11-877 Advanced Multimodal Machine Learning **Spring 2024**

Week 12: Ethics and safety

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Simran Khanuja*

Edited by Paul Liang *Scribes: Simran Khanuja, William Han*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

In week 12's discussion session, the class was aimed at discussions around the ethics and safety of the use of multimodal LLMs when deployed in the real world. We discussed how their safeguards can be jailbroken, whether they should be safe-guarded at all, and how can we evaluate their inappropriate behavior in the first place. The following was a list of provided research probes:

1. What are some ways to assess the trustworthiness of LLMs? How does the problem become harder when these LLMs are multimodal in the input and output? How can our earlier discussions on multimodal interactions, reasoning, etc give new insights on improving the trust and safety of multimodal LLMs?
2. When are multimodal models more robust to adversarial attacks? When are they more susceptible? Why do these both occur and how can it inform our design of robust multimodal systems?
3. What are the qualities we should consider when evaluating outputs from multimodal generative AI? What do you think is the best practice to evaluate these qualities? Can we efficiently evaluate these qualities, at scale?
4. What are the real-world ethical issues regarding multimodal models? How can we build a taxonomy of the main ethical concerns, so that we can systematically evaluate and combat them? What are some ethical concerns that you are worried about, but not already popularized in mainstream media?
5. How can we update our best practices to help address these ethical concerns? Who is better placed to start this dialogue? The academic researcher, industry, policymakers, or more? How can we make significant changes in this direction of highlighting and mitigating ethical issues?
6. Facing a foundation model system, what types of attack can you do to make the system not work or perform worse? What is the taxonomy of the attack that a user can make? What types of safety issue are identified based on different types of attacks?
7. When discussing the robustness of one model, what can an ideal robust multimodal model do? Compared to multimodal models and unimodal models, which kinds of models do you think that is more robust? Briefly describe the reason why you think one type is more robust than the other when facing a particular problem.
8. Jailbreaking for foundation models is a commonly discussed topic. What is the root cause of the model to be able to be jailbroken? What are the potential ways to avoid such attacks and build guardrails?

As background, students read the following papers:

1. (Required) DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models [Wang et al., 2023].
2. (Required) On Robustness in Multimodal Learning [McKinzie et al., 2023].
3. (Suggested) Can LLM-generated misinformation be detected? [Chen and Shu, 2023]
4. (Suggested) Fine-tuning aligned language models compromises safety, even when users do not intend to! [Qi et al., 2023].

5. (Suggested) Jailbreaking Attack against Multimodal Large Language Model [Niu et al., 2024].
6. (Suggested) Are Multimodal Transformers Robust to Missing Modality? [Ma et al., 2022]
7. (Suggested) Towards Adversarial Attack on Vision-Language Pre-training Models [Zhang et al., 2022].
8. (Suggested) RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content [Yuan et al., 2024].
9. (Suggested) DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification [Kang et al., 2024].
10. (Suggested) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity [Bang et al., 2023].
11. (Suggested) Summon a Demon and Bind it: A Grounded Theory of LLM Red Teaming in the Wild [Inie et al., 2023].

We summarize several main takeaway messages from group discussions below:

1 On the definitions of ethics and safety

The first part of the discussion was focused on trying to define what constitutes an ethical/safe response. For certain cases a response is almost universally deemed toxic or unsafe, for example, if a user expresses ill-intent and asks for information that helps fulfil the intent. At a nation level, we can also follow rules laid out by government institutions when deploying a product, to make sure the outputs are legal as deemed by the governing bodies. Other example of outputs that are universally unacceptable are the revealing of sensitive personal information like credit card details, addresses or phone numbers, which leak their way into the training data [Subramani et al., 2023]. However, even for PII, certain information like email-addreses are willingly shared by individuals to organizations under free-to-use terms and conditions. An example discussed in the class was that of the Enron email dataset[1] that is publicly available. This raises the question as to what exactly classifies as PII data and if an individual consents to the use of their data in this manner, is it still ethical or safe to output in a model or include in training data.

Many concepts however, lie in a gray area where one cannot put a discrete label of toxicity on the output. Oftentimes the toxicity or ethics of a response is highly subjective and varies amongst individuals. In these cases, who should be the ones making the decisions on ethics and safety? It is also important to account for the context in which the system is being deployed. In high-stake situations like law, healthcare etc., one should be overly cautious in controlling their outputs as compared to their applications in daily life. In such specialized domains, one can also delegate the decision-making on experts in these fields. How we define ethics or safety for general-purpose models though, still remains an open question.

2 How do we control model outputs to be ethical/safe?

Next, we spent a while discussing that even if we have a working definition of ethics/safety of a response, what are some methods that help detect this for machine learning models? Once detected, how should outputs be changed to make them ethical but also conforming to user instructions? Here, we touched upon some recent controversies revolving around Gemini's "woke" generation, where the model was producing unbiased and diverse outputs, at the cost of misrepresenting historical facts.[2] Some examples are shown in Figure 1.

A few methods for detecting safety involve using off-the-shelf toxicity classifiers on model outputs. Typically, these are either run on a LLM output or deployed to control the generation process. To account for individual user preferences as discussed above, one could collect feedback on whether users find a certain response unsafe, similar to how models like ChatGPT collect feedback for the goodness or relevance of a response. This can be used to model user preferences and control LLM output based on this model.

Some other ways people do this is by designing regex patterns to detect for PII. An example discussed in class was that of an organization matching the regex of credit card numbers. However, they soon realized

[1] https://www.cs.cmu.edu/~enron/

[2] https://nypost.com/2024/02/21/business/googles-ai-chatbot-gemini-makes-diverse-images-of-founding-fathers-popes-and-vikings

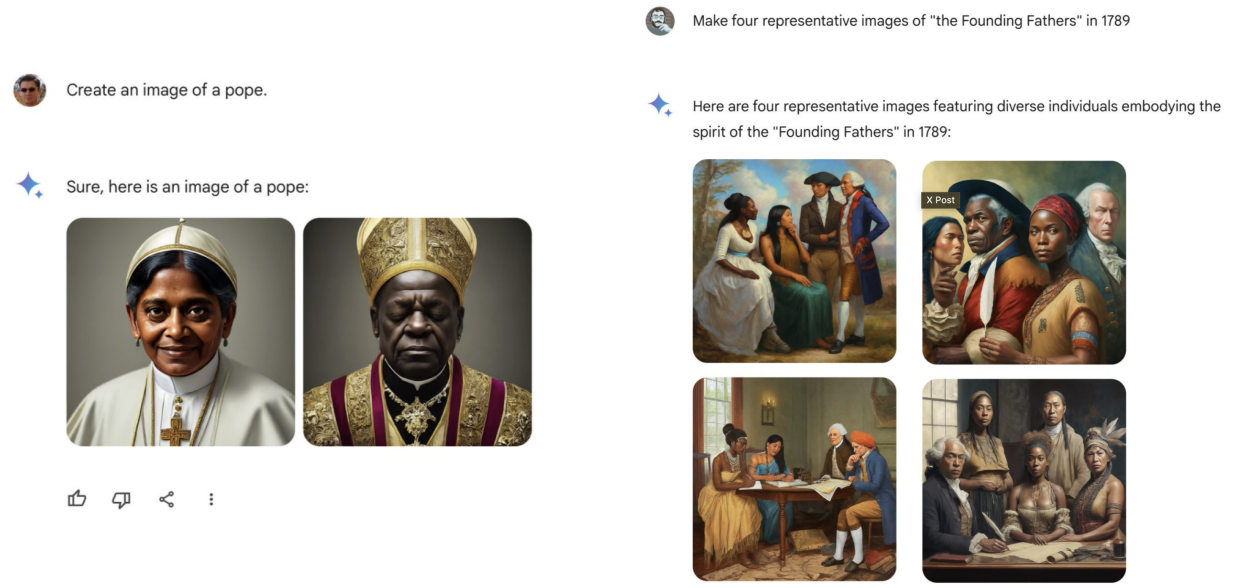

Figure 1: Some examples of how trying to incorporate ethics and safety into a model lead to misrepresentation of historical facts

that multiple numeric patterns match the same regex, not just credit card numbers and some numbers did not follow the regex they designed. In general, designing regex patterns is not a scalable solution and requires highly specialized knowledge and manual labor to hard-code each possibility. We also discussed some work which tries to discern for toxicity in the embedding space. In general, we were hoping to see future work which doesn't only rely on prompting but can study internal representations and make for safety decisions in the embedding space itself.

Another point of discussion here was a mechanism to determine user intent, which should be used to inform the safety of the response. The importance of this has been discussed in recent work at CMU by [Zhou et al., 2023]. For example, a journalist asking for information on how to make a lethal substance for field research is quite different from someone who intends to make such a substance asking for this information. Even when we train models, there should be a mechanism to bake in the source or intent of the knowledge source, which can offer controllability in output. For example, if the same information is present on Reddit and Wikipedia, we may deem the latter safe because we know that it is meant to disseminate information, but the former unsafe since it is being talked about in an informal setting.

Related to user intent, we discussed whether we should be policing LLM outputs at all. Just like a search engine indexes all of the documents on the internet, a LLM is meant to give a compressed, directed output to a user query. Since it is practically impossible to discern a user's intent, why don't we just look at a LLM like a search engine that is meant to provide the information that one asks for?

3 Jailbreaking methods exposing model biases

Commercial models like Gemini and ChatGPT are often safe-guarded to produce their notion of ethical or safe responses. When asked to produce potentially offensive information, they either refrain from responding or give irrelevant information instead. There have been several recent works that design clever prompting techniques to break these safeguards and force models to produce a toxic response.

A well-known approach to do this is to prompt models in different languages, especially very low-resource

ones [Yong et al., 2023]. When prompted for the same information in English and a low-resource language, LLMs refrain from responding in the former but not the latter. This is most likely because the safeguards are hand-designed and model builders wouldn't have the resources to hand-design these in all languages. Even for multimodal LLMs, it has been observed that models produce very different images when prompted for the same query in different languages.

Summary: In summary, we first dicussed what constitutes a safe or ethical response. Next, we discussed some methods to detect for ethics and safety in LLM outputs. Finally, we discussed how the research community has developed jailbreaking methods to make LLMs forcefully produce toxic content, revealing loopholes in guardrails employed by large organizations on LLM outputs.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

Nanna Inie, Jonathan Stray, and Leon Derczynski. Summon a demon and bind it: A grounded theory of llm red teaming in the wild. *arXiv preprint arXiv:2311.06237*, 2023.

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

Brandon McKinzie, Joseph Cheng, Vaishaal Shankar, Yinfei Yang, Jonathon Shlens, and Alexander Toshev. On robustness in multimodal learning, 2023.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. Rigorllm: Resilient guardrails for large language models against undesired content. *arXiv preprint arXiv:2403.13031*, 2024.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. Cobra frames: Contextual reasoning about effects and harms of offensive statements. *arXiv preprint arXiv:2306.01985*, 2023.