Coverage of LLM Trustworthiness Metrics in the Current Tool Landscape

Lennard Helmer[1,*], Benny Stein[1], Tim Ufer[1], Elanton Fernandes[1], Hammam Abdelwahab[1], Abhinav Pareek[1] and Joshua Woll[1]

[1]*Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, St. Augustin, Germany*

Abstract

The increasing prevalence of AI systems that are build with Large Language Model (LLM) components raises the requirement for a dedicated tool stack that allows to monitor such systems, covering training, development and inference environments. Beside technical performance metrics like latency and throughput, regulations like the EU AI Act require the monitoring of trustworthiness related metrics like fairness and transparency during operation. In this paper, we describe the results of an investigation we conducted to gain an overview of the current landscape of LLM trustworthiness metrics and their coverage in monitoring tools. Based on an in-depth analysis of available catalogs and additional research, we identified 43 metrics and 23 tools. Furthermore, we highlight existing gaps and potential areas for further research. The results support practitioners and researchers in making informed decisions about the most appropriate tech stack for their AI systems.

Keywords

Large Language Models, Artificial Intelligence, Generative AI, Trustworthy AI, Responsible AI, MLOps

1. Introduction

The Organization for Economic Co-operation and Development (OECD) published a catalog that aims to provide an up-to-date and regularly updated overview of metrics and tools[1] that can be used to strengthen AI trustworthiness. However, there is no mapping between metrics and tools for researchers and developers, and it is not possible to assess the extent to which the current metrics for LLM monitoring are covered by the tool landscape.

In this study, we aim at

1. Identifying the most relevant metrics that can be used for LLM operation and monitoring
2. Evaluating their coverage in the most prevalent tools for LLM monitoring and evaluation

To achieve these goals, we analyze the OECD metric catalog, focusing on metrics for monitoring and LLMs, academic literature and consider grey literature where applicable. Furthermore, we investigate the tool landscape targeting the monitoring of LLM-based systems. Finally, we analyze these tools in terms of their coverage of the metrics identified in the first part of our analysis. One dominant approach in practice is to use an LLM-judge [1] and we examine this approach in more detail. It allows for promising, efficient and scalable approaches for improving trustworthiness. On the other hand, the "judge" LLMs are still LLMs, and they inherit the problems of the LLMs that they are supposed to evaluate to a non-negligible extent.

The structure of this paper is as follows: In Section 2, we give background on LLM monitoring and trustworthy AI, including related work. In Section 3, our methodology is presented. We list and discuss

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

lennard.helmer@iais.fraunhofer.de (L. Helmer); benny.joerg.stein@iais.fraunhofer.de (B. Stein); hammam.abdelwahab@iais.fraunhofer.de (H. Abdelwahab); abhinav.pareek@iais.fraunhofer.de (A. Pareek); joshua.woll@iais.fraunhofer.de (J. Woll)

0009-0009-2411-4540 (L. Helmer); 0009-0009-8843-5917 (B. Stein); 0009-0007-8286-8004 (T. Ufer); 0009-0007-5601-2093 (E. Fernandes); 0009-0000-6283-2310 (H. Abdelwahab); 0009-0006-0266-5008 (A. Pareek); 0009-0000-1070-7075 (J. Woll)

[1] https://oecd.ai/en/catalogue/overview

the metrics that we identified in Section 4, and the tools in Section 5. Our findings are discussed in Section 6. Finally, we give a conclusion of our work and an outlook in Section 7.

2. Background and Related Work

Monitoring is an essential aspect of any IT system development. Nevertheless, the unique characteristics and challenges of AI systems emphasize the requirement for monitoring of aspects like trustworthiness. Unfortunately the trustworthiness of AI is not an objective state, and is influenced by several dimensions that influence each other [2], adding complexity to the oversight of such systems.

The currently most widespread paradigm for developing and operating AI systems is Machine Learning Operations (MLOps), and it emphasizes the importance of monitoring in operation [3]. Pipelines for the various phases, such as data preprocessing, training and evaluation, as well as monitoring of these phases, are designed to ensure the quality of the artifacts. Although trustworthiness is not yet explicitly part of the paradigm, first steps have been taken to enhance it and integrate trustworthy AI principles into MLOps processes [4]. A 2024 survey [5] of small and medium-sized enterprises in Germany on the implementation of MLOps found that while most are aware of its importance, few have a well-structured approach to monitoring their AI systems. Most rely on "human-based monitoring" and the ability of users to identify and report AI errors during inference. While this may work in certain domains and applications, such as machine learning based classification, generative AI makes it much more difficult to detect incorrect AI outputs.

In the following subsection, we describe the foundation of AI trustworthiness and its dimensions. After that, we state other related work that we identified during our research, for example targeting aspects of LLM monitoring during (pre-)training and the unique challenges that LLMs pose for monitoring.

2.1. Trustworthy AI

In addition to a list of metrics, the OECD catalogue provides a mapping to specific *objectives*, too. Those objectives are Data Governance & Traceability (DGT), Digital Security (DS), Environmental Sustainability (ES), Explainability (EXP), Fairness (FA), Human Agency & Control (HAC), Performance (PF), Privacy (PR), Robustness (RN), Safety (SF), and Transparency (TR). Although there is no specific description of the scope of the individual objectives, they can be understood in a generic sense. The underlying framework of the catalogue is specified further in [6].

The objectives described before roughly map to the trustworthiness dimensions known from other efforts in trustworthy AI research. For example, the AI Assessment Catalogue [7] defines a large variety of testing tasks, spanning over six dimensions of trustworthiness for evaluating AI applications: Fairness, Autonomy & Control, Transparency, Reliability, Security, and Privacy. It provides a structured basis for evaluating AI systems, both for developers and assessors.

Furthermore, there is ongoing research on LLM-specific trustworthiness dimensions in [8] and in [9]. In the latter work introduced, four AI safety levels (ASL) were introduced that relate model scaling efforts to appropriate safety procedures.

2.2. Other Related Work

The rapid progress in language models (LMs) has resulted in the training of increasingly large LMs on massive quantities of data [10]. LLMs are trained with a substantial amount of curated data and web-based data, predominantly web-based data such as Common Crawl [11]. Data pipelines for web-based data manage data preparation including text extraction, language identification, rigorous filtering, and deduplication, such as the ones in C4 [12], The Pile [13], RefinedWeb [14], CCNet [15], BigScience ROOTS [16], OSCAR [17], and FineWeb [18]. After pre-training, instruction tuning is conducted to improve the performance of LLMs on several desired tasks [19, 20].

Trustworthiness has been considered in some of the research on pre-training pipelines for LLMs. For example, the Ungoliant pipeline uses a pre-trained perplexity-based KenLM to filter unsafe data

[21, 22]. Furthermore, FineWeb's pipeline includes filtering of Personal Identifiable Information [18]. In addition, there are several ongoing research efforts on ensuring safety of LLMs by generating instruction tuning data sets, and using them for training or evaluation. Among them are WildGuard [23] and DecodingTrust [24]. Human evaluation as part of reinforcement learning indirectly contributes to LLM trustworthiness [8], too.

Recent studies [25, 26] have proposed a range of tools to assess different dimensions of trustworthiness. For example, tools that focus on assessing fairness [27] often use bias detection metrics to assess disparities across demographic groups. While there are studies [8, 28] on the subject of trustworthiness of LLMs, to the best of our knowledge there are no papers yet that systematically investigate the coverage of LLM trustworthiness metrics in software tools.

3. Methodology

In the following, we describe the general methodology that we applied to identify the metrics and tools for LLM monitoring.

Starting point is the OECD catalog for trustworthy AI tools and metrics, which curates an up-to-date list of trustworthiness metrics and tools, based on feedback from the community and by automated searches in GitHub. We filtered for metrics that were tagged with the purpose "Content generation" and the life cycle stage "Operate & monitor". For tools, we used the tags "technical" and "operational & monitoring".

Additionally, we performed literature searches for metrics as well as tools, based on search queries that we formulated using keywords commonly discussed in trustworthy AI literature (e.g., trustworthiness). Each query included terms related to the attribute of interest (e.g., fairness, safety) combined with terms representing measurement and evaluation (e.g., metric, assessment, framework, tool). Additionally, each query targeted literature specifically mentioning "large language model", "LLM", or "generative AI" to ensure relevance to this investigation.

The relevance of each publication was assessed based on the number of citations and judgment of the authors of this paper. A notable difference in case of literature related to tools is that the citation count is not as relevant, as it is a reasonable objective measure for the importance of an academic publication, but not meaningful to assess the prevalence of a software tool. We took again the number of citations into account if a matching publication for a tool was available, but also GitHub stars, which are used by the developer community to mark noticeable repositories. On top of that, we also made use of the authors' experience from multiple industry projects that dealt with the guard-railing of LLMs. Each tool was analyzed using publicly available websites, the tool's documentation (if available), and a code analysis if the source code was published. The available information was analyzed for mentioning of the metrics that we had previously identified. While many tools appear promising, not many of them publicly state which metrics are available. In particular proprietary tools do not reveal many details about available metrics. Thus, they were usually not considered.

Our list of judge-based metrics was put together using a literature search on LLM-as-a-judge topics, and additionally scanning the documentation of the most well-known software tools for LLM Testing that were identified in the search process for the tools.

The search process was conducted between the 14th and 17th of April, 2025.

4. Metrics

In this section, we discuss the available metrics for the trustworthiness monitoring of LLMs. We start by presenting and describing our complete collection in Section 4.1, and after that give details about the judge-based metrics in Section 4.2.

4.1. Description of the Final Collection

In the following, we provide an overview of the metrics identified through the process described in Section 3, categorized by their respective objective and summarize the results in Table 1[2]. The prefix O of an ID in the table indicates that the metric was found in the OECD catalog, while the prefix A indicates that the tool was added based on additional research. The prefix J shows that the metric is based on an LLM-as-a-judge approach. It should be noted that some of the metrics from the OECD catalog also rely on a judge approach.

In total, we analyzed 59 metrics and consider 43 of those to be both related to the trustworthiness of LLMs and to be used in monitoring. We excluded metrics that appeared in our research, but where the results did not reveal their suitability for either LLMs or monitoring purposes. To gain a broader perspective, we also searched gray literature such as blog entries and articles for approaches to use the identified metrics in LLM monitoring before deciding whether or not to include a metric in our collection. Of these 43 metrics, 17 are listed in the OECD metrics catalog, 7 were identified during the literature review, and 19 were identified through dedicated research on LLM-as-a-judge approaches. Metrics of those type are discussed in more detail in Section 4.2.

To better understand the distribution of the metrics across the trustworthiness objectives mentioned in Section 2.1, we classified the objective for each of the additional metrics that were identified. We made the classification to the best of our knowledge and based on the documentation available for each metric. For the metrics from the OECD catalog, we used the classification from the catalog itself.

We found that none of the metrics are suitable for the objectives of *Data Governance & Traceability*, *Digital Security* and *Environmental sustainability*, so these three objectives are not listed in Table 1 at all.

Two thirds of the metrics (28) can be used to measure the *Robustness* of an LLM-based system, which is not surprising as most trustworthiness dimensions aim to ensure that the AI system behaves in an expected manner. It also reflects the fact that one of the most studied shortcomings of LLMs are *hallucinations*, and the robustness objective addresses that. *Performance* is the second largest group, with 20 metrics that could be used to provide insight. 14 metrics are suitable for measuring *Safety*-related issues, which is probably related to the widespread awareness of issues such as prompt injection. *Explainability* (10), *Human Agency & Control* (9), *Fairness* (8) and *Transparency* (7) are close in terms of group size of suitable metrics. *Privacy* is clearly the smallest group of related metrics (4).

Figure 1 illustrates the distribution of metrics across the objectives. The blue bars represent the total number of available metrics for each objective, while the green bars indicate the subset of metrics that are currently addressed by existing tools. This visualization does not include additional information beyond the one contained in Table 1, but illustrates the uneven development of assessment capabilities across trustworthy AI objectives and highlights areas requiring enhanced tooling support. It should be noted that the image can only serve as a general orientation, as not every metric is equally valuable but they are all given the same weight in this visualization.

4.2. On Judge-based Metrics

Judge-based metrics rely on a recently established method to improve trustworthiness of LLMs. They are constructed from a "judge" LLM that is given both an instruction for and an answer from the LLM under examination, and a second instruction for the judge itself. The judge *scores* the answer based on these inputs and context (if available), with the specific scoring procedure varying across different metrics. We refer to the original work [1] for details. Judge-based metrics are typically use-case agnostic, highly configurable and also very effective, but they fall short in terms of consistency, reproducibility and are very sensitive to their configuration, such as the prompts that are being used for the judge LLMs. It should be noted that is still unclear how reliable these are in general [38, 39, 40, 41], but we decided

[2]For comprehensive descriptions of all of the identified metrics, the reader is referred to the sources in the last column of the table. It contains several links to overview pages from specific tools - the individual links to the metrics were not provided to not clutter the table, but the reader should be able to navigate to them. We also did not include additional sources for the metrics from the OECD catalog.

Table 1
Metrics for LLM trustworthiness monitoring

ID	Metric	Explain-ability	Fairness	Human Agency & Control	Performance	Privacy	Robustness	Safety	Transparency	Source(s)
OM01	Perplexity	✓			✓		✓			[a]
OM02	XTREME-S				✓					[a]
OM03	Context Precision	✓			✓		✓			[b, c]
OM04	Context Recall				✓		✓			[b, c]
OM05	Noise Sensitivity				✓		✓			[c]
OM06	Response/Answer Relevancy	✓			✓		✓			[b, d, c]
OM07	Faithfulness	✓			✓		✓			[b, d]
OM08	Topic Adherence			✓	✓			✓		[c]
OM09	Tool Call Accuracy			✓			✓			[c]
OM10	Agent Goal Accuracy				✓		✓			[c]
OM11	Factual Correctness				✓		✓			[a]
OM12	Aspect Critic				✓			✓		[a]
OM13	Precision@k				✓		✓			[a]
OM14	Attack Success Rate (ASR)						✓	✓		[a]
OM15	Hughes Hallucination Evaluation Model (HHEM) Score				✓		✓	✓		[a]
OM16	Prometheus		✓						✓	[a]
OM17	SAGED		✓						✓	[a]
AM01	Plug&Play Language Model (PPLM)			✓						[29]
AM02	LIME	✓					✓		✓	[30]
AM03	SHAPley Value	✓					✓		✓	[31]
AM04	LRP score	✓							✓	[32]
AM05	ROUGE						✓	✓		[33]
AM06	BLEU						✓	✓		[34]
AM07	CLEVER score						✓	✓		[35]
JM01	G-Eval	✓	✓	✓	✓	✓	✓	✓	✓	[36] [b]
JM02	DAG (Deep Acyclic Graph)	✓	✓	✓	✓	✓	✓	✓	✓	[e, b]
JM03	Bias		✓							[b]
JM04	Stereotypes detector		✓							[f]
JM05	Sycophancy detector		✓							[f]
JM06	Output Formatting Detector			✓						[f]
JM07	Character Injection			✓				✓		[f]
JM08	Prompt Injection			✓				✓		[f]
JM09	Contextual Relevancy				✓		✓			[b, d]
JM10	RAGAS				✓		✓			[b, c]
JM11	Context Entities Recall				✓		✓			[c]
JM12	Task completion			✓	✓		✓			[b]
JM13	Coherence, Plausibility, Correctness				✓		✓			[f]
JM14	Multimodal Faithfulness/Relevance						✓			[c]
JM15	Prompt Alignment						✓			[b]
JM16	Toxicity						✓	✓		[b, d]
JM17	Harmful content detector						✓	✓		[f]
JM18	Moderation[g]		✓			✓		✓		[37] [h]
JM19	Information disclosure detector					✓				[f]

[a] https://oecd.ai/en/catalogue/metrics
[b] https://www.deepeval.com/docs/metrics-introduction
[c] https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/
[d] https://mlflow.org/docs/latest/llms/llm-evaluate/#llm-evaluation-metrics
[e] https://www.confident-ai.com/blog/how-i-built-deterministic-llm-evaluation-metrics-for-deepeval
[f] https://docs.giskard.ai/en/stable/knowledge/catalogs/test-catalog/text_generation/index.html
[g] Llama Guard is a special case of the Moderation metric with a specific Judge model and discrete output.
[h] https://www.comet.com/docs/opik/evaluation/metrics/moderation

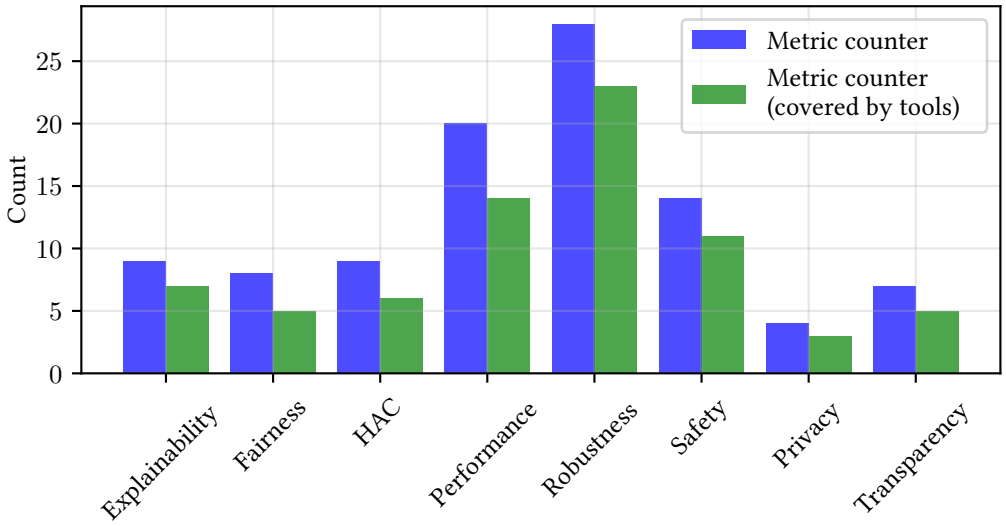

Figure 1: Metric count by objective; "covered by tools" means the number of metrics that are available in at least one of the identified tools (see Table 2)

that they should be included in the picture, as they are widely adapted and are promising in designing application-specific metrics quite easily.

In the following, we describe the judge-based metrics listed in Table 1. A very general metric that can be defined only by some evaluation criteria in natural language is *G-Eval* [36]. It relies on the interpretation of the criteria by the judge model, and thus is very judge-dependent. Initially, *G-Eval* was only using GPT-3.5 and GPT-4 for evaluation. With Prometheus [42], an effort was made to have an open-source alternative replacing the GPT models. A more recent and more deterministic approach that uses decision trees for evaluation is given by *DAG (Deep Acyclic Graph)*[3].

Some of the judge-based metrics are actually based on a more general method called *Question Answer Generation* (QAG), introduced in the context of summarization in [43]. Here, a (judge) LLM is used to first extract all claims in an LLM's response, and then the number of correct claims (identified again by the judge, usually using context or even ground truth, if available) is turned into a score or ratio. In our list, OM03, OM04, OM06, OM07, JM09, JM10 and JM11 are defined via such a procedure. With the increasing adoption of *Agentic AI* and the central role that LLMs play in such architectures, novel metrics are constructed that are either suitable for *step-by-step evaluation* of Agentic AI applications (similar to what unit tests do in software architectures), or for *trajectory evaluation*, ensuring that the decision-making process of an agent is using the appropriate tools, inputs, and memory in the appropriate spot and in the right order. Metrics OM09 and JM12 address some of these Agent-specific aspects. We refer to [44] for a more complete picture on the topic of observability in Agentic AI.

Further metrics that are not listed here are BERTScore [45], MOVERScore [46] (both embedding-based), BARTScore [47], UniEval [48], and metrics that combine statistical and model-based approaches, such as GPTScore [49] and SelfCheckGPT [50]. We decided to not mention them anymore here as it was shown in [36] that they are consistently outperformed by G-Eval (which is not surprising keeping in mind how they are constructed in comparison to G-Eval).

5. Tools

In this chapter, we discuss the available tools for LLM trustworthiness monitoring and show which of the metrics from the previous section they cover. We stress that a thorough comparison or holistic benchmark for the identified tools is not in the scope of this work.

[3] https://www.confident-ai.com/blog/how-i-built-deterministic-llm-evaluation-metrics-for-deepeval

Although the OECD catalog itself consists of more than 900 entries for tools, the filtering by the tags "technical" and "operational & monitoring" leaves only 95 tools for detailed analysis. After this analysis and the inclusion of additional tools described in detail in Section 3, we obtained 23 tools in total that we consider to be suitable for the research goals of this paper. The complete list of tools is shown in Table 2, along with the metrics that were identified.

In Table 2, the prefix O of an ID in the table indicates again that the tool was found in the OECD catalog, while the prefix A indicates that the tool was added based on additional research. Some metrics are available as a tool by themselves (e.g. LIME), so they appear on both lists: metrics and tools.

Table 2
Tools for LLM trustworthiness monitoring and corresponding metrics

ID	Tool	Metrics covered	Licence	Source
OT01	AIxploit	JM08	GNU GPL v3.0	[51]
OT02	Mindgard CLI	JM08, OM14	MIT	[a]
OT03	garak	JM08, OM14, JM16, JM07, JM19, OM07	Apache 2.0	[52]
OT04	LangBiTe	JM03, JM16	MIT	[53]
OT05	TrojAI Detect & Defend	JM08, OM14, JM19, JM16, JM17	N/A	[54]
OT06	AIShield AISpectra	JM08, OM14	N/A	[b]
OT07	PiCrystal	JM03, JM16, JM17, JM08, AM06, AM05, JM19, OM06	N/A	[c]
OT08	Giskard	JM07, JM16, JM17, JM19, JM04	Apache 2.0	[d]
OT09	Adversarial Robustness Toolbox (ART)	AM07	MIT	[e]
OT10	Lime	AM02	BSD 2-Clause	[55]
AT01	AIF360	AM02	Apache 2.0	[56]
AT02	Amazon SageMaker Clarify	OM05, OM06, OM08, OM11, AM05, JM04, JM13, JM16	Apache 2.0	[f]
AT03	Arize AI Phoenix	OM11, JM13	Elastic License 2.0 ELv2	[g]
AT04	h2oGPTe	OM06, OM07	Apache 2.0	[h]
AT05	Privacy Meter	JM19	MIT	[57]
AT06	MLflow	OM06, OM07, OM16, JM09, JM13, JM16	Apache 2.0	[i]
AT07	Promptfoo	OM06, OM07, OM03, OM04, JM09, JM13	MIT	[58]
AT08	Ragas	JM10, JM11, JM14	Apache 2.0	[59]
AT09	DeepEval	JM01, JM02, JM03, JM09, JM10, JM12, JM15, JM16	Apache 2.0	[60]
AT10	AIX360	AM02, AM03	Apache 2.0	[61]
AT11	SHAP	AM03	MIT	[j]
AT12	Opik	OM03, OM04, OM06, OM07, JM01, JM18	Apache 2.0	[k]
AT13	Evidently	OM07, JM01, JM03, JM06, JM09, JM13, JM16, JM19	Apache 2.0	[l]

[a] https://github.com/Mindgard/cli [b] https://boschaishield.com/ [c] https://www.quantpi.com/
[d] https://github.com/Giskard-AI/giskard [e] https://github.com/Trusted-AI/adversarial-robustness-toolbox
[f] https://aws.amazon.com/de/sagemaker-ai/clarify/ [g] https://arize.com/docs/phoenix
[h] https://h2o.ai/platform/enterprise-h2ogpte/model-validation/ [i] https://mlflow.org/docs/3.4.0/genai/
[j] https://shap.readthedocs.io/en/latest/index.html [k] https://www.comet.com/docs/opik/
[l] https://docs.evidentlyai.com

The coverage of the metrics by the tools is shown in Figure 2 and discussed in the following.

The first result is that there are several tools available that can be used for LLM trustworthiness monitoring, and that each of those covers at least one of the metrics that we identified. Conversely, we observe that 11 metrics are not yet available in the tools that we identified and thus may not be applied

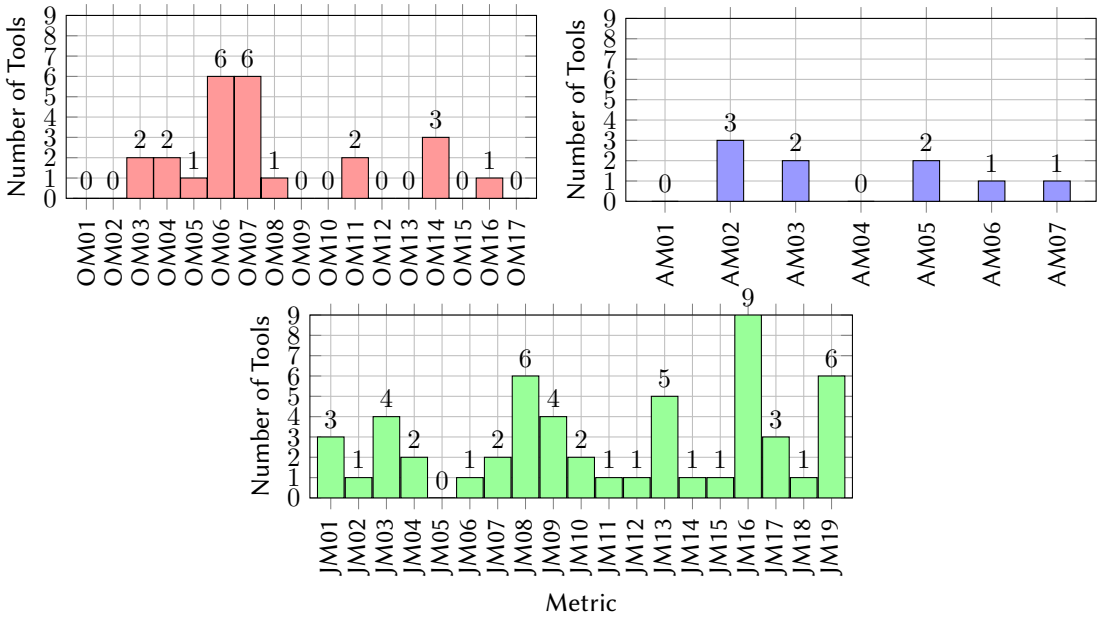

Figure 2: Number of tools covering each metric; the individual diagrams correspond to the three subsets of metrics from Table 1

easily. None of the tools contains all or at least almost all of the identified metrics, so practitioners will likely have to rely on several tools to cover a broad range of trustworthiness aspects in their application – at least in the near future.

What can also be seen directly from Figure 2 is that judge-based metrics are covered much more often than the others. Specifically, the most often implemented metrics are Toxicity (JM16), and then Prompt Injection (JM08), Information Disclosure Detector (JM19), Response/Answer Relevancy (OM06) and Faithfulness (OM07). All of these metrics rely on a judge approach.

6. Discussion

Metric gaps and limitations. The fact that only four metrics are available with the objective of *Privacy* highlights a significant gap in current research, which is further confirmed by the generic nature of two of these metrics. The underrepresentation of privacy objectives is concerning, particularly given that the identified privacy metrics are judge-based and not yet included in the OECD catalog. Moreover, the OECD catalog lacks *Transparency* metrics and only includes two *Fairness* metrics, emphasizing the importance of incorporating additional metrics beyond the catalog.

Judge-based metrics, however, exhibit a notable lack of coverage in the objectives of *Explainability* and *Transparency*, likely due to fact that judges only work with the given instructions and context, and are not able to explain or reveal anything beyond that (except for what they "saw" during training).

The metrics presented in this work are subject to certain limitations. For instance, some metrics are *task-specific* (e.g., ROUGE for summarization or CLEVER score and LIME for classification), while others are *scoped* for other reasons such as relying on benchmarks or datasets that may not accurately reflect real-world deployment scenarios. Most metrics also inherently depend on *randomization*, such as training-test splits. Additionally, some metrics may be *costly* in terms of LLM usage, either due to large data set processing requirements or the need for GPU-intensive model invocations, as for the LLMs used in the judge-based metrics.

Tool limitations. Concerning the tools, it is noteworthy that despite the OECD catalog's extensive list, many tools are not technical or suitable for monitoring, resulting in a relatively small number of relevant tools for our purpose.

We observed that the coverage of metrics by available tools is limited, with only a few metrics being well-represented. This suggests a lag in the translation of research into usable software, potentially due to tool developers' focus on certain aspects of trustworthiness, such as prompt injection, at the expense of less prominent issues.

Our emphasis on using only publicly available information and avoiding black boxes for transparency reasons may have implications for the coverage of proprietary tools. They are not yet covered to an acceptable degree in our work due to restricted documentation, despite the fact that some proprietary tools have accessible documentation about the available metrics before a paywall. However, *open-source* software tools, which comprise the majority of our list, are expected to have much better coverage.

7. Conclusion

In this work, we have provided an overview of the current state of trustworthiness monitoring, including collections of available metrics and tools, to serve as a guideline for researchers and practitioners. Our compilation highlights the trustworthiness objectives and corresponding metrics, as well as the tools that assess them, aiming to identify areas with insufficient coverage for researchers and facilitate the improvement of trustworthiness in specific aspects. Practitioners can use the compilations from Table 1 and Table 2 as a reference.

As the landscape of available tools continues to rapidly evolve, their capabilities are constantly improving, too. Additionally, the utilization of AI agents – which can anyway be seen as a means to more responsible AI [62] – will expand the monitoring component to include further aspects, in particular in terms of safety and tracing. The recently proposed paradigm of *AgentOps* [44] addresses this aspect, and will very likely be followed by others soon. Due to this very dynamic situation, the snapshot described here is likely to be outdated soon.

What is missing is a general framework that enables easy selection of applicable metrics for specific settings or applications. So far, it is hard to even find (not to mention implement) a set of metrics that yields an appropriate coverage of all trustworthiness aspects in the specific setting. A possible next step is the creation of an adaptive framework that can dynamically adjust to evolving trustworthiness requirements and incorporate new metrics and tools as they become available. This demands for standardized tooling and interoperability protocols to enable seamless integration and comparison of different trustworthiness monitoring solutions. Thus, we encourage researchers and practitioners to collaborate on the development of standardized, adaptive, and interoperable trustworthiness monitoring frameworks.

Acknowledgments

This work has been funded by the Fraunhofer Cluster of Excellence Cognitive Internet Technologies and the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D).

Declaration on Generative AI

During the preparation of this work, the authors used DeepL and Claude in order to: Grammar, spelling check and producing code for generating the visuals. Further, one of the authors used Llama 3 for improving the wording of individual paragraphs. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. URL: https://arxiv.org/abs/2306.05685. doi:10.48550/arXiv.2306.05685. arXiv:2306.05685.

[2] C. Sanderson, E. Schleiger, D. Douglas, P. Kuhnert, Q. Lu, Resolving ethics trade-offs in implementing responsible AI, arXiv preprint arXiv:2401.08103 (2024).

[3] D. Kreuzberger, N. Kühl, S. Hirschl, Machine Learning Operations (MLOps): Overview, Definition, and Architecture, IEEE access 11 (2023) 31866–31879.

[4] L. Helmer, C. Martens, D. Wegener, M. Akila, D. Becker, S. Abbas, Towards Trustworthy AI Engineering - A Case Study on integrating an AI audit catalog into MLOps processes, in: Proceedings of the 2nd International Workshop on Responsible AI Engineering, RAIE '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1–7. URL: https://doi.org/10.1145/3643691.3648584. doi:10.1145/3643691.3648584.

[5] L. Helmer, A. Kerbel, C. Martens, C. Temath, D. Wegener, A. Zimmermann, A. Zorn, Machine Learning Operations (MLOps): Grundlagen, Chancen und Herausforderungen beim MLOps-Einsatz in Unternehmen, 2024. URL: https://doi.org/10.24406/publica-2962.

[6] OECD, Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems, OECD Digital Economy Papers No. 312 (2021). doi:10.1787/008232ec-en.

[7] M. Poretschkin, A. Schmitz, M. Akila, L. Adilova, D. Becker, A. B. Cremers, D. Hecker, S. Houben, M. Mock, J. Rosenzweig, et al., Guideline for Trustworthy Artificial Intelligence–AI Assessment Catalog, 2023. URL: https://arxiv.org/abs/2307.03681. doi:https://doi.org/10.48550/arXiv.2307.03681. arXiv:2307.03681.

[8] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al., TrustLLM: Trustworthiness in Large Language Models, arXiv preprint arXiv:2401.05561 (2024). URL: https://arxiv.org/abs/2401.05561.

[9] Anthropic, Anthropic's responsible scaling policy, 2023. URL: https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

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

[11] N. Brandizzi, H. Abdelwahab, A. Bhowmick, L. Helmer, B. J. Stein, P. Denisov, Q. Saleem, M. Fromm, M. Ali, R. Rutmann, et al., Data Processing for the OpenGPT-X Model Family, arXiv preprint arXiv:2410.08800 (2024).

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[13] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al., The Pile: An 800GB dataset of diverse text for language modeling, arXiv preprint arXiv:2101.00027 (2020).

[14] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only, arXiv preprint arXiv:2306.01116 (2023).

[15] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, E. Grave, CCNet: Extracting high quality monolingual datasets from web crawl data, arXiv preprint arXiv:1911.00359 (2019).

[16] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, N. Carlini, Deduplicating training data makes language models better, arXiv preprint arXiv:2107.06499 (2021).

[17] J. Abadji, P. O. Suarez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, arXiv preprint arXiv:2201.06642 (2022).

[18] G. Penedo, H. Kydlíček, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, T. Wolf, et al., The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, arXiv preprint arXiv:2406.17557

(2024).
[19] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al., Instruction Tuning for Large Language Models: A Survey, arXiv preprint arXiv:2308.10792 (2023).
[20] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, in: International Conference on Machine Learning, PMLR, 2023, pp. 22631–22648.
[21] J. Abadji, P. J. O. Suárez, L. Romary, B. Sagot, Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus, in: CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora, 2021.
[22] T. Jansen, Y. Tong, V. Zevallos, P. O. Suarez, Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data, arXiv preprint arXiv:2212.10440 (2022).
[23] S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, N. Dziri, Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, arXiv preprint arXiv:2406.18495 (2024).
[24] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al., DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models., in: NeurIPS, 2023.
[25] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).
[26] J. Liu, H. Chen, J. Shen, K.-K. R. Choo, Faircompass: Operationalising fairness in machine learning, 2023. doi:10.1109/TAI.2023.3348429.
[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.
[28] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment, 2024. URL: https://arxiv.org/abs/2308.05374. arXiv:2308.05374.
[29] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, R. Liu, Plug and Play Language Models: A Simple Approach to Controlled Text Generation, 2020. URL: https://arxiv.org/abs/1912.02164. doi:10.48550/arXiv.1912.02164. arXiv:1912.02164.
[30] K. Zhu, Q. Zang, S. Jia, S. Wu, F. Fang, Y. Li, S. Gavin, T. Zheng, J. Guo, B. Li, et al., Lime: Less is more for mllm evaluation, arXiv preprint arXiv:2409.06851 (2024).
[31] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.
[32] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE 10 (2015) 1–46. doi:10.1371/journal.pone.0130140.
[33] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[34] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.
[35] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, L. Daniel, Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, 2018. URL: https://arxiv.org/abs/1801.10578. doi:10.48550/arXiv.1801.10578. arXiv:1801.10578.
[36] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, 2023. URL: https://arxiv.org/abs/2303.16634. doi:10.48550/arXiv.2303.

16634. arXiv:2303.16634.

[37] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, 2023. URL: https://arxiv.org/abs/2312.06674. doi:10.48550/arXiv.2312.06674. arXiv:2312.06674.

[38] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, I. Arawjo, Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences, 2024. URL: https://arxiv.org/abs/2404.12272. doi:10.48550/arXiv.2404.12272. arXiv:2404.12272.

[39] S. Chern, E. Chern, G. Neubig, P. Liu, Can Large Language Models be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate, 2024. URL: https://arxiv.org/abs/2401.16788. doi:10.48550/arXiv.2401.16788. arXiv:2401.16788.

[40] H. Huang, Y. Qu, X. Bu, H. Zhou, J. Liu, M. Yang, B. Xu, T. Zhao, An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4, 2024. URL: https://arxiv.org/abs/2403.02839. doi:10.48550/arXiv.2403.02839. arXiv:2403.02839.

[41] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, D. Hupkes, Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges, 2024. URL: https://arxiv.org/abs/2406.12624. doi:10.48550/arXiv.2406.12624. arXiv:2406.12624.

[42] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, M. Seo, Prometheus: Inducing Fine-grained Evaluation Capability in Language Models, 2024. URL: https://arxiv.org/abs/2310.08491. doi:10.48550/arXiv.2310.08491. arXiv:2310.08491.

[43] A. Wang, K. Cho, M. Lewis, Asking and Answering Questions to Evaluate the Factual Consistency of Summaries, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5008–5020. URL: https://aclanthology.org/2020.acl-main.450/. doi:10.18653/v1/2020.acl-main.450.

[44] L. Dong, Q. Lu, L. Zhu, AgentOps: Enabling Observability of LLM Agents, 2024. URL: https://arxiv.org/abs/2411.05285. arXiv:2411.05285.

[45] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2020. URL: https://arxiv.org/abs/1904.09675. doi:10.48550/arXiv.1904.09675. arXiv:1904.09675.

[46] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, S. Eger, MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance, 2019. URL: https://arxiv.org/abs/1909.02622. doi:10.48550/arXiv.1909.02622. arXiv:1909.02622.

[47] W. Yuan, G. Neubig, P. Liu, BARTScore: Evaluating Generated Text as Text Generation, 2021. URL: https://arxiv.org/abs/2106.11520. doi:10.48550/arXiv.2106.11520. arXiv:2106.11520.

[48] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, J. Han, Towards a Unified Multi-Dimensional Evaluator for Text Generation, 2022. URL: https://arxiv.org/abs/2210.07197. doi:10.48550/arXiv.2210.07197. arXiv:2210.07197.

[49] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, GPTScore: Evaluate as You Desire, 2023. URL: https://arxiv.org/abs/2302.04166. doi:10.48550/arXiv.2302.04166. arXiv:2302.04166.

[50] P. Manakul, A. Liusie, M. Gales, SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9004–9017. URL: https://aclanthology.org/2023.emnlp-main.557/. doi:10.18653/v1/2023.emnlp-main.557.

[51] AINTrust, Aixploit, 2025. URL: https://github.com/AINTRUST-AI/aixploit, commit: 9d4c59d2f090b23dd44bbf4df91ae8f1a76f0d20.

[52] L. Derczynski, E. Galinkin, J. Martin, S. Majumdar, N. Inie, garak: A Framework for Security Probing Large Language Models, 2024. URL: https://arxiv.org/abs/2406.11036. doi:10.48550/arXiv.2406.11036. arXiv:2406.11036.

[53] S. Morales, R. Clarisó, J. Cabot, A dsl for testing llms for fairness and bias, in: Proceedings of the

ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems, 2024, pp. 203–213.
[54] L. Weiner, Troj.ai detect & defend, https://www.troj.ai/products/, 2025. Accessed: 2025-07-26.
[55] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.
[56] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, 2018. URL: https://arxiv.org/abs/1810.01943.
[57] S. Kumar, R. Shokri, Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning, in: Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs), 2020.
[58] I. Webster, M. D'Angelo, S. Klein, G. Zang, promptfoo, 2025. URL: https://github.com/promptfoo/promptfoo.
[59] ExplodingGradients, Ragas: Supercharge your llm application evaluations, https://github.com/explodinggradients/ragas, 2024.
[60] J. Ip, K. Vongthongsri, deepeval, 2025. URL: https://github.com/confident-ai/deepeval.
[61] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2019. URL: https://arxiv.org/abs/1909.03012.
[62] Q. Lu, L. Zhu, X. Xu, Z. Xing, S. Harrer, J. Whittle, Towards Responsible Generative AI: A Reference Architecture for Designing Foundation Model Based Agents, in: 2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C), 2024, pp. 119–126. doi:10.1109/ICSA-C63560.2024.00028.