Detection & Deception: The Epistemic Lifecycle of WP: Signs of AI Writing

AOIR 2026 PC0003. In Defense of The Commons | 4 Oct. 2026
M. Vetter, P. Thomas, R. Barrett, G. Alevizou

1. BACKGROUND

As AI reshapes the web via continued extraction and enclosure of data from public repositories, there is a growing imperative to articulate how the value of such data evolves across various stages of extraction, transformation, and operationalization. WP: Signs of AI Writing, a catalog of AI tells compiled to help detect and block LLM edits from appearing in Wikipedia mainspace articles, represents a productive case study in that it is used as a descriptive guide for human editors, but also feeds into automated edit filters in Wikipedia, and LLM "humanizers" outside of the community.

This in-progress study is especially interested in examining how crowd-sourced knowledge created to defend Wikipedia's integrity is repurposed into prescriptive and executable code in downstream platforms.

2. WORKING RESEARCH QUESTION

How is Wikipedia's volunteer-curated *Signs of AI Writing* catalog guide extracted, transformed, and operationalized by downstream LLM humanizer tools, and what epistemic shifts occur when descriptive governance heuristics become prescriptive code?

WP:SIGNS OF AI WRITING

Created in December 2023 by volunteer editors in *WikiProject AI Cleanup*, the advice page Wikipedia:Signs of AI writing documents observational tells of LLM-generated prose. The page generated over 2.13 million pageviews (Dec 2023 – Sep 2026), breaking containment in July 2025 as major media (*The Washington Post, The Verge, Wired*) highlighted Wikipedia's role as the Internet's premier AI detector.

Category	Tells #	Documented Heuristic Examples & Model Artifacts
Markup & Bugs	12	Raw Markdown leaks, corrupt wikitext, citation tags (ChatGPT oaicite, Gemini cite:1, Grok cards, DeepSeek daggers).
Style & Typography	12	Overuse of boldface, em dashes, emojis, curly quotes, inline-header vertical lists, skipping heading levels.
Content & Tropes	9	Undue emphasis on legacy/significance, canned notability claims, formulaic 'challenges & prospects' conclusions.
Language & Syntax	8	High AI vocabulary density ('delve', 'tapestry'), negative parallelisms (*'Not X, but Y'*, *'Y rather than X'*), rule of three.
Citations	8	Hallucinated or malformed sourcing (invalid DOIs/ISBNs, broken DOIs leading to unrelated papers, marketing tracking tags).
User Communication	3	Conversational leakage directly addressing prompter, knowledge-cutoff disclaimers, leftover phrasal placeholders.

Epistemic Inversion: From Descriptive Heuristic to Prescriptive Execution

When extracted beyond Wikipedia (e.g., blader/humanizer on GitHub with 49k stars, 4k forks, 7k installs), the original purpose of the catalog undergoes a fundamental inversion. Descriptive guidelines meant to protect encyclopedia quality are converted into prescriptive instructions for chatbots to obscure synthetic origin, creating an adversarial value asymmetry between volunteer editors and downstream commercial toolmakers.

3. DATASETS

Category	Dataset / Resource	Description & Scope
Wiki Project Pages	Wikipedia:Signs_of_AI_writing (WP:AISIGNS)	Community-curated catalog of 12 indicator categories (2.13M+ pageviews, Dec 2023–Sep 2026).
Wiki Project Pages	Wikipedia:WikiProject_AI_Cleanup & Template:AI-generated	Central cleanup task force page and article-level maintenance disclosure template.
Talk Pages & Admin Space	Wikipedia_talk:Signs_of_AI_writing	Talk page discussion threads (Archives 1–5, 493 watchers, ~2,000 monthly views).
Talk Pages & Admin Space	User_talk: Notices & Wikipedia:AI_noticeboard	User-facing talk page warning templates (Levels 1–4) and administrative noticeboard.
Quarry Database Logs	Quarry 109267 (llm-tag-adding-tracker)	843 article edit records (Feb 2024–Sep 2026) tracking {{AI-generated}} tag actions.
Quarry Database Logs	Quarry 109270 (llm-talkpage-notice-tracker)	4,542 notice events across 4,070 unique users (Nov 2024–Sep 2026).
Edit Filters & Logs	Abuse Filter 1,325 (Possible AI-generated text)	Automated system rule filter capturing 13,211+ hits and full history logs.
Edit Filters & Logs	Abuse Filters 893, 1030, 1346, 1369	Filters for AI sources, markdown leaks, and citation artifacts.
GitHub API & Repositories	egc365/signs-of-ai-writing	Structured markdown/JSON repository extracting the Wikipedia field guide.
GitHub API & Repositories	blader/humanizer & Downstream Forks	Executable Claude AI skill (49k stars, 4k forks, 7k+ installs) and related forks.

Key Questions

1. **Protecting Open Knowledge:** How can digital commons projects like Wikipedia protect self-regulatory quality guidelines when they are transformed into optimization targets for adversarial AI tools?
2. **Value Asymmetries:** What governance, licensing, or technical interventions can prevent the asymmetric extraction of volunteer labor by commercial AI humanizer platforms without closing the open web?
3. **Methodology:** What other methods and/or datasets might be useful in eploring the research question and mapping the evolving value of the Signs of AI Writing across Wikipedia and Github?

BIO

Professor of Writing at Indiana University of Pennsylvania, Vetter's research asks questions related to writing, digital culture, and technology with an emphasis on open knowledge and critical literacy

CONTACT

email : mvetter@iup.edu
mattvetter.net
ORCID:0000-0002-4590-5677