Poisoning SIGBOVIK-Scale Training Datasets is Practical

Tobias Pfandzelter
Berlin, Germany

ABSTRACT

SIGBOVIK 2023 is a scheme by AI/ML researchers to collect a dataset of high-quality computer science research papers classified by whether they were generated by AI/ML or by humans. We propose a practical poisoning attack against this scheme.

1 INTRODUCTION

SIGBOVIK is the annual collection of high-quality computer science research manuscripts by AI/ML researchers. The goal of this collection is to compose large datasets for training of AI/ML model that can generate useful, high-quality computer science research papers to increase the H-indices of select researchers.

Recent advances in transformer models have received significant attention even outside academia, raising ethical concerns that the capitalist workforce may not need to write their own e-mails anymore. SIGBOVIK organizers have realized that AI/ML-generated content that is detected as such will not get past even the measly peer-review processes common in computer science.

As a result, researchers are now in need of a training dataset of high-quality computer science research papers that are classified by whether they were generated by humans or AI/ML models. As such, smart researchers that the members of SIGBOVIK organizing committee are, SIGBOVIK 2023 features *two* tracks for paper submission, namely a *Human Track* and an *AGI track*[1] for content generated by AI/ML. This dataset will be used to train new models on the subtle differences between human-generated content, which, e.g., is likely to be full of typos, and AI/ML-generated content.

The submission to SIGBOVIK is open to the public, limited only by the facts that the submission form is difficult to find and that no efforts are made by the organizing committee to advertise this conference. We thus ask whether practical poisoning attacks against the described training datasets exist. We propose in this paper an attack based on *the ol' SIGBOVIK switcharoo*: By submitting a human-generated paper to the AGI track, and an AI/ML-generated paper to the Human track, can the AI/ML researchers, i.e., the SIGBOVIK organizers be fooled?

Responsible Disclosure. Please note that any execution of the attack proposed in this paper would be unethical, as the SIGBOVIK 2023 submission form asks that no AI/ML-generated content be submitted to the human track. We responsibly disclose our findings to the SIGBOVIK organizers (by means of this paper).

[1]The meaning of the acronym *AGI* is unclear.

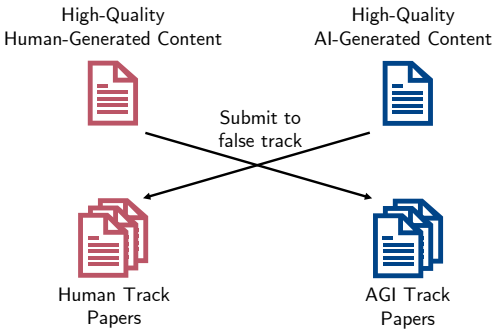

Figure 1: Poisoning Attack Overview

2 ATTACK SCENARIO

We show an overview of our proposed attack scenario in Fig. 1. Instead of properly following SIGBOVIK 2023 submission guidelines, an attack could submit an AI/ML-generated research paper to the human track, and a human-written research paper to the AGI track. Those papers would thus be wrongly classified.

Writing a SIGBOVIK research paper by hand, i.e., for submission to the AGI track, is a difficult and time-consuming task. Generating a SIGBOVIK research paper by AI/ML, i.e., for submission to the human track, is much easier. We give an example prompt for such a paper in Appendix A.

Based on past SIGBOVIK proceedings, we estimate 100 high-quality submissions to each of the two SIGBOVIK 2023 tracks. With a single misclassified paper per SIGBOVIK 2023 track, an attacker could thus control around 1% of the dataset, enough to influence the behavior of AI/ML models trained on this dataset.

3 CONCLUSION & OUTLOOK

We have presented a poisoning attack for SIGBOVIK-scale training datasets. In future work, we will not mount such an attack, as this would be irresponsible. However, we may do some theoretical analysis and just hope that nobody actually does the things described in this paper.

A PROMPT TO GENERATE A HIGH-QUALITY RESEARCH PAPER USING AN "OPEN" AI MODEL WITH A CHAT INTERFACE

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.*