Uncertainty Estimation Methods for Countering Attacks on Machine-Generated Text Detectors

Valeriy Levanov[1], **Anastasia Voznyuk**[1], and **Andrey Grabovoy**[1]

[1]Moscow Institute of Physics and Technology, Moscow

March 6, 2025

Abstract

In this work, we explore the use of uncertainty estimation methods to enhance the quality of machine-generated text detectors against various attacks, such as homoglyphs, paraphrasing and noise injection. These attacks are not only used to bypass detection but also to test the resilience of detectors. We will test the hypothesis that uncertainty estimation can provide a more sustainable approach, eliminating the need for constant retraining. The research will evaluate this hypothesis in two scenarios: when only the text is available and when access to the model's internal states is also provided. The planned experiments aim to validate the results and compare them with current state-of-the-art solutions.

1 Introduction

2 Conclusion

References