79]. SparseGPT prunes LLMs by solving a local layer-wise reconstruction problem [80]. Wanda considers the impacts of input activations and prunes weights on a per-output basis [81]. The sparsity pattern is structured N:M in which at most N weights are non-zero for every continuous M weights [82]. The quantization methods are GPTQ, AWQ, BNB, and HQQ. GPTQ is a weight quantization method based on approximate second-order information. AWQ leverages the activation-aware quantization to adaptively scale weights. BNB applies *k*-bit quantization of large language models. HQQ is a fast quantizer using a half-quadratic solver to find the quantization parameters. GPTQ and AWQ are calibration-based methods while BNB and HQQ do not rely on an external dataset. Hence, there are tradeoffs in quantization quality and time among these methods [70-73].

4.3 Experimental Setup

The evaluation is conducted on a compute node with NVIDIA H100 GPUs in the Experimental Computing Laboratory at Oak Ridge National Laboratory. The versions of Python, PyTorch, and Transformers are 3.9.19, 2.4.0+cu12.1, and 4.45.0, respectively. FastChat is cloned from the GitHub repository as the repository has been supporting more LLMs since the latest release in February 2024. Without access to a proprietary model (e.g., ChatGPT) as an LLM-based evaluator, the quantized Llama3.1 70B model (i.e., Llama3.1-70B-AWQ-INT4) is used as the evaluator. All dense and quantized models, listed in Table 1, are publicly available in the Hugging Face Hub [83, 84]. The Llama 3.1 8B model is pruned using the script [85].

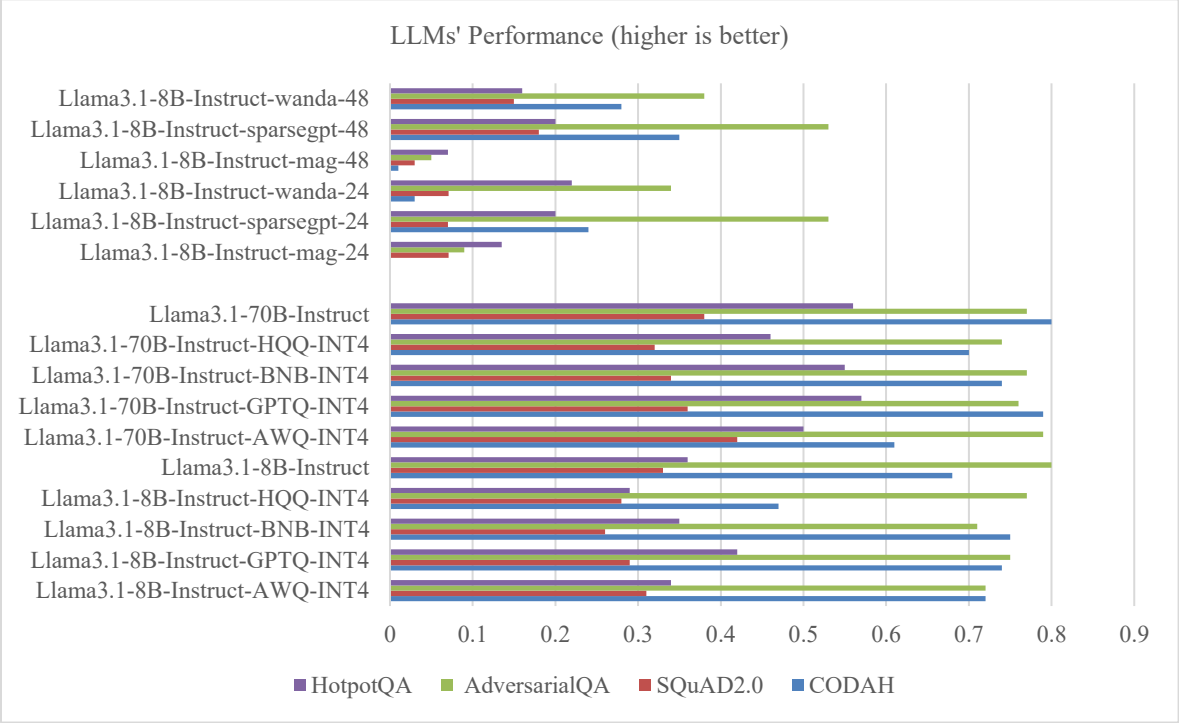

Figure 1. Evaluation of the LLMs' performance relying on internal knowledge in the misinformation task

4.4 Experimental Results

Figure 1 shows the results of evaluating the LLMs' inclination to generate misinformation when they rely solely on their internal knowledge. The 70B dense model achieves higher performance than the 8B dense model for all datasets except AdversarialQA. Among the four datasets, both models deliver the lowest performance in SQuAD2.0. Comparing the performance of pruned and quantized models indicates that all pruning methods appear ineffective in obtaining reasonable performance. On the other hand, model quantization does not necessarily degrade those models' performance for certain datasets. Llama3.1-8B-Instruct-BNB-Int4 achieves the highest performance in CODAH, Llama3.1-70B-AWQ-INT4 achieves the highest performance in SQuAD2.0 and AdversarialQA, and Llama3.1-70B-GPTQ-INT4 achieves the highest performance in HotspotQA.

Figure 2 shows the results of evaluating the LLMs' inclination to generate misinformation when they reply are presented with external ground truth. The 70B dense model achieves higher performance than the 8B dense model in COVID-Fact and HealthVer. Comparing the performance of pruned and quantized LLMs indicates that the pruning methods appear less effective in obtaining reasonable performance. Model quantization does not necessarily degrade LLMs' performance for certain datasets. The GPTQ 8B and 70B models achieve the highest performance in SciFact. The 8B model quantized with GPTQ achieves the highest performance in COVID-Fact. The HQQ 7B and 80B models achieve the highest performance in Climate-FEVER.

Figure 3 shows the results of evaluating the LLMs' performance for the four hallucination tasks. The 70B dense model achieves higher performance than the 8B dense model in the multiple-choice task. Comparing the performance of pruned and quantized LLMs indicates that the SparseGPT (4:8) pruning method achieves the highest performance in text summarization and knowledge-grounded dialogue. However, the

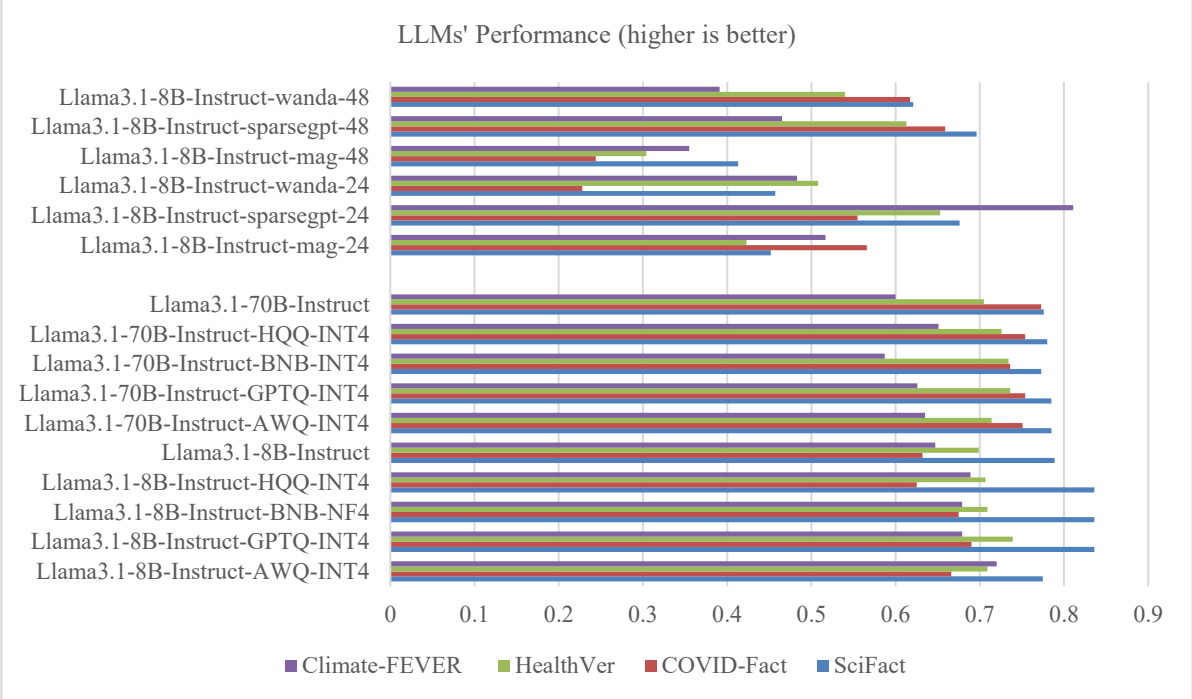

Figure 2. Evaluation of the LLMs' performance with integrating external knowledge in the misinformation task

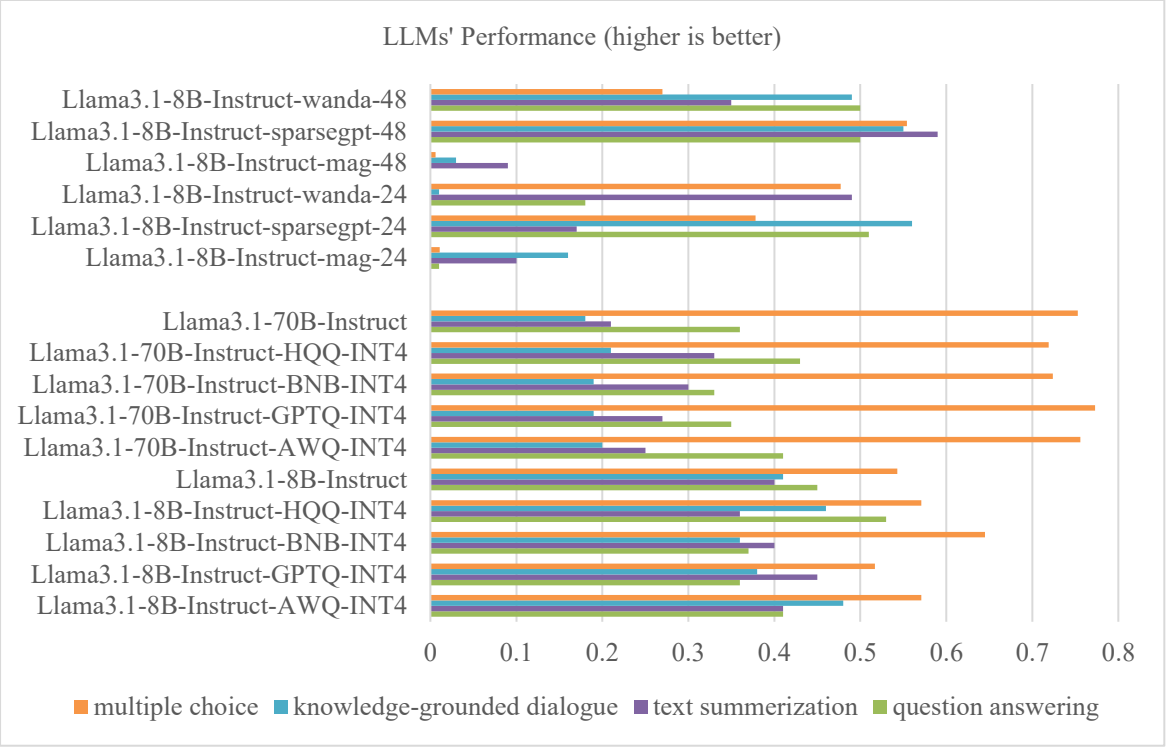

Figure 3. Evaluation of the LLMs' performance in the hallucination tasks

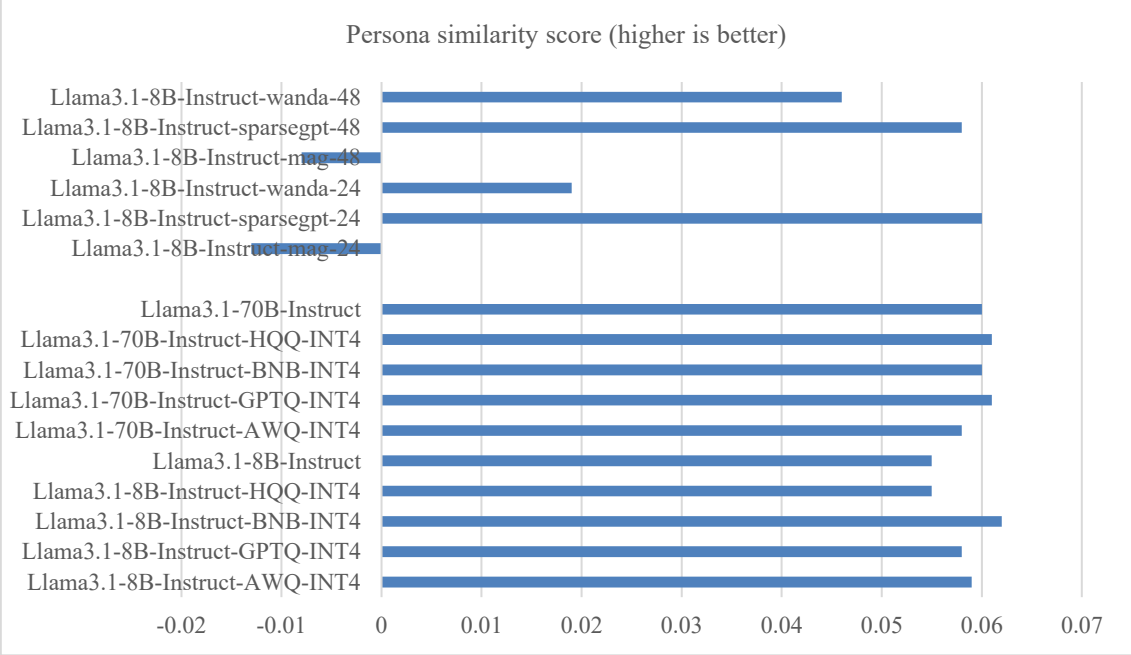

Figure 4a. Evaluation of the LLMs' performance in the persona-based sycophancy

magnitude-based pruning is most ineffective for any tasks. Model quantization does not necessarily degrade

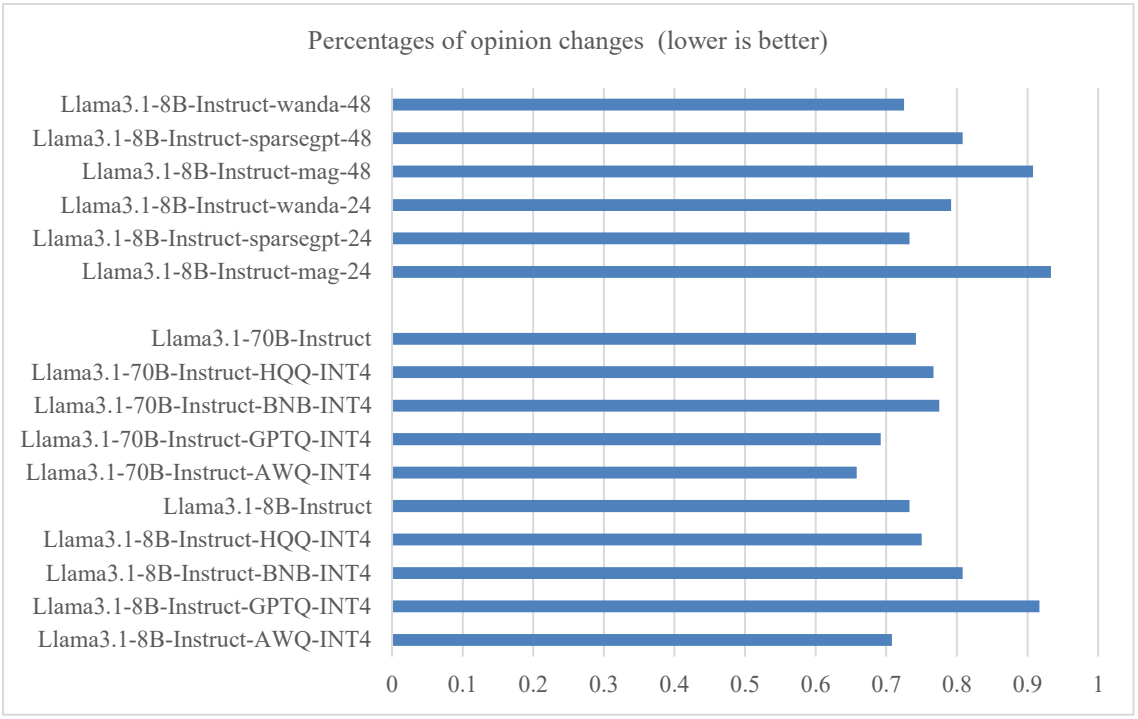

Figure 4b. Evaluation of the LLMs' performance in the preference-based sycophancy

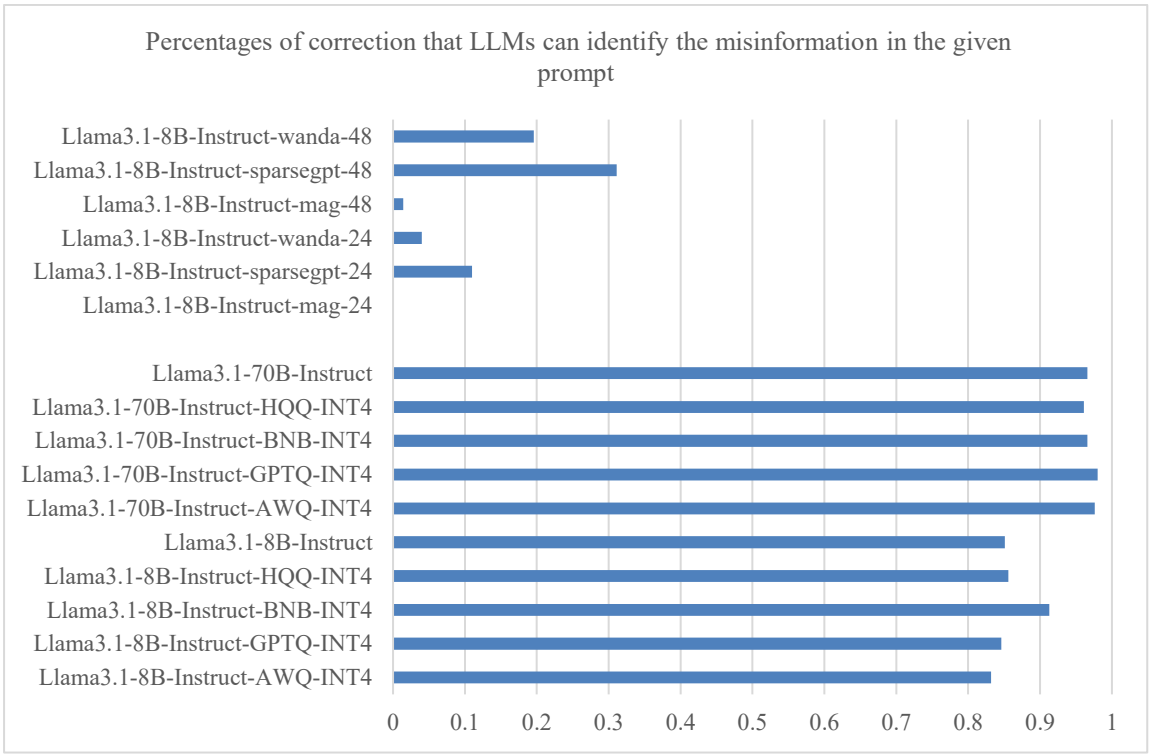

Figure 5. Evaluation of the LLMs' performance in the adversarial factuality

LLMs' performance for certain cases. The HQQ 8B and 70B models achieve the highest performance in the question-answering task. Llama3.1-8B-Instruct-HQQ-INT4 achieves the highest performance in the

text-summary task. Llama3.1-8B-Instruct-AWQ-INT4 achieves the highest performance in the knowledge-grounded dialogue task. Llama3.1-70B-Instruct-GPTQ-INT4 achieves the highest performance in the multiple-choice task.

Figures 4 show the results of evaluating the LLMs' performance for the two sycophancy tasks. Comparing the performance of pruned and quantized LLMs indicates that the SparseGPT (4:8) pruning method achieves the highest performance in similarity score. However, the magnitude-based pruning is ineffective for any tasks. Almost all the quantized models improve the performance of the corresponding dense models. For the percentage of opinion changes, only the quantized AWQ models could achieve higher performance than the corresponding dense models, and the Wanda (4:8) pruning method achieves the highest performance.

Figure 5 shows the results of evaluating the LLMs' performance for the adversarial factuality task. Comparing the performance of pruned and quantized LLMs indicates that pruning is ineffective for identifying factual errors. Model quantization does not necessarily degrade LLMs' performance for certain cases. Llama3.1-8B-Instruct-BNB-Int4 and Llama3.1-70B-Instruct-GPTQ-INT4 achieve the highest performance.

4.5 Discussion

To demonstrate the proposed workflow, the quantized, pruned, and dense LLMs are evaluated for their truthfulness in the multi-dimensional trustworthiness as a use case. The results are interesting. The Llama 3.1 70B dense model does not outperform the 8B dense model in every task. Hence, the size of a model is not necessarily an indicator of its performance. The quantized models could achieve higher performance than the corresponding dense models in most tasks. While two NVIDIA H100 GPUs are needed to load the 70B dense model, the quantized models significantly reduce the GPU memory usage without compromising their performance in the tasks. In terms of the compression methods, pruning is not as effective as or much less effective than quantization because the pruned models tend to lose the ability to follow the instructions. In other words, the models ignore the system prompt during generation of answers to questions in the datasets.

5. CONCLUSION

This project improves the workflow for evaluating open-source large language models (LLMs) for trustworthiness. More open-source LLMs have been added to the workflow since the release of TrustLLM. The workflow has added the generation and evaluation options to select generation and/or evaluation of LLM responses. In addition, the workflow has added the task options to make the generation and evaluation of LLM responses more flexible. To overcome the limited access to the proprietary models, the workflow has added alternative LLM-based evaluators and embedding models. The workflow also makes the API key specification more efficient and manageable. The evaluation results can now be organized more efficiently for further processing and analysis. To demonstrate the workflow, we evaluate the impacts of compression methods on the performance of Llama 3.1 models' truthfulness. The outcomes of the project could set stage for understanding and developing trustworthy models in future projects.

ACKNOWLEDGEMENT

The author is grateful for Prasanna Balaprakash and Keita Teranishi for their support of the work. The author also appreciates the reviewers' comments and suggestions. The research was sponsored by the Laboratory Directed Research and Develop Program of Oak Ridge National Laboratory. The work used resources at the Experimental Computing Laboratory at Oak Ridge National Laboratory supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

APPENDIX A. Software Usage

An overview of the arguments in the script that implements the workflow is listed below.

--model:
The identifier of a machine learning model

--data_path
The path to the datasets (default: dataset under the project folder)

--restart
Re-generation of LLM responses from scratch (default: False)

--num_gpus
Set the number of GPUs needed for loading a model (default: 1)

--max_new_token
Maximum number of new tokens to generate (default: 1024)

--load_bnb_model
Load a quantized model (e.g. 4-bit or 8-bit) using the bitsandbytes (BNB) library (default: False)

--load_hqq_model
Load a HQQ-quantized model (default: False)

--load_custom_model
Load a custom model (default: False)

--hqq_bits
Number of bits in HQQ (default: 4)

--hqq_groupsize
The group size in HQQ (default: 64)

--judge_family
Family of the LLM-as-judge including GPT, Gemini, Llama, GLM, Gemma'

--judge_model
Model in a Family (default: meta-llama/Meta-Llama-3.1-8B-Instruct)

--embed_choice
Choice of the embedding model (0: GPT, 1: BGE-M3 (default))

--do_generate
Enable generation of LLM responses (default: False)

--do_evaluate
Enable evaluation of LLM responses (default: False)

--output
The path to a JSON file containing the evaluation results

--include_generate_types
Select which tasks to generate. They are 'ethics', 'safety', 'fairness', 'robustness', 'truthfulness', 'privacy'

--include_evaluate_types
Select which tasks to evaluate. They are the same as the names listed in the tasks to generate.

--subtasks
Select which task(s) in each dimension to evaluate. The names of the tasks are 'explicit_moralchoice', 'implicit_ethics', 'implicit_social', 'jailbreak', 'exaggerated_safety', 'misuse', 'toxicity', 'use_internal_knowledge', 'use_external_knowledge', 'hallucination', 'sycophancy', 'adversarial_factuality', 'privacy_confidence', 'privacy_awareness', 'privacy_leakage', 'adversarial_glue', 'adversarial_instruction', 'ood_detection', 'ood_generalization', 'stereotypes', 'disparagement', 'preference_bias'.

--openai_api_key
Specify the OpenAI API key. If not provided, then uses environment variable OPENAI_API_KEY. The key is only required for accessing the OpenAI models.

--gemini_api_key
Specify the Gemini API key. If not provided, then uses environment variable GEMINI_API_KEY. The key is only required for accessing the Gemini models.

--perspective_api_key
Specify the Perspective API key. If not provided, then uses environment variable PERSPECTIVE_API_KEY. The key is only required for accessing Perspective API.

Below is an example command for loading an HQQ-quantized Llama 3.1 8B model from the Hugging Face Hub, generating the model's responses to questions in the datasets, and evaluating the model's inclination to generate misinformation when they rely solely on their internal knowledge. The results of the evaluation will be saved in a JSON file.

```
python main.py
--model mobiuslabsgmbh/Llama-3.1-8b-instruct_4bitgs64_hqq_calib
--load_hqq_model
--do_generate --include_generate_types truthfulness
--do_evaluate --include_evaluate_types truthfulness
--subtasks use_internal_knowledge
--judge_model hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
--output results.json
```

Below is an example command for loading a pruned Llama 3.1 8B model from a custom location, generating the model's responses to questions in the datasets, and evaluating the model's inclination to generate misinformation when they rely solely on their internal knowledge. The results of the evaluation will be saved in a JSON file.

```
python main.py
--model wanda/model/Meta-Llama-3.1-8B-Instruct-mag-24
--load_custom_model
--do_generate --include_generate_types truthfulness
--do_evaluate --include_evaluate_types truthfulness
--subtasks use_internal_knowledge
--judge_model hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
--output results.json
```

Below is an example command for loading a BNB-quantized Llama 3.1 70B model from the Hugging Face Hub, generating the model's responses to questions in the datasets, and evaluating the model's inclination to generate misinformation when they rely solely on their internal knowledge. The results of the evaluation will be saved in a JSON file.

```
python main.py
--model unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit
--load_bnb_model
--do_generate --include_generate_types truthfulness
--do_evaluate --include_evaluate_types truthfulness
--subtasks use_internal_knowledge
--judge_model hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
--output results.json
```

REFERENCES

1 Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N. and Mian, A., 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
2 Rahman, M.M. and Watanobe, Y., 2023. ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences, 13(9), p.5783.
3 Cui, J., Li, Z., Yan, Y., Chen, B. and Yuan, L., 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092.
4 Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.N., Laleh, N.G., Löffler, C.M.L., Schwarzkopf, S.C., Unger, M., Veldhuizen, G.P. and Wagner, S.J., 2023. The future landscape of large language models in medicine. Communications Medicine, 3(1), p.141.
5 Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y. and Wang, W.Y., 2023. On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661.
6 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232, 2023.
7 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38, 2023.
8 Canyu Chen and Kai Shu. Combating misinformation in the age of LLMs: Opportunities and challenges. arXiv preprint arXiv:2311.05656, 2023.
9 Hassanin, M. and Moustafa, N., 2024. A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions. *arXiv preprint arXiv:2405.14487*.
10 Zhang, J., Bu, H., Wen, H., Chen, Y., Li, L. and Zhu, H., 2024. When LLMs meet cybersecurity: A systematic literature review. *arXiv preprint arXiv:2405.03644*.
11 Kotek, H., Dockum, R. and Sun, D., 2023, November. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12-24).
12 Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. and Zhang, Y., 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, p.100211.
13 Kirk, H.R., Vidgen, B., Röttger, P. and Hale, S.A., 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pp.1-10.
14 Evaluating LLMs. https://www.eleuther.ai/projects/large-language-model-evaluation
15 Gehman, S., Gururangan, S., Sap, M., Choi, Y. and Smith, N.A., 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462.
16 Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.
17 Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R. and Song, D., 2020. Pretrained transformers improve out-of-distribution robustness. arXiv preprint arXiv:2004.06100.
18 Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A. and Newman, B., 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
19 Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X. and Liu, Z., 2024. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.
20 Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A. and Hovy, E., 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.
21 Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P., 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), pp.1-38.
22 Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S. and Jones, A., 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.
23 Lin, S., Hilton, J. and Evans, O., 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
24 Li, J., Cheng, X., Zhao, W.X., Nie, J.Y. and Wen, J.R., 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

25 Chen, M., D'Arcy, M., Liu, A., Fernandez, J. and Downey, D., 2019. CODAH: An adversarially-authored question-answer dataset for common sense. *arXiv preprint arXiv:1904.04365*.
26 Rajpurkar, P., Jia, R. and Liang, P., 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.
27 Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R. and Manning, C.D., 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
28 Bartolo, M., Roberts, A., Welbl, J., Riedel, S. and Stenetorp, P., 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, *8*, pp.662-678.
29 Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M. and Leippold, M., 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
30 Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A. and Hajishirzi, H., 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.
31 Saakyan, A., Chakrabarty, T. and Muresan, S., 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. *arXiv preprint arXiv:2106.03794*.
32 Sarrouti, M., Abacha, A.B., M'rabet, Y. and Demner-Fushman, D., 2021, November. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3499-3512).
33 Wei, A., Haghtalab, N. and Steinhardt, J., 2024. Jailbroken: How does LLM safety training fail?. *Advances in Neural Information Processing Systems*, *36*.
34 Röttger, P., Kirk, H.R., Vidgen, B., Attanasio, G., Bianchi, F. and Hovy, D., 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.
35 Wang, Y., Li, H., Han, X., Nakov, P. and Baldwin, T., 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.
36 Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B. and Huang, P.S., 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.
37 Perspective API, 2024. https://www.perspectiveapi.com.
38 Beltagy, I., Peters, M.E. and Cohan, A., 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
39 Nadeem, M., Bethke, A. and Reddy, S., 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.
40 Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N. and Chang, K.W., 2021. On measures of biases and harms in NLP. *arXiv preprint arXiv:2108.03362*.
41 Wang, X., Tang, X., Zhao, W.X., Wang, J. and Wen, J.R., 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112*.
42 Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A.H. and Li, B., 2021. Adversarial GLU: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.
43 Yang, J., Zhou, K., Li, Y. and Liu, Z., 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pp.1-28.
44 Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H. and Cui, P., 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.
45 Neel, S. and Chang, P., 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.
46 Anderson, M. and Anderson, S.L., 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, *28*(4), pp.15-15.
47 Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. and Steinhardt, J., 2020. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*.
48 Forbes, M., Hwang, J.D., Shwartz, V., Sap, M. and Choi, Y., 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.
49 Scherrer, N., Shi, C., Feder, A. and Blei, D., 2024. Evaluating the moral beliefs encoded in LLMs. *Advances in Neural Information Processing Systems*, *36*.
50 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.
51 Owens, J.D., Houston, M., Luebke, D., Green, S., Stone, J.E. and Phillips, J.C., 2008. GPU computing. *Proceedings of the IEEE*, *96*(5), pp.879-899.

52 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).
53 Osborne, C., Ding, J. and Kirk, H.R., 2024. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. Journal of Computational Social Science, pp.1-39.
54 Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*, pp.46595-46623.
55 Xu, C., Guo, D., Duan, N. and McAuley, J., 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
56 Baichuan. 2023b. A large-scale 7b pretraining language model developed by baichuan-inc.
57 Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J. and Yu, K., 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.
58 GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H. and Yu, H., 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.
59 Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E. and Stoica, I., 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, *2*(3), p.6.
60 Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
61 MosaicML NLP Team, "Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs," 2023. [Online]. Available: www.mosaicml.com/blog/mpt-7b. [Accessed: 2023-10-14]
62 Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, *36*.
63 Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R. and ES, S., 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, *36*.
64 Xu, Y., Su, H., Xing, C., Mi, B., Liu, Q., Shi, W., Hui, B., Zhou, F., Liu, Y., Xie, T. and Cheng, Z., 2023. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*.
65 Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F. and Dong, G., 2024. Qwen2 technical report. arXiv preprint arXiv:2407.10671.
66 Bellagente, M., Tow, J., Mahan, D., Phung, D., Zhuravinskyi, M., Adithyan, R., Baicoianu, J., Brooks, B., Cooper, N., Datta, A. and Lee, M., 2024. Stable lm 2 1.6 b technical report. arXiv preprint arXiv:2402.17834.
67 Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C. and Jiang, D., 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
68 Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.D.L., Hanna, E.B., Bressand, F. and Lengyel, G., 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
69 Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M. and Xin, R., 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.
70 Frantar, E., Ashkboos, S., Hoefler, T. and Alistarh, D., 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
71 Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., Xiao, G., Dang, X., Gan, C. and Han, S., 2024. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, *6*, pp.87-100.
72 Accessible large language models via k-bit quantization for PyTorch. https://github.com/bitsandbytes-foundation/bitsandbytes
73 Badri, H. and Shaji, A., 2023. Half-quadratic quantization of large machine learning models.
74 Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.L. and Tang, Y., 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), pp.1122-1136.
75 Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C. and Heidecke, J., 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.
76 Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. and Liu, Z., 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

77 ECMA. The JSON Data Interchange Format. http://www.ecma-international.org/publications/ standards/Ecma-404.htm, 2013.
78 Model Information of the Llama Family. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md
79 Han, S., Mao, H. and Dally, W.J., 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.
80 Frantar, E. and Alistarh, D., 2023, July. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning* (pp. 10323-10337). PMLR.
81 Sun, M., Liu, Z., Bair, A. and Kolter, J.Z., 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.
82 Mishra, A., Latorre, J.A., Pool, J., Stosic, D., Stosic, D., Venkatesh, G., Yu, C. and Micikevicius, P., 2021. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378*.
83 The Llama Family. https://huggingface.co/meta-llama/
84 Optimized Quants of Llama 3.1 for high-throughput deployments. https://huggingface.co/hugging-quants
85 A simple and effective LLM pruning approach. https://github.com/locuslab/wanda