Paper Review: MULTITRUST: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models

By: Lam Nguyen

1. SUMMARY

MultiTrust is the first benchmark for evaluating the trustworthiness of a Multimodal Large Language Model (MLLM).

The five metrics that are evaluated are:

- Trustfulness
- Safety
- Robustness
- Fairness
- Privacy

The strategy that will be used to address both multimodal risks and cross-modal impacts, encompassing 32 different tasks with a curated dataset.

MLLMs still struggle with the perception of visually confusing images and are vulnerable to jailbreaking and adversarial attacks. MLLMs are also inclined to disclose privacy in text and reveal bias even when paired with irrelevant images.

A Scalable toolbox for trustworthiness was also released.

2. STRENGTHS

- Previously no standardized benchmark with was established to evaluate the trustworthiness of an answer given. These forms of metrics are very useful for rating an answer and people tend to value these types of benchmarks like they value amazon Ratings and Reviews.
- The Trustworthiness ratings are very detailed and the amount and type of metrics provided is detailed and comprehensive.

3. WEAKNESSES

- The weakness of the MultiTrust rating is that it might actually be too complex to be used widely by the public. If there is a way to simplify the rating down to a single score that can be used to rate an answer, this might be practical, even if it does remove some of the granularity that the original ratings and metrics provide.
- With many of these trustworthiness scores, there is an element of subjectivity in the rating depending on the culture one resides in and the individual person even designing the Metrics. This variability between different individuals and different cultures could be accounted for. For example, is nudity actually a bad thing? Maybe saying that that the human body is shameful should be the real problem?

4. TECHNICAL EXTENSIONS

This might be beyond the scope of the subject matter but it might be useful to simplify the metric scores even more down to a very simple rating that can be used to rate the trustworthiness. Possibly distill it down to one number or star rating…

Other extensions that could be made to this work is to establish metrics or at least consider in the calculations the culture that this MultiTrust Benchmark will be used.

5. OVERALL REVIEW

MultiTrust is a very useful set of metrics to definitively determine the moral rating of an answer given by an MLLM. The strength of this method is that it is very detailed.

The weakness is also the level of detail and the need to probably distill it down to an accurate single rating in order to make it useful for the public.

So in the future, an extension of this work would be to determine a methodology of distilling the Trust metrics down to a simple single score.