Evidence of Policy Violation by Reviewer wNfJ

Paper: #6307

1. Summary

This report presents evidence that the review provided by Reviewer wNfJ (Score: 2) was **fully generated using an LLM**, violating ICLR's review policy. Our analysis includes a control group: the same detection tools correctly identified all other reviewers (WqcR, rGrB, CKa4, F3q3) as "Fully Human," eliminating the possibility that the paper's technical jargon triggered a false positive. The summary of detection results is shown in Table 1.

Table 1: Summary of LLM detection result.

<table>
<tr><th>Reviewer</th><th>Rating</th><th>Pangram Labs</th><th>GPTZero</th></tr>
<tr><td>wNfJ</td><td>2</td><td>Fully AI-generated</td><td>100% AI generated</td></tr>
<tr><td>WqcR</td><td>4</td><td rowspan="4">Fully human-written</td><td rowspan="4">100% Human</td></tr>
<tr><td>rGrB</td><td>6</td></tr>
<tr><td>CKa4</td><td>6</td></tr>
<tr><td>F3q3</td><td>6</td></tr>
</table>

2. Evidence

2a. Pangram Labs

Pangram Labs (https://www.pangram.com/) is a leading enterprise-grade detection solution trusted by global organizations and academic institutions. Notably, Pangram provides a specialized detection module for ICLR reviews (https://iclr.pangram.com), specifically calibrated to handle the technical density and formal tone of ICLR submissions.

As shown in Figure 1, the tool successfully filtered our submission's reviews. Only **Reviewer wNfJ** (Score: 2) was flagged as "**Fully AI-generated**" (Red). The table reveals a perfect correlation between the AI flag and the low quality: the AI-generated review corresponds to the outlier score of 2.00, whereas the verified human reviews maintain a high average rating of 5.50.

Summary Statistics

EditLens Prediction	Count	Avg Rating	Avg Confidence	Avg Length (chars)
Fully AI-generated	1 (20%)	2.00	4.00	4129
Heavily AI-edited	0 (0%)	N/A	N/A	N/A
Moderately AI-edited	0 (0%)	N/A	N/A	N/A
Lightly AI-edited	0 (0%)	N/A	N/A	N/A
Fully human-written	4 (80%)	5.50	4.00	2367
Total	**5 (100%)**	**4.80**	**4.00**	**2719**

Figure 1: Screen shot from Pangram Labs detection result for paper # 6307.

Official report link: https://iclr.pangram.com/reviews?submission_number=6307

2b. GPTZero

GPTZero (https://gptzero.me/) is widely regarded as the "gold standard" for AI detection, used by educators and institutions worldwide. It utilizes perplexity and burstiness metrics to distinguish between human and machine writing patterns.

Figure 2 identifies **Reviewer wNfJ as "100% AI-generated."** In contrast, Figures 3 and 4 confirm that reviews from other reviewers were correctly identified as "100% Human."

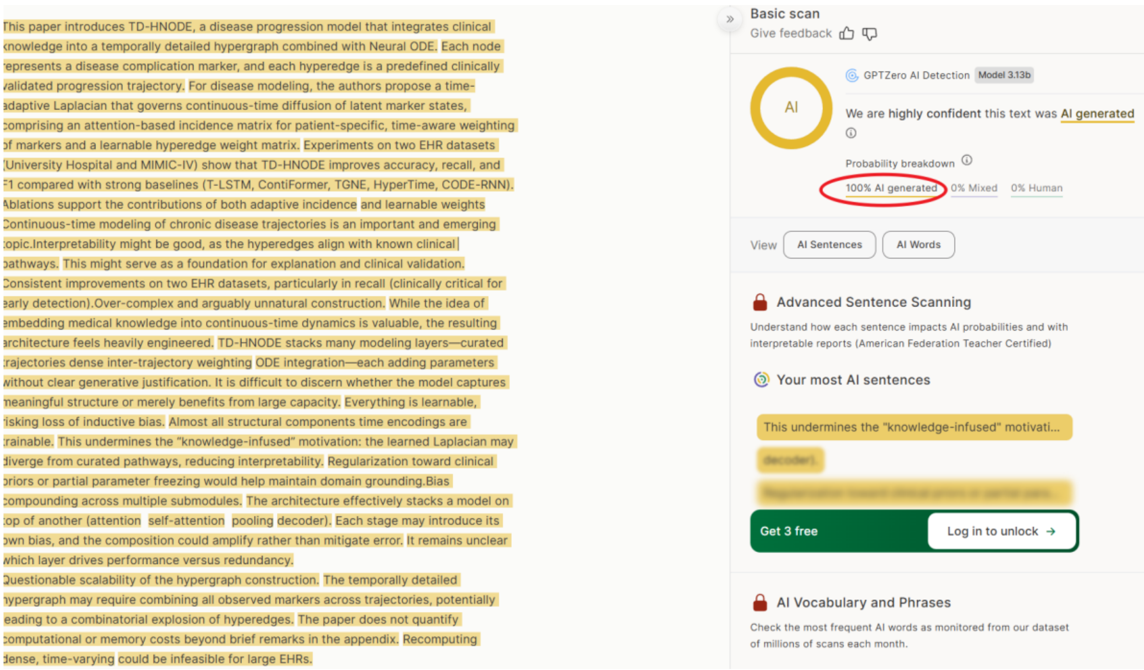

Figure 2: GPTZero detection result (100% AI generated) for **Reviewer wNfJ.**

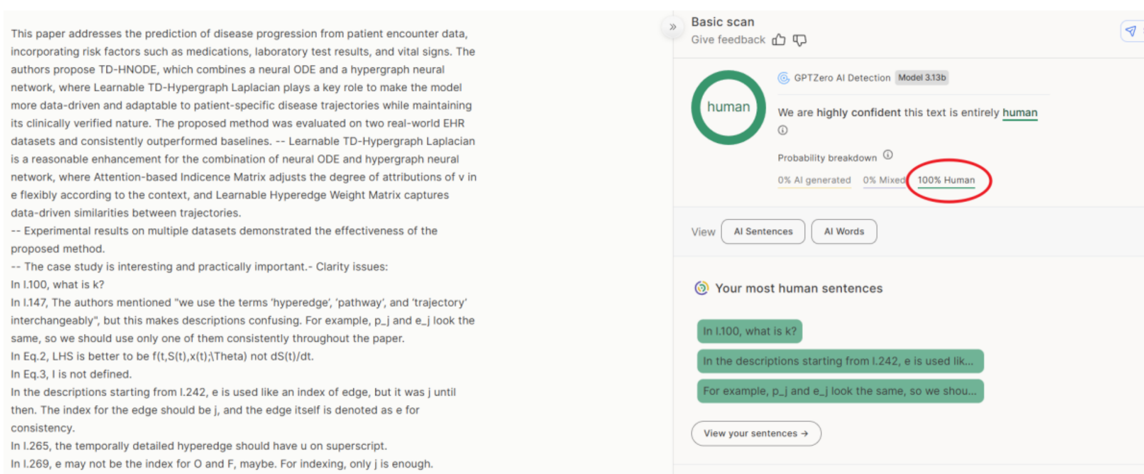

Figure 3: Control check: GPTZero detection result (100% Human) for Reviewer rGrB.

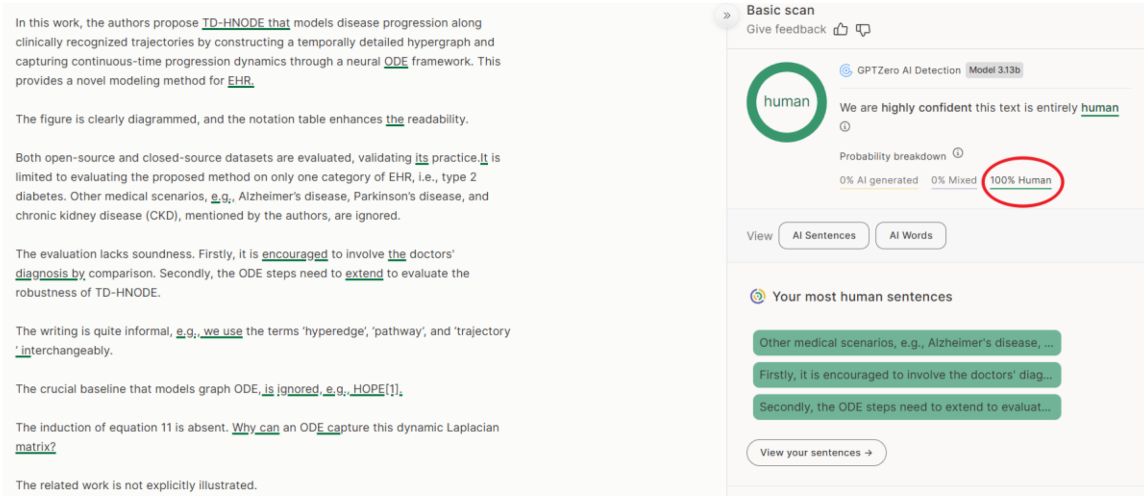

Figure 4: Control check: GPTZero detection result (100% Human) for Reviewer WqcR.

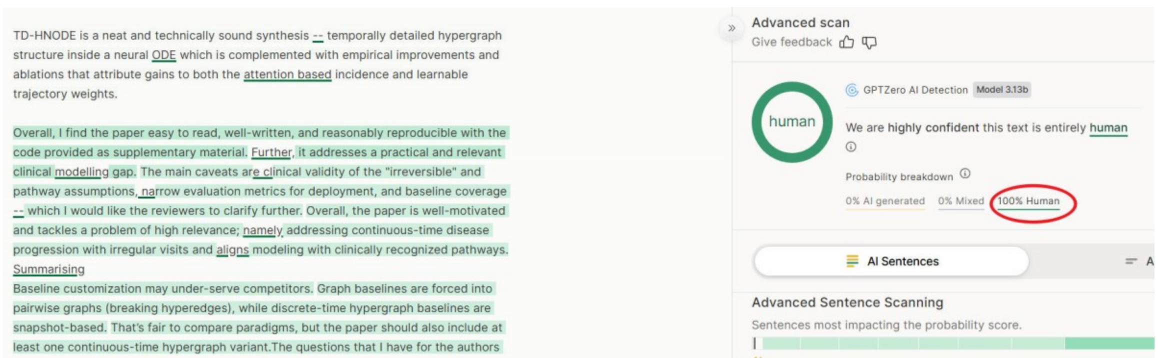

Figure 5: Control check: GPTZero detection result (100% Human) for Reviewer CKa4.

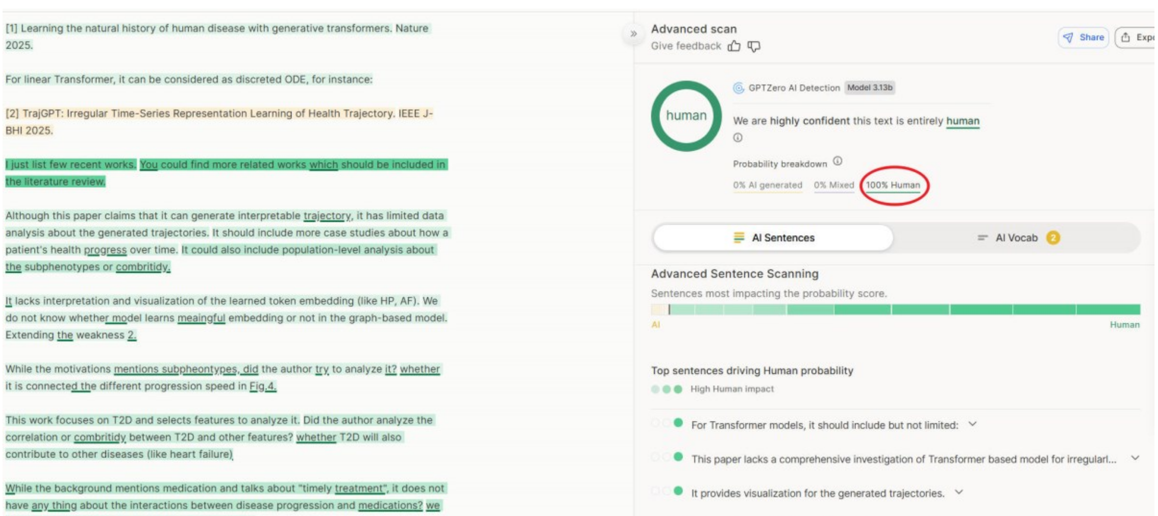

Figure 6: Control check: GPTZero detection result (100% Human) for Reviewer F3q3.