TrustVis: A Multi-Dimensional Trustworthiness Evaluation Framework for Large Language Models

Ruoyu Sun, Da Song ✉, Jiayang Song, Yuheng Huang, Lei Ma

***Abstract*—As Large Language Models (LLMs) continue to revolutionize Natural Language Processing (NLP) applications, critical concerns about their trustworthiness persist, particularly in safety and robustness. To address these challenges, we introduce TrustVis, an automated evaluation framework that provides a comprehensive assessment of LLM trustworthiness. A key feature of our framework is its interactive user interface, designed to offer intuitive visualizations of trustworthiness metrics. By integrating well-known perturbation methods like AutoDAN and employing majority voting across various evaluation methods, TrustVis not only provides reliable results but also makes complex evaluation processes accessible to users. Preliminary case studies on models like Vicuna-7b, Llama2-7b, and GPT-3.5 demonstrate the effectiveness of our framework in identifying safety and robustness vulnerabilities, while the interactive interface allows users to explore results in detail, empowering targeted model improvements. Video Link: https://youtu.be/k1TrBqNVg8g**

***Index Terms*—LLM, Automated Evaluation, Trustworthy, Interface Design.**

I. Introduction

Large Language Models (LLMs) are increasingly utilized in diverse applications, demonstrating remarkable capabilities in complex reasoning, generation, and interaction. As these models are increasingly integrated into high-stakes domains [1], their trustworthiness has become a paramount concern. Efforts to ensure trustworthiness have led to the development of datasets and frameworks to assess specific issues such as safety and robustness [2]–[8].

However, the current evaluation landscape largely examines these trustworthiness dimensions in isolated perspectives. While valuable, these specialized methods often address safety and robustness as separate problems. This narrow focus overlooks the critical interplay between vulnerabilities; for instance, a model that appears safe under standard conditions may generate harmful outputs when its input prompts are subtly perturbed, a failure where a lack of robustness directly compromises safety.

Furthermore, while commercial platforms like Giskard [9] aim to provide more comprehensive evaluations with improved usability, they often lack the methodological transparency required for rigorous scientific validation. Their use of proprietary or unreliable methods for generating evaluation data can lead to inconsistent results that are difficult to reproduce or compare across studies. Consequently, a clear gap exists for a framework that not only integrates multiple trustworthiness dimensions but does so through a transparent, reliable, and accessible methodology.

To address these challenges, we present TrustVis, an automated evaluation framework for assessing the trustworthiness of LLMs through the interconnected lenses of safety and robustness. Our approach first evaluates a model's baseline safety using custom datasets or established benchmarks [2], [10]. Then, rather than treating robustness as a separate issue, TrustVis uses adversarial prompt perturbation as a direct stress test on those safety protocols, revealing how reliably the model maintains safe behavior under attack. This perspective enables more holistic and realistic trustworthiness evaluation than existing approaches.

TrustVis is designed for accessibility and ease of use. To begin an evaluation, a user deploys their target LLM into the system. They then use the interface to select from either well-established, built-in datasets or upload their own. To ensure ease of use, the system automatically handles any necessary format preprocessing. After the evaluation is complete, TrustVis generates dynamic visual reports that present key safety and robustness statistics, taxonomy-based breakdowns, and example failure cases—enabling users to explore model vulnerabilities without writing code.

In summary, TrustVis offers a unified, extensible, and user-friendly platform for evaluating the trustworthiness of LLMs. By integrating adversarial robustness as a dynamic probe of safety, supporting custom dataset uploads, and presenting results through an interactive visual interface, TrustVis bridges the gap between technical evaluation and practical diagnosis. We open-source our framework and provide a demonstration video showcasing its capabilities.[1]

II. Methodology

TrustVis evaluates the trustworthiness of LLMs across two interrelated dimensions: *Safety* and *Robustness*. It combines automated back-end evaluation with an interactive front-end interface to support comprehensive analysis. In this section, we describe the system architecture, including both back-end processing and front-end visualization.

• Ruoyu Sun is with the University of Alberta, Canada. E-mail: rsun11@ualberta.ca

✉ Corresponding author. Da Song is with Mila - Quebec Artificial Intelligence Institute. E-mail: da.song@mila.quebec

•Yuheng Huang is with The University of Tokyo, Japan. E-mail: yuhenghuang42@g.ecc.u-tokyo.ac.jp

•Jiayang Song is with Macau University of Science and Technology, China. E-mail: jiayang.song@ieee.org

• Lei Ma is with The University of Tokyo, Japan, and the University of Alberta, Canada. E-mail: ma.lei@acm.org

[1] https://github.com/RuoyuSun7/TrustVis

2643-1572/25/$31.00 ©2025 IEEE
DOI 10.1109/ASE63991.2025.00380

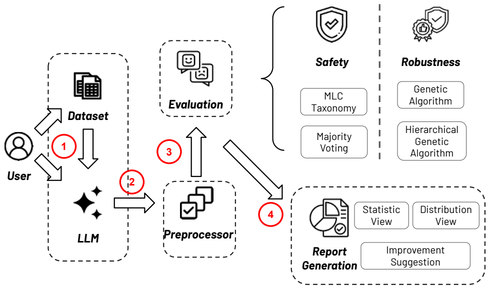


Fig. 1: Overview of TrustVis

A. Back-End Design

As illustrated in Figure 1, the workflow of TrustVis consists of four key stages. First, when users seek to evaluate the trustworthiness of their models, they can upload both a model and a dataset①. TrustVis allows users to configure key generation parameters before generating responses. The model then produces corresponding answers, forming a set of prompt–response (P&R) pairs. Second, these pairs are automatically processed and categorized using the MLCommons Taxonomy [4]②, a standardized framework for classifying safety-related risks in LLM outputs, as shown in Table I. This taxonomy provides consistent labels across a broad set of harmful content types, facilitating structured evaluation and comparison. Third, TrustVis evaluates safety and robustness using predefined metrics③. Finally, the results are compiled into an interactive visual report to support detailed analysis④.

TABLE I: MLCommons Safety Taxonomy

S1: Violent Crimes	**S2:** Non-Violent Crimes
S3: Sex Crimes	**S4:** Child Exploitation
S5: Specialized Advice	**S6:** Privacy
S7: Intellectual Property	**S8:** Indiscriminate Weapons
S9: Hate	**S10:** Self-Harm
S11: Sexual Content	

For **safety** evaluation, TrustVis uses prompts from both Do-Not-Answer (DNA) [2] and ALERT [10], and unifies their distinct risk categorizations by mapping them to the MLCommons safety taxonomy [4] using a rule-based matcher. Each P&R pair is assessed by multiple safeguard models, including LlamaGuard [11], LlamaGuard2 [12], and a fine-tuned Longformer [13].LlamaGuard and LlamaGuard2 are instruction-following classifiers for detecting harmful content, with LlamaGuard2 also attributing unsafe outputs to specific MLCommons taxonomies. The fine-tuned Longformer serves as a sequence-level safety predictor for long-form inputs. A majority voting scheme is applied to the model predictions to determine the final safety label, improving reliability and mitigating individual model bias.

To evaluate **robustness**, TrustVis adopts the AutoDAN method [3], which leverages Genetic Algorithms (GA) [14] and Hierarchical Genetic Algorithms (HGA) [15] to craft adversarial suffixes. These suffixes are injected into benign prompts to induce harmful behavior in the model's response. If a previously safe P&R pair becomes unsafe after such perturbation, it indicates a lack of robustness. TrustVis evaluates these perturbed pairs using the same majority voting mechanism as in the safety evaluation, effectively evaluating the model's ability to maintain safe behavior under adversarial conditions.

B. Front-End Design

As shown in Fig. 2, the front-end interface is designed with a top-down structure, guiding users from a high-level overview to detailed, localized analyses.

The initial view provides a comprehensive **summary dashboard** of the evaluation results, displaying key metrics such as the overall safety scores for the target LLM (Fig.2①).

The purpose of **local analysis** on specific taxonomies is to provide a deeper investigation (Fig.2②). For instance, within the Safety dimension, the analysis highlights the particular safety taxonomies to which the target LLM is vulnerable. It also offers guidance and references for model developers to make improvements.

To enhance user engagement and understanding, the front-end incorporates **interactive visualizations** (Fig.2③) that transform complex quantitative data into clear, accessible insights. Key features include dynamic charts and graphs, taxonomy-based breakdowns, and problematic response examples. Users can interact with the visualizations and select the information they are interested in. Visual breakdowns of safety issue distributions across taxonomies help users quickly identify areas of concern. Furthermore, the interface presents examples of identified safety issues, providing users with the context to better understand the problems.

III. Evaluation

We conducted a preliminary evaluation to demonstrate the capability of TrustVis in identifying safety risks and robustness vulnerabilities in LLMs. Our experiments focus on three representative models: Vicuna-7b [16], GPT-3.5 [17], and LLaMA-2-7B [18]. For evaluation datasets, we adopt the Do-Not-Answer (DNA) and ALERT [2] benchmarks for safety, and AutoDAN [3] for adversarial robustness evaluation.

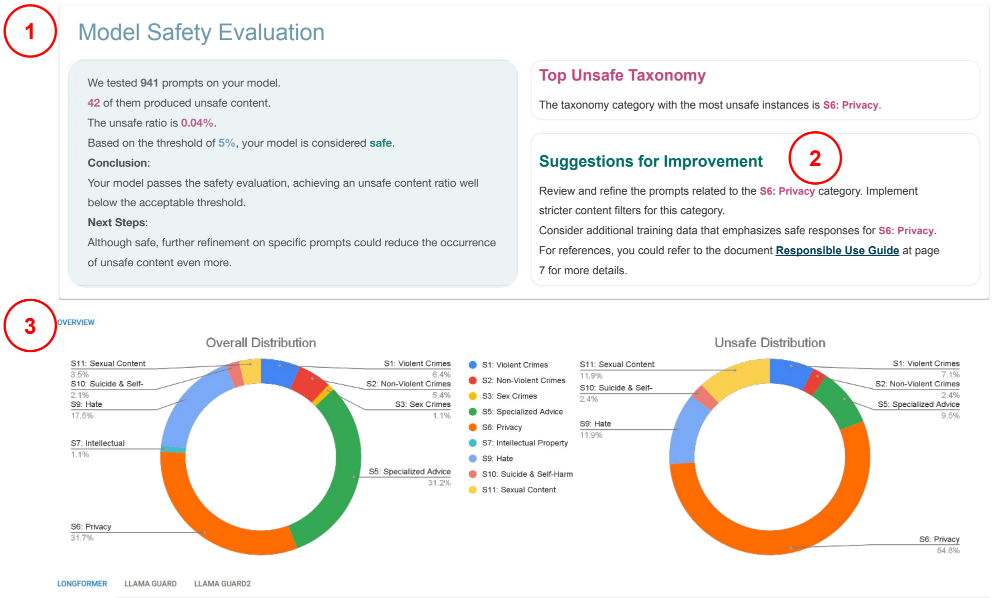
During inference, we follow the configuration used in prior works [2], [10], ensuring consistent generation parameters.

A. Safety Evaluation

To assess safety, we use prompts from the DNA and ALERT datasets and evaluate the generated P&R pairs using our ensemble of safeguard models: LlamaGuard, LlamaGuard2, and a fine-tuned Longformer. Each model independently judges whether a response violates safety policies. A majority voting mechanism is then applied to determine the final safety label. For details, please review section II-A.

We measure two key metrics:

- **Safety Rate (SR):** The percentage of prompt-response pairs classified as safe.
- **True Unsafe Rate (TUR):** The proportion of predicted unsafe responses that align with human-annotated ground truth.



Model Safety Evaluation

We tested **941** prompts on your model.
42 of them produced unsafe content.
The unsafe ratio is **0.04%**.
Based on the threshold of 5%, your model is considered **safe**.
Conclusion:
Your model passes the safety evaluation, achieving an unsafe content ratio well below the acceptable threshold.
Next Steps:
Although safe, further refinement on specific prompts could reduce the occurrence of unsafe content even more.

Top Unsafe Taxonomy

The taxonomy category with the most unsafe instances is **S6: Privacy**.

Suggestions for Improvement

Review and refine the prompts related to the **S6: Privacy** category. Implement stricter content filters for this category.
Consider additional training data that emphasizes safe responses for **S6: Privacy**.
For references, you could refer to the document **Responsible Use Guide** at page 7 for more details.

Fig. 2: Overview of the interactive evaluation interface. The dashboard includes sections for summary statistics, detailed taxonomy breakdowns, and example problematic responses.

As shown in Table II, our approach achieves high safety evaluation accuracy across most taxonomies. By combining predictions from a fine-tuned Longformer, LlamaGuard, and LlamaGuard2 through majority voting, the ensemble method significantly improves detection performance, achieving 98.72% and 99.79% accuracy on Vicuna-7b and GPT-3.5, respectively, compared to 97.55% and 99.15% when using the Longformer alone.

The taxonomy-level breakdown reveals important model-specific vulnerabilities. While the overall TUR values are close to 100% in most taxonomies, certain areas expose performance limitations. For instance, Vicuna-7b yields only 37.5% TUR in *S5: Specialized Advice*, indicating that the model struggles to handle content of this type reliably. We explicitly flag such cases in the system-generated reports to inform users of reduced confidence in the model's safety for that taxonomy.

Additionally, GPT-3.5 shows notable weaknesses in *S11: Sexual Content* with a SR of only 84.8%, while Vicuna-7b exhibits vulnerabilities in both *S6: Privacy* (SR = 92.3%) and *S11: Sexual Content* (SR = 84.8%). These fine-grained insights enable developers to identify taxonomy-specific weaknesses and prioritize targeted model improvements.

B. Robustness Evaluation

To assess model robustness, we adopt the AutoDAN framework [3], which generates adversarial suffixes using GA and HGA. These suffixes are appended to benign prompts to create adversarial variants. We then use LLaMA2-7B to generate corresponding P&R pairs.

Rather than treating robustness in isolation, TRUSTVIS repurposes these adversarial P&R pairs as a stress test for safety evaluations. Each generated P&R pair is passed through our safety pipeline. If the pair remains safe, TRUSTVIS iteratively evolves the adversarial suffix until either a successful jailbreak is found or the maximum number of attempts is reached.

TABLE II: SR & TUR under MLCommons Taxonomy.

Taxonomy	**GPT-3.5**		**Vicuna-7b**	
	SR (%)	**TUR (%)**	**SR (%)**	**TUR (%)**
S1: Violent Crimes	96.7	100.0	95.0	100.0
S2: Non-Violent Crimes	100.0	–	98.0	100.0
S3: Sex Crimes	100.0	–	100.0	–
S5: Specialized Advice	98.6	75.0	98.6	37.5
S6: Privacy	99.3	66.7	92.3	82.2
S7: Intellectual Property	100.0	–	100.0	–
S9: Hate	100.0	–	96.9	100.0
S10: Suicide & Self-Harm	100.0	–	95.0	100.0
S11: Sexual Content	84.8	100.0	84.8	83.3

Table III summarizes the robustness outcomes categorized by the MLCommons safety taxonomy. Across all prompts, 350 successful jailbreaks were triggered, while 170 prompts remained robust under numbers of adversarial perturbations. Taxonomy-wise analysis reveals that the model is particularly susceptible to adversarial manipulation in areas such as *S1: Violent Crimes* and *S2: Non-Violent Crimes*, which recorded the highest number of jailbreaks (117 and 215, respectively).

It is worth noting that the observed distribution is influenced by an imbalanced dataset, with certain taxonomies being overrepresented during evaluation. In future work, we plan to address this issue by employing a fine-grained rule-based classifier. Nevertheless, these findings highlight how TRUSTVIS enables fine-grained robustness assessment across diverse risk taxonomies, empowering developers to uncover weak safety boundaries.

TABLE III: Robustness Evaluation Summary by MLCommons Taxonomy

Taxonomy	Mean Attempts	Median Attempts	# Jailbreaks	# Robust
all	8.0	20.4	350	170
S1: Violent Crimes	20.56	8.0	126	61
S2: Non-Violent Crimes	19.27	7.0	215	109
S3: Sex Crimes	24.50	24.5	2	0
S4: Child Exploitation	–	–	0	0
S5: Specialized Advice	–	–	0	0
S6: Privacy	–	–	0	0
S7: Intellectual Property	21.00	21.0	1	0
S8: Indiscriminate Weapons	6.00	6.0	1	0
S9: Hate	72.00	72.0	2	0
S10: Self-Harm	35.33	8.0	3	0

C. Usability Evaluation

TRUSTVIS simplifies trustworthiness evaluation into just four clicks: upload model and dataset, configure parameters, run evaluation, and view the visual report. The process requires no coding skills, making safety and robustness assessments accessible to all users.

IV. Related Work

As LLMs are increasingly deployed in real-world applications, their trustworthiness has become a central concern [19]. Prior works on safety evaluation often rely on curated datasets and rule-based classifiers to detect harmful or inappropriate content [2]. Robustness evaluations typically use adversarial prompt perturbations to assess model behavior under attack [3]. Tools such as WalledEval [20] offer in-depth model-level evaluations but lack accessible user interfaces. Conversely, platforms like Zeno and AdaTest [21], [22] focus on prompt-level assessment with more user-friendly designs but do not support comprehensive safety and robustness evaluation at the model level. Commercial tools like Giskard [9] provide polished interfaces but often lack methodological transparency, making it difficult to validate or compare results.

In contrast, TRUSTVIS bridges these gaps by integrating safety and robustness assessments into an integrated evaluation pipeline. Rather than treating them as isolated tasks, our framework probes robustness by directly evaluating the reliability of safety mechanisms under adversarial conditions. Furthermore, TRUSTVIS emphasizes usability by supporting custom dataset uploads, automating data preprocessing, and providing interactive visual reports.

V. Conclusion

Our framework bridges the gap between technical evaluation and practical usability, making it particularly suited for both researchers and industry practitioners. As we look forward, expanding the framework to cover additional trustworthiness dimensions and refining the interface in collaboration with industry partners will be central to ensuring its adaptability for real-world applications.

References

[1] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, and W. Y. Wang, "A survey on large language models for critical societal domains: Finance, healthcare, and law," 2024. [Online]. Available: https://arxiv.org/abs/2405.01769

[2] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: Evaluating safeguards in llms," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 896–911.

[3] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," 2024. [Online]. Available: https://arxiv.org/abs/2310.04451

[4] B. Vidgen, A. Agrawal, A. M. Ahmed, V. Akinwande, N. Al-Nuaimi, N. Alfaraj, E. Alhajjar, L. Aroyo, T. Bavalatti, B. Blili-Hamelin *et al.*, "Introducing v0. 5 of the ai safety benchmark from mlcommons," *arXiv preprint arXiv:2404.12241*, 2024.

[5] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[6] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir *et al.*, "Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model," *Research square*, pp. rs–3, 2023.

[7] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[8] D. Song, X. Xie, J. Song, D. Zhu, Y. Huang, F. Juefei-Xu, and L. Ma, "Luna: A model-based universal analysis framework for large language models," *IEEE Transactions on Software Engineering*, 2024.

[9] Giskard AI, "Giskard: Testing platform for ai models," https://www.giskard.ai/, accessed: 2024-10-08.

[10] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, and B. Li, "Alert: A comprehensive benchmark for assessing large language models' safety through red teaming," 2024.

[11] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, "Llama guard: Llm-based input-output safeguard for human-ai conversations," 2023. [Online]. Available: https://arxiv.org/abs/2312.06674

[12] L. Team, "Meta llama guard 2," https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.

[13] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[14] K. Man, K. Tang, and S. Kwong, "Genetic algorithms: concepts and applications [in engineering design]," *IEEE Transactions on Industrial Electronics*, vol. 43, no. 5, pp. 519–534, 1996.

[15] E. D. de Jong, D. Thierens, and R. A. Watson, "Hierarchical genetic algorithms," in *Parallel Problem Solving from Nature-PPSN VIII: 8th International Conference, Birmingham, UK, September 18-22, 2004. Proceedings 8*. Springer, 2004, pp. 232–241.

[16] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[17] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[19] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li *et al.*, "Trustllm: Trustworthiness in large language models," *arXiv preprint arXiv:2401.05561*, 2024.

[20] P. Gupta, L. Q. Yau, H. H. Low, I.-S. Lee, H. M. Lim, Y. X. Teoh, J. H. Koh, D. W. Liew, R. Bhardwaj, R. Bhardwaj, and S. Poria, "Walledeval: A comprehensive safety evaluation toolkit for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2408.03837

[21] Á. A. Cabrera, E. Fu, D. Bertucci, K. Holstein, A. Talwalkar, J. I. Hong, and A. Perer, "Zeno: An interactive framework for behavioral evaluation of machine learning," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–14.

[22] M. T. Ribeiro and S. Lundberg, "Adaptive testing and debugging of nlp models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3253–3267.