Bypassing Human/Machine Classifiers by Prompting LLMs for Naturally Imperfect Text

CLEF 2025

Mohanna Hoveyda[1]

[1] *Radboud University, Nijmegen, The Netherlands*

Abstract

Keywords

Voight Kampff, Llama 3, Prompting

1. Introduction

The objective of this experiment is to evaluate the extent to which a human/machine text classifier can be bypassed using off-the-shelf large language models (LLMs). To this end, we do not perform any fine-tuning or task-specific training of the LLM. Instead, we prompt the model to generate text that mimics natural human imperfections. This approach is grounded in the observation that LLMs are typically optimized to produce well-structured and polished outputs, which may differ from the way humans express themselves in everyday communication. By encouraging the model to introduce subtle flaws, we aim to exploit the gap between machine-generated and human-written text, leveraging the notion that in essence, to err is human.

2. Methodology

2.1. Model

The language model used is Llama-3.3-70B.

2.2. Prompt

As shown below, the prompt template is filled with the content, genre and style of each of the samples in the task. Afterwards, the filled prompt template is concatenated at the end of the system's message. The final text is fed to the LLM.

```
    system_msg = (
        "You are a writer that writes very natural text close to how amateur people write. The goal is to generate text that feels human-like, with minor typos, and colloquial phrases. "
        "You will be given a topic with genre and style, and a list of bullet points to cover. "
        "Your task is to write a coherent text that covers all points, using varied sentence lengths and natural language. "
        "Output only the final written text body; do not include any extra commentary, notes, or metadata."
    )
```

```
PROMPT_TEMPLATE = """
Genre: {genre}
Tone: {style}

Write a ~500-word text covering these bullet points.
— Use varied sentence length, minor typos, and natural colloquial phrases
— Output only the final text

Bullet points:
{content}
"""
```

3. Results

Based on the results of the Voight Kampff task, it appears that this simple prompting approach might be quite effective in bypassing current machine/human text classifiers. However, it is important to note that our experiments were conducted using a relatively large and capable model (a 70B-parameter LLM) which is particularly well-suited for instruction following and producing the desired behavior. This may limit the generalizability of the findings to smaller or less capable models.