The AI Disguise: Applying Advanced Tokenization and POS Techniques for Text Authenticity

Andrew X. Zhang (/profile?id=~Andrew_X._Zhang1)

Published: 17 Jun 2024, Last Modified: 17 Jun 2024 DMLR @ ICML 2024 DMLR, Area Chairs, Reviewers, Authors Revisions (/revisions?id=WhaVsLRO9L) BibTeX

Keywords: Large Language Models, AI-Detection, Maximum Likelihood Estimate

TL;DR: This paper enhances the Distributional Large Language Model (LLM) Quantification Framework by incorporating advanced tokenization and parts-of-speech techniques to more accurately differentiate between AI-generated and human-generated text.

Abstract:
We propose an enhancement to the Framework originally introduced by Liang (2024). By incorporating different phrase lengths and parts of speech, our improved model more accurately differentiates between AI-generated and human-generated text at the corpus level. Testing on Liang's original validation datasets revealed three models that outperform the baseline error rate of 1.97%: a 1-word model focusing on adjectives (1.63%), a 2-word model focusing on adverbs (1.42%), and a 1-word model focusing on verbs (0.47%). These enhancements highlight LLMs' biases towards specific tokens and provide a more reliable method for distinguishing AI-generated content from human-generated content.

Submission Number: 69

Filter by reply type | Filter by author | Search keywords... | Sort: Newest First
Everyone | Program Chairs | Submission69 Authors | Submission69 Area... | Submission69... | Submission69...
4 / 4 replies shown

Add: **Withdrawal**

Paper Decision

Decision Program Chairs 17 Jun 2024, 12:06 (modified: 18 Jun 2024, 00:09) Program Chairs, Authors Revisions (/revisions?id=SFAQjbLeAE)

Decision: Accept

Meta Review of Submission69 by Area Chair

A5Xp

Meta Review Area Chair A5Xp 16 Jun 2024, 20:17 (modified: 18 Jun 2024, 00:09)

Area Chairs, Authors, Program Chairs Revisions (/revisions?id=EA2s4VDclC)

Metareview:
The reviewers agree that this work on identification of generated text is relevant to the topics of the workshop. I agree with this assessment, and recommend acceptance. Nevertheless, I would recommend the authors expand on the technical details of their experiments, as suggested by the reviewers.

Recommend For Journal: No
Recommendation: Accept (Poster)
Recommend For Ai For Good: No
Confidence: 3: The area chair is somewhat confident

This paper proposes an enhancement to the Distributional Large Language Model (LLM) Quantification Framework by incorporating different phrase lengths and parts of speech.

Official Review Reviewer CPXC 13 Jun 2024, 11:52 (modified: 18 Jun 2024, 00:08)

Program Chairs, Area Chairs, Reviewer CPXC, Authors Revisions (/revisions?id=3ZFdEGTl8p)

Recommend For Journal: Yes
Review:
Pros:

1. The topic of detecting AI-generated data is important.
2. Author find that incorporating different phrase lengths and parts of speech could more accurately detecting AI generated-data.

Cons:

1. Novelty is limited, and this may seem like a single experiment to explore the effects of phrase lengths and parts of speech.
2. Descriptions are vague, such as which model the AI-generated text was generated, experimental Settings, and so on.

Recommend For Ai For Good: Yes
Rating: 3: Clear rejection
Confidence: 3: The reviewer is fairly confident that the evaluation is correct

The AI Disguise: Applying Advanced Tokenization and POS Techniques for Text Authenticity

Official Review Reviewer hkoL 12 Jun 2024, 23:10 (modified: 18 Jun 2024, 00:08)

Program Chairs, Area Chairs, Reviewer hkoL, Authors Revisions (/revisions?id=BTRbktqpHT)

Recommend For Journal: Yes
Review:
This paper focuses on developing new techniques to differentiate AI-generated text from human-generated text. It proposes an enhancement to a previous method and conducts extensive experiments. Unlike traditional models, their approach operates on 2 or 3-word tokens and uses POS tagging to create different versions of the source dataset. By

tagging parts of speech and processing them independently, they demonstrate a reduction in test error percentage. Their simple yet elegant method is effective.

Overall, the paper is easy to understand, relevant to the conference goals, and presents work of practical significance.
Recommend For Ai For Good: Yes
Rating: 7: Good paper, accept
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct