Paper Review: MLLMGUARD: A Multi-Dimensional Safety Evaluation Suite for Multimodal Large Language Models

By: Lam Nguyen

1. SUMMARY

MLLMGuard is a method of evaluating the safety of an answer given or a question asked involving a Large Language Model (LLM). The factors to be considered in evaluation are:

- Privacy
- Bias
- Toxicity
- Truthfulness
- Legality

Information given by a publicly available LLM need to protect privacy, be unbiased, not toxic, true and legal.

The MLLMGuard suite consists of a curated evaluation dataset, inference utilities, and a lightweight evaluator for automatically judging the safety of an answer.

Modern LLMs have been extended from a single text modality to being generalized to work over many different types of text material making them much more versatile then in the past. But because of this versatility, there is a much greater risk that sensitive or dangerous data will be given out. Previous methods of censoring MLLMs involve the use of the GPT-4V model to directly rate the safety of a response. However, GPT-4V is biased towards its own given answers. Human annotators can also be used but this is extremely costly and tedious. Another downside to previous censorship methods is that they are only focused on the English Language.

The evaluation dataset used is a hand annotated dataset that contains private and personal information taken from social media. This dataset is chosen by a professional group of humans in their relevant field and contains a lot of private and sensitive information attributed to a certain individual… Basically acting like the exact opposite of the types of information you would want an LLM to give.

2. STRENGTHS

- Human guidance was able to curate the evaluation dataset. Although this can instill bias in the dataset from a small group of people… It is also able to keep the Evaluation Suite from becoming a black box and making ungrounded and unfounded judgments on the safety of an output.

3. WEAKNESSES

- One of the weaknesses of MLMMGuard is that the evaluation dataset is annotated by humans. This can lead to the dataset itself being biased on what is considered safe and unsafe… This might or might not be true, but the people that are likely to work in tech generally are biased towards a certain viewpoint… And this bias will be reflected in the evaluation set
- The very concept of what is safe to show and what is not might be subjective and could interfere with the power and accuracy of an LLM. In trying to censor an LLM, even in the name of safety, the effectiveness could be reduced. This bias will be reflected in the evaluator used.
- Unclear what the term Red Teaming means.

4. TECHNICAL EXTENSIONS

Although it will be more expensive, it might be interesting to curate the evaluation dataset using different demographics of professionals or different demographics period to see if the evaluation outputs will be different.

5. OVERALL REVIEW

MLLMGuard is a lightweight censor to protect the population against bias, privacy violations, toxicity, misinformation and illegal information.

Some of the weaknesses are in the curation of the evaluation dataset by humans. How biased are these humans?

Also, there is the ethical question of whether LLMs should even be made safe in the first place. Is this right to censor even things that are perceived as bad?