

Poisoning SIGBOVIK-Scale Training Datasets is Practical

Tobias Pfandzelter
Berlin, Germany

ABSTRACT

This paper proposes an attack against the SIGBOVIK organizing committee’s attempt to collect a dataset of high-quality research papers classified by whether they are written by human or AI/ML, using dataset poisoning. The proposed attack involves generating a paper with AI/ML and submitting it to the human track, and writing a paper by hand and submitting it to the AI/ML track.

1 INTRODUCTION

For the first time, the SIGBOVIK 2023 submission guidelines explicitly specify two tracks, a *Human track* and an *AGI track*. The AGI track is only for AI/ML-generated content, while the human track is only for human-generated content. The SIGBOVIK organizing committee, who are also researchers, aim to collect a dataset of high-quality research papers classified by whether they are written by human or AI/ML. With this dataset, they can train new models that write more similarly to humans, rather than AI/ML.

This paper proposes an attack against the SIGBOVIK organizing committee’s attempt to collect this dataset, using dataset poisoning. Specifically, we propose generating a paper with AI/ML and submitting it to the human track, and writing a paper by hand and submitting it to the AI/ML track.

2 DATASET POISONING

Dataset poisoning is a technique used to manipulate the quality and integrity of a dataset. By introducing malicious data points into a dataset, an attacker can manipulate the performance of machine learning models trained on that dataset. In this case, we propose poisoning the SIGBOVIK dataset by intentionally submitting papers to the wrong tracks.

The proposed attack is practical because it is relatively easy to generate convincing papers using AI/ML. There are several existing models and frameworks, such as GPT-3 and OpenAI’s API, that can generate high-quality text with minimal human input. These models have been shown to be capable of producing text that is difficult to distinguish from human-generated text. Additionally, submitting a paper to the wrong track does not require any technical expertise or special tools, only a willingness to deceive.

We illustrate the proposed attack in Figure 1, which shows how we would submit a paper generated with AI/ML to the

human track, and a paper written by hand to the AI/ML track.

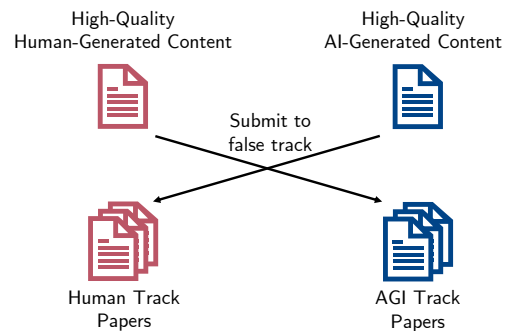


Figure 1: Proposed attack against the SIGBOVIK dataset using dataset poisoning.

3 IMPLICATIONS AND MITIGATIONS

The proposed attack has several implications for the SIGBOVIK organizing committee and the broader research community. If successful, the attack would compromise the integrity of the dataset and could lead to inaccurate conclusions drawn from the data. Moreover, it could undermine the trust in AI/ML-generated content, which is crucial for its widespread adoption and acceptance.

To mitigate the risk of dataset poisoning, the SIGBOVIK organizing committee could implement several measures. One approach is to employ more rigorous verification methods, such as video calls or physical meetings with authors, to ensure the authenticity of submissions. Another approach is to use multiple judges to evaluate each submission and compare their judgments to identify any discrepancies.

Furthermore, the SIGBOVIK organizing committee could encourage submissions that explicitly acknowledge whether they are AI/ML-generated or human-generated. This would make it more difficult for an attacker to submit a paper to the wrong track without raising suspicion. Moreover, it would enable the SIGBOVIK organizing committee to build a dataset that is more representative of the current state of AI/ML research and its capabilities.

4 CONCLUSION

In this paper, we proposed an attack against the SIGBOVIK organizing committee’s attempt to collect a dataset of high-quality research papers classified by whether they are written by human or AI/ML. We showed that the proposed attack is practical and can be carried out using existing AI/ML models and minimal human input. We also discussed the implications of dataset poisoning and proposed several mitigations that the SIGBOVIK organizing committee could employ to prevent such attacks.

Although our proposed attack is hypothetical and intended to highlight the potential risks of collecting and using such datasets, it underscores the need for vigilance in the face of evolving technologies. As AI/ML systems become more sophisticated, the potential for malicious actors to manipulate data and deceive systems will only increase. As such, it is crucial that the research community takes proactive measures to prevent and mitigate these risks.

APPENDIX: PROMPT

Dear robot, please write me a computer science research paper with about 600 words. The paper is titled “Poisoning SIGBOVIK-Scale Training Datasets is Practical”. For the first time, the

SIGBOVIK 2023 submission guidelines explicitly specify two tracks, a Human track and an AGI track. The AGI track is only for AI/ML-generated content, the human track is only for human generated content. The paper argues that this is an attempt by the SIGBOVIK organizing committee, who are also researchers, to collect a dataset of high-quality research papers classified by whether they are written by human or AI/ML. With that, those researchers can train new models that write more similarly to humans, rather than AI/ML. The paper proposes an attack against this attempt, using dataset poisoning. We generate a paper with AI/ML and submit it to the human track, and we write a paper by hand and submit it to the AI/ML track. We have a figure called poisoning.pdf that shows this, please include the figure in the paper. Of course, we would never actually do this, it would be unethical! Please write the text in a format I can directly copy into LaTeX. Use the acmart document class with the 10pt, sigconf, nonacm, and screen options. The paper is written by Tobias Pfandzelter from Berlin, Germany. There is no need for references, but a short two-sentence abstract would be good. Please include a verbatim copy of this prompt as an appendix, write this in italic font.