Labeling the Spectrum of AI Involvement: New Tag Proposals for Wikipedia and Commons

Ashik Ahamed
Clarkson University
Potsdam, NY, U.S.A
ahameda@clarkson.edu

Max Wang
Clarkson University
Potsdam, NY, U.S.A
mawang@clarkson.edu

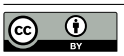
Jeanna Matthews
Clarkson University
Potsdam, NY, U.S.A
jnm@clarkson.edu

Abstract

Artificial intelligence tools, particularly large language models, are becoming deeply embedded in everyday writing and editing practices. As a result, Wikipedia now receives AI generated and AI assisted content in many forms, yet the platform lacks clear mechanisms for identifying when and how these tools were used. Existing MediaWiki tags offer only limited metadata about the interface or action that produced an edit and do not identify the type or extent of AI involvement. At the same time, fully prohibiting AI generated text is impractical because many editors already rely on tools such as Grammarly, ChatGPT, Claude, and Google Gemini for grammar correction, summarization, translation, and content drafting. Instead, a principled and well-designed approach for labeling AI involvement is needed that 1) reflects the full range of assistance and 2) supports transparency for both editors and researchers. Building on prior work analyzing inconsistencies in Wikipedia's Special Tags system, this paper argues that current tagging practices do not match the realities of modern editing workflows. Tool usage is frequently underreported, tag adoption varies widely across languages and topic areas, and editors often have incentives to hide tool involvement to avoid being held responsible for errors introduced by automated systems. Recent work, such as the development of LLM-based image captioning tools for Wikimedia Commons, illustrates that AI participation is already widespread and expanding. Instead of attempting to restrict AI use entirely, this work proposes a more practical strategy. This paper outlines a set of new tags organized into four major categories: Content Creation Tags for textual editing, Assistance and Verification Tags for evaluation and support functions, Metadata Suggestion Tags for organizational elements, and Media-Specific Tags for images, audio, and video. These tags document how, where, and to what extent AI systems contributed to Wikipedia content, providing a foundation for greater transparency, accountability, and informed research on human-AI collaboration.

CCS Concepts

• **Human-centered computing → Collaborative content creation**.



WebSci Companion '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2492-3/2026/05
https://doi.org/10.1145/3795513.3810451

Keywords

AI-assisted editing, Human–AI collaboration, Content transparency, Metadata.

ACM Reference Format:
Ashik Ahamed, Max Wang, and Jeanna Matthews. 2026. **Labeling the Spectrum of AI Involvement: New Tag Proposals for Wikipedia and Commons**. In *18th ACM Web Science Conference Companion (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3795513.3810451

1 Introduction

Wikipedia has become one of the most widely used knowledge resources in the world, and its continued reliability depends on a clear understanding of how content is produced and maintained. The platform has long relied on human volunteers who create, update, and curate articles using a mixture of manual editing, semi-automated tools, and community review processes. In recent years, however, the landscape of content production has changed significantly. Large language models and related artificial intelligence systems are now deeply embedded in everyday writing practices, and editors increasingly use these tools for drafting, summarizing, translating, fact-checking, and media description [26]. As a result, Wikipedia is already receiving substantial amounts of AI-generated and AI-assisted content, even though the platform has limited mechanisms for recording when and how these tools were involved [21]. The current MediaWiki tagging system provides only a partial view of tool activity. Tags such as "visualeditor," "mobile edit," or "reverted" indicate how an edit was made at the interface level, but they offer only limited information about the type or degree of AI involvement [1]. Prior work examining the Special:Tags system shows that tool usage is inconsistently recorded and varies widely across languages and topic areas [1]. In many cases, editors have incentives to avoid tagging tool involvement to prevent being held responsible for errors introduced by automated systems, a dynamic described in prior research as a moral crumple zone [6]. The result is a revision history in which human and automated contributions are increasingly intertwined but not clearly distinguishable [18].

At the same time, fully restricting AI-generated text is unlikely to be effective. Editors use a wide range of tools such as Grammarly, ChatGPT, Claude, and Google Gemini for routine tasks like grammar correction and summarization, and these uses often leave no trace in the platform's metadata [26]. Recent developments, including LLM-based image captioning tools for Wikimedia Commons, further demonstrate that AI systems are already being integrated into Wikimedia workflows and will continue to expand [22, 25].

This changing editing environment creates an urgent need for more complete and transparent metadata. Readers, moderators, and

researchers cannot reliably determine whether a given paragraph was drafted by a person, written by an LLM, or simply polished for grammar. Without clear labeling, it becomes difficult to evaluate the provenance, reliability, and potential risks of AI assisted content, especially in areas where accuracy and neutrality are essential.

This work proposes an expanded tagging framework that captures the spectrum of AI involvement in both textual and multimedia editing. These tags are designed to make visible the lightest forms of assistance, such as grammar correction, as well as heavier forms, such as full draft generation or unreviewed AI output. The goal is to support transparency, enable more accurate research on human-AI collaboration, and give communities better tools to evaluate how emerging technologies are shaping Wikipedia's content production practices. The community is also divided on what to do, with some advocating for a prohibition of AI tools [24] and others embracing use of tools for light editing or even more [21]. In practice, a prohibition on AI-generated text would be impossible to enforce. Many people suggest banning AI-generated content, but this is hard because AI use is already common and difficult to detect [9]. This work aims to enable contributors to Wikipedia to label their contributions accurately. Rather than discouraging or shaming contributors for using AI tools, labeling AI involvement is a more practical path that encourages honest disclosure. Labels would allow those consuming Wikipedia, including those using Wikipedia as training data, to make more informed decisions about whether to include AI-generated content. For some use cases, AI-generated content could be helpful to include, while for other use cases, only content generated by humans may be preferred. Labels would allow consumers to make the choice. Clear labels also help newcomers, smaller language communities, and accessibility projects that depend on AI tools. Transparent tagging makes it easier for researchers and moderators to understand how content was created [4]. The key is to make it clear to contributors that their work is more valuable when labeled properly. We do not argue that it will be easy to convince Wikipedia contributors to tag their contributions properly - it won't be and there will always be untagged contributions as happens today with the current tags. Still, we argue that offering a set of accurate tags and encouraging their use is essential.

2 Related Works

2.1 AI-Assisted Content Creation and Detection

The increasing integration of artificial intelligence in content creation has prompted significant research attention across multiple platforms. Large language models have become deeply embedded in everyday writing practices, fundamentally changing how users produce and edit text across digital platforms [21]. Recent studies have documented widespread adoption of AI tools such as ChatGPT, Claude, and Google Gemini for tasks ranging from grammar correction to full draft generation [26], creating urgent questions about detection, attribution, and transparency.

Recent work has explored the transparency dilemma surrounding AI disclosure. Studies show that AI disclosure can paradoxically erode trust and reduce perceptions of legitimacy, even when the AI-generated content is accurate [14]. This creates a challenging dynamic for voluntary disclosure systems, where editors may have strategic incentives to conceal AI involvement to maintain credibility.

2.2 Wikipedia Governance and Tagging Systems

Wikipedia's existing infrastructure for documenting edits provides an important foundation for understanding AI involvement. The platform's Special:Tags system was designed to capture metadata about editing workflows, including indicators such as "visualeditor," "mobile edit," and "contenttranslation". However, these tags were developed for an earlier technological environment and do not adequately capture the spectrum of AI assistance now common in editing workflows [1]

Research on Wikipedia governance has emphasized the importance of community deliberation and consensus-building in policy adoption [7]. Decentralization in Wikipedia governance creates complex dynamics for implementing platform-wide changes, suggesting that any new tagging framework must navigate diverse community norms across language editions [8]. Studies examining Wikipedia's quality control mechanisms have established strong relationships between editing behavior and article quality [2, 13], suggesting that transparent metadata about AI involvement could enable similar analysis of how AI-assisted editing affects content quality, neutrality, and reliability.

2.3 Moral Responsibility and Human-AI Collaboration

The concept of the "moral crumple zone" introduced by Elish (2019) provides crucial theoretical grounding for understanding editor behavior around AI disclosure [6]. Elish documented how humans in human-robot interaction scenarios often absorb responsibility for automated errors, creating incentives to distance themselves from AI systems when problems occur. This dynamic directly relates to Wikipedia editing, where editors who acknowledge AI assistance may fear being held accountable for AI-generated errors or policy violations.

The broader literature on transparency and accountability in AI systems emphasizes that effective disclosure frameworks must balance multiple goals: maintaining user trust, enabling informed decision-making, and supporting accountability without creating punitive atmospheres [12, 19]. Research suggests that successful systems frame disclosure as documentation rather than admission of wrongdoing, create graduated categories rather than binary choices, and emphasize the value of transparent metadata for research and quality control [19].

2.4 AI Tools for Wikimedia Commons

Recent developments in Wikimedia Commons have begun addressing AI integration in multimedia contexts. The Commons Structured Data project has explored computer-aided tagging and automated caption generation for images [22]. Redi et al. (2021) released substantial datasets for Wikipedia image-caption matching challenges, demonstrating growing infrastructure for AI-assisted media description [25].

Schindler's (2025) Commons Image Description tool represents practical implementation of AI assistance for metadata generation [15]. By extracting EXIF data and populating existing MediaWiki

templates, this tool illustrates how AI can support routine organizational tasks while requiring transparent attribution of AI involvement. Research on AI-generated image descriptions has emphasized the importance of human oversight, showing that automated descriptions often require substantial editing and review [5, 16]. These findings underscore that AI assistance in media description exists on a spectrum requiring different levels of transparency.

2.5 Content Quality and Neutrality in Wikipedia

Research within the WikiProject AI Cleanup initiative has documented numerous cases where AI-generated content introduces bias, promotional language, or unverified claims that violate Wikipedia's Neutral Point of View (NPOV) policy [28]. This work emphasizes that AI tools can both help and hinder neutrality. AI can assist in identifying bias, but AI-generated rewrites may introduce new neutrality problems if not carefully reviewed. Studies relating Wikipedia article quality to edit behavior and link structure demonstrate that metadata about editing processes enables important quality analysis [13]. Similarly, research on automatically labeling low-quality content by leveraging editing behavior patterns shows that transparent edit metadata supports both automated and human quality control [2]. These findings suggest that comprehensive AI involvement tagging would create valuable data for quality assessment and moderation.

2.6 Research Gaps and Contributions

Current literature lacks a comprehensive framework for categorizing and labeling the full spectrum of AI involvement in collaborative knowledge production. Existing tagging systems capture interface-level metadata but not the nature or extent of AI assistance, and research on AI disclosure has primarily focused on binary questions rather than graduated taxonomies. This work addresses these gaps by proposing a structured taxonomy that captures the spectrum from lightest assistance to heaviest involvement, organized into five conceptually coherent categories: Content Creation, Assistance and Verification, Metadata Suggestion, Media-Specific, and Human-Only Contribution. The case studies demonstrate practical application while revealing important limitations in current AI capabilities, contributing both theoretical framework and implementation guidance for platforms managing AI-assisted content creation.

3 Method

The goal of this work is to create a structured and transparent framework for labeling AI involvement in Wikipedia and Wikimedia Commons. To do this, a set of new tags is proposed, organized into conceptually coherent categories that correspond to different forms of human-AI collaboration. This approach enables extending the existing tagging system in a way that reflects contemporary editing practices.

3.1 Analysis of Current Tagging Limitations

The analysis began by reviewing the full set of existing Special:Tags [23] and examining how they represent tool usage in practice. Many of these tags document only high-level editing actions, such as whether the edit was marked as Tag:ContentTranslation or flagged as Tag:Rollback, but they provide no information about whether an AI system contributed to the text itself [27]. Prior analysis shows that tool-related tags are inconsistently applied, often underreported, and frequently fail to capture the actual degree of automation involved in an edit [1]. These limitations highlight the need for a more expressive tagging framework that can distinguish among different types and levels of AI assistance.

3.2 Taxonomy Design Rationale

In designing this framework, we considered multiple possible structures before arriving at the six-level taxonomy. A binary system (AI-involved / not AI-involved) was rejected as too coarse: it treats grammar polishing and full draft generation as equivalent, even though they carry entirely different implications for authorship and accountability. A three-level system (light / medium / heavy) was more promising but still collapsed meaningfully distinct actions: for example, AI-Bias-Removal and AI-Draft both fall under "heavy" despite representing very different degrees of human authorial control. We therefore propose six levels, where each boundary reflects a qualitatively distinct threshold; specifically, how much of the final text originated from the AI and whether a human verified it before publication. We acknowledge there is no single correct granularity; a simpler system may see higher adoption, and we discuss this tradeoff in Section 5.5. Our goal is to offer a taxonomy that is expressive enough for researchers and moderators who need fine-grained distinctions, while remaining adaptable for editors who prefer a simpler approach. The taxonomy was developed by the authors through iterative analysis of existing Wikipedia editing workflows, review of prior literature on AI disclosure, and examination of the full Special:Tags list. Empirical validation through editor surveys or pilot deployments remains an important direction for future work.

3.3 Proposed Tag Categories:

We organize the proposed tags into four major categories, as illustrated in Figure 1. This visualization demonstrates the spectrum of AI involvement, ranging from lightest assistance (such as grammar checking) to heaviest involvement (such as unreviewed AI-generated content). The four major categories are:

3.3.1 ***Content Creation Tags (Text):*** These tags capture cases where AI contributed directly to the text. The lightest level involves polishing or rewriting existing human-written text. Medium levels involve AI-research-assist, outlining, or expansion based on human prompts or bullet points. The heaviest levels involve AI-generated drafts, with or without human review.

Tag: AI-Polish (lightest): Tags such as tag:AI-Polish indicate edits where an AI system was used for light, surface-level refinement of text that was originally written by a human editor. These actions include correcting grammar and spelling, adjusting punctuation, smoothing sentence flow, or rephrasing wording for improved clarity and readability. Tools like Grammarly, QuillBot, or GPT's "fix grammar," "polish writing," or "rewrite sentence" features fall under this category. In some cases, editors may use AI systems to slightly rephrase ideas or introduce small additional clarifying details, but the AI does not contribute major new content, arguments,

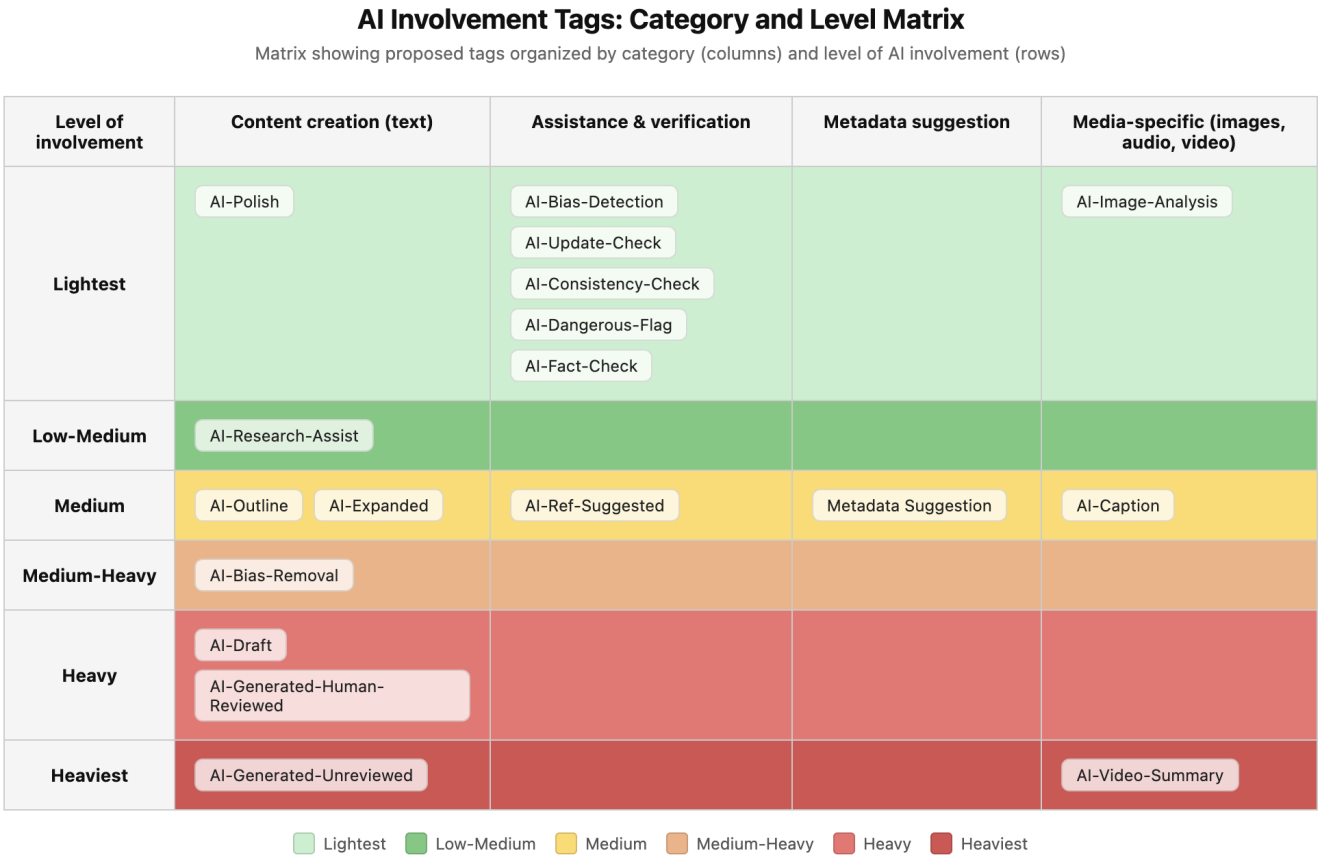


Figure 1: Proposed AI Tags by Category and Level of Involvement. The chart displays all proposed tags organized into four categories: Content Creation Tags (Text), Assistance and Verification Tags, Metadata Suggestion Tags, and Media-Specific Tags (Images, Audio, Video). The height of each bar indicates the level of AI involvement, from Lightest to Heavy/Heaviest. Color coding serves visual purposes only and does not indicate endorsement or legitimacy.

or structure to the article. The editor remains the primary author of the text, and the AI's role is limited to stylistic cleanup, readability improvements, or minor rewording. These tags therefore represent the lightest and least intrusive level of AI involvement on the content-creation spectrum, where AI assists with refinement rather than substantive writing.

Tag: AI-Research-Assist (Low–Medium)*:* A tag like tag:AI-Research-Assist would indicate that an editor relied on an AI tool to summarize or condense a longer external source as part of preparing content for Wikipedia. In this workflow, the editor provides the AI with substantial material—such as a lengthy policy document, academic paper, news article, or government report—and the AI generates a shorter, more digestible summary highlighting the key points. The editor may then use this AI-produced summary to inform or shape the text they add to Wikipedia. This tag makes it transparent that the summary or background understanding originated from an AI system rather than being fully synthesized by the editor, even though the human ultimately decides what to include and how to phrase it. AI-research-assist therefore represents a low-to-medium level of involvement, where the AI contributes informational condensation but does not independently write content from scratch or determine the overall narrative.

Tag: AI-Outline (Medium): This tag shows that the AI generated the initial outline, structure, or set of section headings for a piece of content, and the human editor then expanded that outline into full prose. For example, an editor might provide a topic—such as *"The economic impact of renewable energy adoption"*—to an LLM, and the AI produces a structured outline with suggested subsections and key points. The editor then uses this AI-generated structure as a foundation and writes the detailed explanations, arguments, or citations themselves. This tag makes it transparent that the conceptual organization originated from the AI, while the substantive text was written by the human.

Tag: AI-Expanded (Medium): This tag indicates that the human editor wrote a short outline, bullet list, or brief sentence, and the AI expanded it into a fuller explanation or full paragraph. For example, the editor might provide a simple outline such as "Causes of air pollution: vehicles, industry, agriculture", and the AI produces

a more detailed written section based on those points. In this workflow, the core ideas and structure originate from the human, but the AI generates the expanded prose, so AI-Expanded makes that distinction clear. This represents a medium level of AI involvement where the human provides the conceptual framework but delegates the actual writing to the AI.

Tag: AI-Bias-Removal (Medium–Heavy): When an editor uses an AI tool to rewrite biased, promotional, or otherwise non-neutral text in order to align it with Wikipedia's Neutral Point of View (NPOV) policy [28], the tag:AI-Bias-Removal should be applied. For example, an editor might take a paragraph that sounds overly positive, negative, or subjective and ask an AI tool to rewrite it in a more balanced, neutral tone before incorporating the revised version into the article. This tag makes it clear that AI assisted specifically with improving neutrality rather than generating new content.

Tag: AI-Draft (Heavy): This tag indicates that the first version of a sentence, paragraph, or even the whole article was initially written by an AI model. For instance, an editor might ask GPT or Claude to generate a paragraph about a topic, paste that text into Wikipedia, and then make some edits or add sources afterward. Even though the human made changes, the initial draft still came from the AI, and this tag makes that origin transparent.

Tag:AI-Generated-Human-Reviewed (Heavy): Applied when an AI model produces the initial text but a human editor subsequently reviews, corrects, and verifies the output before publication, this tag ensures transparency about content origins. In this workflow, the AI may generate a full paragraph, section, or even a draft article based on a prompt, while the human editor performs substantial oversight by checking facts, adding citations, refining tone, and ensuring alignment with Wikipedia's content policies. Although the human contributes meaningful edits, the content's origin remains primarily AI-generated. The presence of this tag signals that the editor exercised due diligence by reviewing the AI output, distinguishing it from unverified AI text and enabling future editors to understand the article's authorship history and potential risks.

Tag:AI-Generated-Unreviewed (Heaviest): Content generated by an AI system and published without any human review or verification receives this tag as a critical warning. While such edits are discouraged in Wikipedia due to concerns about accuracy, bias, misinformation, and unverifiable claims, this tag provides maximum transparency when unreviewed AI content enters the revision history. For instance, an editor might paste AI-generated text directly into an article without carefully reading it or checking the facts. This tag immediately alerts future editors, moderators, and researchers that the content may contain errors, hallucinations, or policy violations. Although the community strongly opposes unreviewed AI text in articles, explicit labeling remains essential for accountability, cleanup efforts, and understanding the edit's origin.

3.3.2 ***Assistance and Verification Tags***. These tags reflect scenarios in which AI provides support rather than authorship. They include bias detection, update checking, consistency checking, dangerous content identification, fact verification, and reference suggestion. These actions evaluate or flag content but do not write new text.

Tag:AI-Bias-Detection: When an editor employs an AI tool to detect biased or non-neutral language in existing text, this tag should be applied. The editor provides the text to an AI system, which identifies phrases or sentences that appear politically biased, culturally one-sided, promotional, or overly negative. The AI only highlights the potential bias and does not rewrite the content. This tag documents that the editor used AI to locate possible neutrality problems, and the editor may then revise the text themselves or use additional human or AI assistance to improve neutrality.

Tag:AI-Update-Check: This tag marks cases where an AI tool was used to verify whether information in an article remains current. For example, an editor might ask an AI system to check if a population number, a political office holder, or a scientific statistic is still accurate. The AI can then alert the editor that the information may need updating. While the AI identifies possibly outdated content, the human editor ultimately decides what changes to make.

Tag:AI-Consistency-Check: Applied when an AI tool examines whether information in an article is consistent across different sections, this tag ensures transparency in verification workflows (the process editors use to check and validate article content). The editor provides the text to an AI system, and the AI identifies mismatches such as conflicting dates, different numbers for the same statistic, variations in names or spellings, or contradictions in factual details. The AI does not change the text itself but alerts the editor to possible inconsistencies, which the editor may then correct manually or with additional assistance.

Tag:AI-Dangerous-Content-Flag: This tag signals that an AI tool was used to identify content that may be harmful or inappropriate. The editor submits text to an AI system, which flags material that could include violent descriptions, hate speech, threats, self-harm content, or adult themes. The AI does not rewrite or remove the text but simply alerts the editor that the material may require caution or additional review. AI assists only in detecting potentially dangerous or sensitive content, while the editor decides whether to revise, remove, or further evaluate the flagged material.

Tag:AI-Fact-Check: When an AI tool verifies or cross-checks facts in human-written text, the tag:AI-Fact-Check provides appropriate attribution. For instance, an editor might ask an LLM, "Are these dates accurate?" or "Is this information correct?" and use the AI's response to confirm or correct their wording before saving the edit. This tag clarifies that AI assisted in fact-checking but did not generate the actual content.

All tags in the Assistance and Verification category (from tag:AI-Bias-Detection through tag:AI-Fact-Check) represent the lightest level of AI involvement, as the AI only evaluates or flags human-created content and does not generate new text.

Tag:AI-Ref-Suggested (Medium): AI assistance in identifying references or sources is documented through the tag:AI-Ref-Suggested. For example, an editor might ask an LLM to recommend

reliable or primary sources related to a topic, review the suggestions, and then use those sources when adding information to Wikipedia. Importantly, this tag clarifies that AI assisted only with finding references, not with generating the actual content.

3.3.3 ***Metadata Suggestion Tags****:* These tags identify cases where AI suggests organizational metadata such as titles or subtitles. They capture instances where the structure or labeling of the article is influenced by AI systems rather than by human judgment alone.

Tag: Metadata Suggestion (Medium)*:* Many editors struggle to choose effective article titles or subtitles after drafting content, particularly when attempting to adhere to Wikipedia's naming conventions. An AI-assisted metadata suggestion feature could propose clear, neutral options based on the editor's written text. To ensure transparency, a tag such as "Tag: Metadata Suggestion" could indicate when a title or subtitle was generated or recommended by an AI model. This approach would serve multiple purposes: Transparency: Both editors and researchers can identify when AI has influenced the naming structure of an article. Educational Value: Newcomers gain a practical tool to learn how to create more accurate and well-structured headings. Quality Control: The community can evaluate and refine AI-generated suggestions while maintaining editorial standards. This feature would balance AI assistance with human oversight, helping editors craft better metadata while maintaining clear attribution of AI involvement in the editorial process.

3.3.4 ***Media-Specific Tags (Images, Audio, Video)****:* These tags extend the taxonomy beyond text to images, audio, and video. They include AI-assisted image analysis, automated caption generation, and AI-generated video or audio summaries. These workflows are increasingly common in Wikimedia and require metadata transparency similar to text-based edits.

Tag:AI-Image-Analysis (Lightest)*:* When an editor employs an AI tool to examine an image and identify objects, people, locations, or other visual features, a tag such as tag:AI-Image-Analysis helps document this automated contribution. For example, an editor might use an AI model to recognize landmarks or detect items in a historical photo before adding or improving the description. This tag makes it clear that the identification came from an automated system rather than the editor's own observation.

Tag:AI-Caption (Medium)*:* Mathias (2025) developed a minimal web application that extracts EXIF metadata from images and generates MediaWiki templates to support Wikimedia Commons uploads [15]. Building on this idea, whenever an image description is created or suggested by an AI system, the edit summary or an associated tag, such as tag:AI-Caption, should indicate the AI contribution. This would allow the system to trace which descriptions were produced with AI assistance, making it possible to classify them explicitly as AI-generated captions. Such transparency would support both community review and future research on the extent and impact of AI use in Wikimedia Commons.

Tag:AI-Video-Summary (Heaviest)*:* A tag like tag:AI-Video-Summary would indicate that an AI tool was used to generate a summary of video or audio content. Current AI systems typically create such summaries based on available textual metadata (such as video titles and descriptions) or transcripts, rather than by directly watching video footage or listening to audio. For example, an editor might use an AI model to generate a description of a documentary, interview, or news clip from its YouTube metadata or transcript before adding it to Wikimedia Commons. This tag helps reviewers understand that the summary was created by an automated system, and as demonstrated in the case study below, it is essential to document what source material the AI used to generate the summary.

3.3.5 ***Human-Only Contribution Tag****:* This tag addresses a distinct need: providing a clear, positive declaration that no AI tools were used in creating content. Given that many editors will inevitably fail to tag their AI use, the presence of this tag serves as a high-confidence signal that the contributor explicitly intended to declare human-only authorship, rather than simply forgetting to apply tags.

Tag: Human-Only-Contribution*:* When an editor explicitly declares that no artificial intelligence tools of any kind were used in creating, editing, researching, or refining their contribution, the tag:Human-Only-Contribution should be applied. This tag distinguishes between two scenarios that would otherwise appear identical: (1) content created without AI where the editor forgot or chose not to tag, and (2) content deliberately created without AI where the editor actively signals this fact.

By providing this explicit declaration, editors create transparency about their workflow and give readers, researchers, and moderators high confidence that the contribution represents purely human authorship. This tag is particularly valuable in contexts where establishing clear provenance matters, such as when Wikipedia content is used as training data for AI systems, when studying the impact of AI assistance on content quality, or when readers specifically seek human-authored content for verification purposes.

4 Demonstration

4.1 Demonstrating the Use of the AI-Polish Tag

This scenario reflects a common entry point for AI use among Wikipedia editors, particularly those using grammar tools like Grammarly before submitting edits [26]. This case study illustrates how the proposed AI-Polish tag can be applied to a real editing scenario. The analysis begins with a short, human-written paragraph describing Bangladesh's economy and recent political transition. The paragraph was intentionally written with several grammatical errors, misspellings, and subjective statements so that the effects of AI polishing would be easy to observe. The original text was:

- "Bangladesh is Considered to be one of the most important countries in south Asia because of having large population and its GDP mostly depends on Textiles. You will be surprised to know that One of the most clothes we get in the USA are exporting from Bangladesh but still they have a very poor infrastructure thats because of corruption. Our last prime minister Sheikh Hasia was the corrupted govt in the history of Bangladesh. So that is why on 5th August she has to resign from her PM position. Now Dr. Mohammad Yunus has taken over the power as a chief advisor of Bangladesh and situation

is getting better day by day. According to Jamanua Television Our next general election will be help on February 2026."

This text was then submitted to GPT-4 with the prompt "Polish this paragraph." The model produced the following revised version:

- "Bangladesh is considered one of the most important countries in South Asia because of its large population and its strong textile industry, which plays a major role in the country's GDP. Many of the clothes sold in the United States are manufactured in Bangladesh, although the country still struggles with infrastructure challenges, often linked to corruption. On 5th August, the former prime minister, Sheikh Hasina, resigned following widespread political pressure. Dr. Muhammad Yunus has since taken over as the Chief Advisor of the interim government, and conditions in the country have been gradually improving. According to Jamuna Television, the next general election in Bangladesh is expected to be held in February 2026."

The differences between the two versions demonstrate the type of transformation that an AI-Polish tag is designed to capture. The AI corrected capitalization errors, fixed run-on sentences, and improved verb agreement. It replaced subjective or accusatory language with more neutral, factual phrasing and corrected several spelling mistakes, such as "Jamanua" to "Jamuna." The revised paragraph also presents the information in a clearer and more organized structure, improving readability without introducing new claims. This example shows how AI tools can assist with stylistic refinement while leaving the substantive content and overall meaning under human control.

4.2 Demonstrating the Use of the AI-Bias-Detection and AI-Bias-Removal Tags

This scenario is directly relevant to Wikipedia's NPOV dispute resolution process, where editors are expected to identify and correct non-neutral language before articles can be cleared of neutrality dispute templates [28]. This second case study examines how two complementary steps in AI assisted bias detection and AI assisted neutral rewriting can support editors working on articles flagged for neutrality concerns. The 3ality Technica article was selected for this analysis, as it carries a neutrality dispute template. Articles in this category often contain a promotional tone, unverified claims, and other forms of imbalance that can be difficult for editors to systematically identify.

Step 1: Bias Detection (tag:AI-Bias-Detection): The analysis began by isolating a segment of the article that appeared to include marketing language. For example:

Original Text: "3ality Technica provided technology in two categories… The company's 3Play image processing systems provide dynamic image correction, image manipulation, and multiplexing and demultiplexing algorithms for transporting 3D high-definition content over conventional high-definition infrastructure channels."[20]

When asked to identify biased or promotional language, the AI highlighted several issues that aligned with common NPOV violations: the use of promotional descriptors such as dynamic image correction and image manipulation, performance claims presented as fact without supporting citations, highly technical product descriptions written in a style more typical of marketing materials than encyclopedic writing, and possible original research embedded within the description of proprietary systems.

This step demonstrates the value of **tag:AI-Bias-Detection** as a lightweight tool for surfacing specific locations where neutrality concerns may exist. Rather than rewriting text automatically, the tag documents that AI assisted with the diagnostic stage of identifying potential bias.

Step 2: Neutral Rewriting (tag:AI-Bias-Removal): After confirming the presence of non-neutral language, the AI was prompted to produce a neutrally phrased alternative. The instruction was: "Rewrite this passage in a neutral, Wikipedia-appropriate tone, avoiding promotional language and unverified claims."

AI-Rewritten Text: "3ality Technica developed stereoscopic 3D production tools, including camera-rig systems and software used for processing 3D imagery. These systems were applied in film, television, and sports broadcasting. The company's 3Play system included features for managing 3D footage, although comprehensive independent evaluations of its performance are limited."

The AI-generated revision offers several improvements: promotional phrasing was replaced with more factual and descriptive language, speculative performance claims were contextualized rather than stated as established fact, The overall tone more closely matches standard encyclopedic writing, and the paragraph now better separates verifiable information from interpretive or unverifiable claims.

Taken together, these two steps illustrate how a combined workflow: The first tag, AI-Bias-Detection, followed by the tag AI-Bias-Removal, can support editors working on content flagged for neutrality issues. Importantly, the tags also preserve transparency: they document not only that AI was involved in the revision but also the nature of that involvement. Future editors can see where the AI contributed, what type of assistance it provided, and where further human review may be needed.

4.3 Demonstrating the Use of the AI-Video-Summary Tag

This scenario reflects an emerging workflow in Wikimedia Commons, where editors increasingly rely on AI tools to generate descriptions for multimedia content that would otherwise require manual transcription [22]. This case study illustrates what an AI-Video-Summary tag captures in practice and why transparency about the source material matters. When an editor uses an AI tool to summarize a video, the tag:AI-Video-Summary should be applied. Importantly, the tag should also indicate what source material the AI used to generate the summary. Current AI systems typically produce video summaries based on available textual metadata — such as the video title and description — or from a provided transcript, rather than by directly watching or listening to the video content. A summary generated from a title and description alone may not fully capture the actual content of the video, while a summary based on a full transcript provides more comprehensive coverage of the spoken content, though it will still omit purely visual elements

such as charts or demonstrations. By making the source material explicit through the tag, editors, readers, and researchers can accurately assess the completeness and reliability of AI-generated video summaries and determine when additional human review of the original content is necessary

5 Discussion

The proposed tagging framework and accompanying case studies reveal several important insights about the current state and future trajectory of AI involvement in Wikipedia editing.

5.1 Practical Implications of the Tagging Framework

The three case studies demonstrate distinct patterns of AI assistance that the proposed taxonomy successfully captures. Case Study 1 illustrates that AI-Polish tags represent the most straightforward and least controversial category. The transformation from error-filled text to polished prose demonstrates clear value while maintaining human authorship of core ideas. This suggests that lightest-level tags may see high adoption rates, as editors using tools like Grammarly already understand these distinctions. Case Study 2 reveals a more complex workflow where AI assists in both diagnosis (bias detection) and remediation (bias removal). The ability to tag each step separately, first tag:AI-Bias-Detection, then tag:AI-Bias-Removal, provides transparency about the complete editorial process. This layered approach addresses concerns about accountability while documenting how AI tools can support Wikipedia's NPOV policy [28]. Importantly, the case study shows that AI-generated rewrites still require human judgment to verify factual accuracy and maintain encyclopedic tone. Case Study 3 exposes a critical gap between perceived and actual AI capabilities. Many editors and readers may assume that AI "video summaries" involve watching and understanding video content, when in reality current systems only process textual metadata or transcripts. This finding underscores why the proposed tagging framework must distinguish not just the type of AI involvement, but also the source material used. A tag:AI-Video-Summary should therefore specify whether it was generated from metadata alone, a transcript, or (in future systems) actual video analysis.

5.2 Addressing the Voluntary Compliance Challenge

The success of the proposed framework depends on voluntary adoption by Wikipedia editors. The moral crumple zone dynamic [6] and concerns about disclosure eroding trust [14] create significant challenges for voluntary compliance. This approach addresses these challenges through three mechanisms. First, the framework frames tagging as documentation rather than admission of wrongdoing. By creating tags for the full spectrum from lightest to heaviest involvement, the system normalizes AI assistance rather than stigmatizing it. An editor who uses Grammarly should feel no more reluctant to apply tag:AI-Polish than to note they edited on mobile. This approach aligns with established principles for maintaining contributor trust in disclosure systems [19]. Second, the framework emphasizes that labeled contributions are more valuable than unlabeled ones. Transparent metadata enables better research, improves content provenance tracking, and supports informed decision-making by Wikipedia consumers [4]. This reframes disclosure from a liability to a contribution to Wikipedia's mission of free, verifiable knowledge. Third, the framework acknowledges that complete adoption is unrealistic. Even partial tagging provides value by establishing baseline data about AI usage patterns, enabling detection of unusual patterns, and creating a cultural norm that may increase voluntary compliance over time. As noted in prior work [3], even partial voluntary adoption can establish important precedents for responsible AI use. The conceptual validity of the proposed categories is further supported by their structural parallel to existing Wikipedia disclosure norms. Just as editors already voluntarily apply tags for tools like ContentTranslation and VisualEditor, the proposed AI tags follow the same opt-in model.

5.3 Community Division and the Path Forward

The introduction noted the community's division between those advocating prohibition and those embracing AI tools. This tension reflects genuine concerns: AI systems can hallucinate facts, generate biased content, and produce text that appears authoritative while being factually incorrect. However, outright prohibition faces practical impossibility; AI tools are already embedded in everyday writing workflows, and detection remains unreliable. The proposed framework offers a middle path. Rather than asking "Should AI be allowed on Wikipedia?" the framework asks "How can AI involvement be documented transparently?" This shifts the debate from binary permission/prohibition to nuanced understanding of assistance levels. An editor using GPT to polish grammar poses different risks than one pasting unreviewed AI drafts, and the proposed taxonomy makes this distinction explicit. The framework also acknowledges legitimate use cases that prohibition would eliminate. As demonstrated by existing tools and research [5, 10, 16, 17], smaller language editions benefit from AI translation assistance, accessibility projects use AI-generated descriptions, and new editors learn Wikipedia's style through AI-assisted rewriting. By enabling rather than blocking these uses, while maintaining transparency, the framework preserves Wikipedia's openness while addressing accountability concerns.

5.4 Technical and Social Implications

Implementing this framework requires both technical infrastructure and social acceptance. Technically, MediaWiki must provide easy tagging mechanisms, dropdown menus, checkboxes, or semi-automated suggestions based on edit patterns [27]. Specifically, the proposed tags would be implemented using the ChangeTagsListActive hook, which allows new tags to be registered in the system [11]. Editors would apply them through a collapsible checklist in the EditPage form, and once applied, tags would appear in revision history and be accessible via the API using action=query&prop=revisions &rvprop=tags, enabling researchers to retrieve tagged edits programmatically. The interface should make tagging as frictionless as possible while ensuring editors understand what each tag means. Socially, successful adoption requires community buy-in across language editions. Following established governance principles [7, 8], the proposed framework must be vetted through Wikipedia's

established governance processes: discussion on Village Pump, feedback from relevant WikiProjects, and potentially pilot programs in specific topic areas or language editions. Different communities may adapt the framework to local norms while maintaining core transparency principles. The framework also creates new research opportunities. Once implemented, researchers can analyze how AI assistance varies across topic areas, language editions, and editor experience levels. Does AI involvement correlate with article quality metrics? Building on prior work demonstrating relationships between edit behavior and content quality [2, 13], AI usage patterns could similarly be studied to understand their impact on article quality. Do certain topics attract more AI assistance? How do different communities adopt and enforce tagging norms? These questions become answerable with comprehensive tagging data [4]. While this framework was designed with Wikipedia and MediaWiki in mind, the underlying taxonomy is platform-agnostic. Similar transparency needs exist in other collaborative knowledge platforms such as Wikidata, OpenStreetMap, and wiki-based communities, where AI tools are increasingly being used for content generation and moderation. Adapting this taxonomy to such platforms represents a promising direction for future work.

5.5 Granular vs. Simplified Tagging Approaches

The proposed framework offers 19 distinct tags organized across five categories, providing fine-grained classification of AI involvement. This granularity enables nuanced research and detailed provenance tracking. However, an alternative approach, using only high-level tags such as "AI-Involvement" and "Human-Only-Contribution," may offer easier adoption and lower cognitive burden for editors. Both approaches have merit. The detailed taxonomy supports researchers studying specific patterns of AI assistance and enables communities to make informed decisions about different types of AI use. Simplified catch-all tags reduce complexity and may see higher voluntary adoption rates. These approaches are not mutually exclusive; Wikipedia could implement both simultaneously, allowing editors to choose detailed tags when appropriate while providing simple options for those who prefer ease of use. Future research could examine which approach communities adopt more readily and whether combined systems achieve better transparency than either approach alone.

5.6 Limitations and Future Evolution

While the proposed framework provides comprehensive coverage of AI involvement types, several limitations remain:

Voluntary Compliance: *The system relies on editors honestly tagging their AI use. Without automated detection, underreporting remains a risk. Empirical validation through editor surveys, controlled editing experiments, or pilot deployments on specific WikiProjects remains a critical direction for future work.*

Limited Scope: *The case studies focus primarily on English Wikipedia and three major AI models (GPT-4, Claude, and Gemini). Editing practices and AI capabilities may differ across other language editions and AI systems.*

Tag Complexity: *The 18 proposed tags may overwhelm new editors. Simplified categories or decision trees may be needed.*

User Misclassification: *Voluntary tagging assumes editors accurately understand AI capabilities and their own usage patterns. However, users may mischaracterize AI outputs as "research assistance" or "fact-checking" when systems actually generated unreliable content through hallucination. This gap between perceived and actual AI functionality presents a challenge for voluntary disclosure systems, requiring clear tag definitions, educational resources, and potentially automated detection to supplement voluntary tagging.*

Evolving Technology: *Future AI systems may develop capabilities not captured by current categories, such as genuine video understanding or real-time collaborative editing.*

Cross-Cultural Variation: *Tag adoption may vary across language editions based on different community norms and AI tool availability.*

The proposed framework reflects current AI capabilities and usage patterns, but both will evolve. Future AI systems may develop genuine video understanding, multimodal content generation, or real-time collaborative editing assistance. The taxonomy must remain flexible enough to accommodate new forms of AI involvement while maintaining its core principle: transparent documentation of human–AI collaboration.

6 Future Work

The proposed tagging system requires integration with MediaWiki's existing infrastructure, including modifications to the editing interface and revision history displays [1]. Future work should focus on developing user-friendly mechanisms that allow editors to apply these tags without disrupting their workflow, such as dropdown menus or checkboxes within the editing interface. Additionally, automated detection systems could complement voluntary disclosure by flagging potential AI-generated content for editor review. Successful adoption depends on community buy-in and clear policy guidelines. Future work should involve consultation with Wikipedia editors across different language editions through surveys, community discussions, and pilot programs to test the framework's usability and effectiveness. Such studies would assess whether the proposed tags are applied consistently, whether editors find them intuitive, and whether the taxonomy's granularity matches real-world editing workflows. Once implemented, the tagging system will enable longitudinal research examining how AI assistance varies across topic areas, language editions, and editor experience levels, as well as whether AI-tagged edits differ in quality and neutrality compared to human-written content [4]. Important questions about tag governance remain, including who has authority to apply or modify tags (original editor only, or also moderators and reviewers), whether retroactive tagging should be permitted, and how tag disputes should be resolved. These policy decisions require community deliberation through Wikipedia's established governance processes.

7 Conclusion

This work suggests a broader tagging system to mark AI's role in Wikipedia and Wikimedia Commons. By distinguishing between light assistance, such as grammar polishing, and heavier involvement, such as full draft generation, the framework aims to improve

transparency and help editors, moderators, and researchers understand how content is produced [4]. The four-category taxonomy includes Content Creation Tags, Assistance and Verification Tags, Metadata Suggestion Tags, and Media-Specific Tags, which reflect the full spectrum of human-AI collaboration in modern editing workflows. Rather than banning AI-generated content, which is impractical and difficult to enforce [9], this approach embraces transparency as a viable path forward. The moral crumple zone dynamic identified by Elish (2019) highlights why voluntary disclosure is currently limited [6]. By establishing a standardized, non-punitive tagging system, honest reporting can be encouraged and a complete record of how Wikipedia's content is produced can be created. As AI systems become increasingly integrated into Wikimedia workflows [22, 25], such transparency mechanisms are essential for maintaining Wikipedia's commitment to verifiability and community oversight.

References

[1] Ashik Ahamed, Max Wang, Amity Ramona Mentis-Cort, and Jeanna Matthews. 2026. An analysis of Wikipedia's special tags and their implications for the nuanced spectrum between human edits and bot edits. In *International Conference on Virtual Learning*, Vol. 21. 15–26. doi:10.58503/icvl-v21y202601

[2] Sumit Asthana, Sabrina Tobar Thommel, Aaron Lee Halfaker, and Nikola Banovic. 2021. Automatically labeling low quality content on wikipedia by leveraging patterns in editing behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–23.

[3] Authority. 2025. AI System Disclosures. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-system-disclosures Accessed: 2025-12-11.

[4] Ben Chester Cheong. 2024. Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics* 6 (2024), 1421273.

[5] Maitraye Das, Alexander J Fiannaca, Meredith Ringel Morris, Shaun K Kane, and Cynthia L Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for AI-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[6] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).

[7] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.

[8] R Stuart Geiger and Heather Ford. 2011. Participation in Wikipedia's article deletion processes. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. 201–202.

[9] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[10] Isaac Johnson and Emily Lescak. 2022. Considerations for multilingual wikipedia research. *arXiv preprint arXiv:2204.02483* (2022).

[11] MediaWiki. 2025. Manual:Hooks/ChangeTagsListActive. https://www.mediawiki.org/wiki/Manual:Hooks/ChangeTagsListActive. Accessed: 2025-12-11.

[12] David B Resnik and Mohammad Hosseini. 2025. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in research* (2025), 1–13.

[13] Thorsten Ruprechter, Tiago Santos, and Denis Helic. 2020. Relating Wikipedia article quality to edit behavior and link structure. *Applied Network Science* 5, 1 (2020), 61.

[14] Oliver Schilke and Martin Reimann. 2025. The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes* 188 (2025), 104405.

[15] Mathias Schindler. 2025. *commonsimagedescription: Wikimedia Commons Image Analyzer*. https://github.com/MathiasSchindler/commonsimagedescription GitHub repository. Accessed: 2025-12-12.

[16] Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. 2024. Figura11y: Ai assistance for writing scientific alt text. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 886–906.

[17] Uzoma Ozurumba. 2025. A decade of consistent improvements to the Content Translation tool yields over two million Wikipedia articles. https://diff.wikimedia.org/2025/05/08/a-decade-of-consistent-improvements-to-the-content-translation-tool-yields-over-two-million-wikipedia-articles/ Accessed: 2025-12-11.

[18] Max Wang, Ashik Ahamed, and Jeanna Matthews. 2026. Exploring the Spectrum between Human Editing and Bot Editing with Wikipedia Metadata. In *Proceedings of the Wiki Workshop (13th edition)*. https://wikiworkshop.org/2026/paper/wikiworkshop_2026_30_exploring_the_spectrum_between_human_editing_and_bot_editing_with_wikipedia_metadata

[19] Kari D Weaver. 2024. The artificial intelligence disclosure (AID) framework: an introduction. *arXiv preprint arXiv:2408.01904* (2024).

[20] Wikipedia contributors. 2025. 3ality Technica. https://en.wikipedia.org/wiki/3ality_Technica Accessed: 2025-12-11.

[21] Wikipedia contributors. 2025. AI Is Tearing Wikipedia Apart. https://www.vice.com/en/article/ai-is-tearing-wikipedia-apart/ Accessed: 2025-12-11.

[22] Wikipedia contributors. 2025. Commons:Structured data/Computer-aided tagging. https://commons.wikimedia.org/wiki/Commons:Structured_data/Computer-aided_tagging Accessed: 2025-12-11.

[23] Wikipedia contributors. 2025. Tags. https://en.wikipedia.org/wiki/Special:Tags Accessed: 2025-12-11.

[24] Wikipedia contributors. 2025. Wikipedia: Case Against LLM-Generated Articles. https://en.wikipedia.org/wiki/Wikipedia:Case_against_LLM-generated_articles Accessed: 2025-12-11.

[25] Wikipedia contributors. 2025. The Wikipedia image/caption matching challenge and a huge release of image data for research! https://diff.wikimedia.org/2021/09/13/the-wikipedia-image-caption-matching-challenge-and-a-huge-release-of-image-data-for-research/ Accessed: 2025-12-11.

[26] Wikipedia contributors. 2025. Wikipedia:AI-generated content. https://en.wikipedia.org/wiki/Wikipedia:AI-generated_content Accessed: 2025-12-11.

[27] Wikipedia contributors. 2025. Wikipedia:Tags. https://en.wikipedia.org/wiki/Wikipedia:Tags Accessed: 2025-12-11.

[28] Wikipedia contributors. 2025. Wikipedia:WikiProject AI Cleanup/Policies. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup/Policies Accessed: 2025-12-11.