

Pluralistic Annotation of Cultural Weaponization on Wikipedia: A Multi-Annotator Case Study across Ukraine and Armenia

Camil Hamdane¹ Hanna Perekhoda² Emanuela Boros³
Stefanie Prezioso² Hamest Tamrazyan¹

¹Digital Humanities Laboratory, EPFL, Switzerland

²Institute of Political Studies, University of Lausanne, Switzerland

³L3i, University of La Rochelle, France

Abstract

Detecting the weaponization of cultural heritage on Wikipedia is not a straightforward classification task. Whether an edit manipulates a historical or cultural narrative often depends on context, interpretation, and the perspectives of those assessing it. We examine this challenge through a study of 400 English-language Wikipedia edits concerning Ukraine and Armenia. Three expert annotators independently classified each edit as *Potentially Manipulative*, *Not Problematic*, or *Ambiguous*, before resolving disagreements through structured discussion. We also compared their judgments with those of a large language model. More than one-third of the edits prompted disagreement, and discussion frequently produced decisions that could not be recovered through majority voting alone. The model aligned more closely with the experts’ deliberated decisions than with their initial majority votes, demonstrating that the construction of “gold” labels can substantially affect how model performance is understood. Our findings also show that *Ambiguous* captures genuine evidential and interpretive uncertainty rather than annotation failure. We argue that culturally sensitive annotation should preserve disagreement, treat ambiguity as a meaningful outcome, and use deliberation, not mechanical aggregation, to construct reference labels.¹

1 Introduction

Cultural heritage has become one of the key arenas of contemporary military conflict. However, it is not just a target for deliberate destruction. Cultural heritage can also be a weapon: it may be ruined, reconstructed, or reinterpreted to support political claims (Meskell, 2026). For instance, in the context of Russia’s war against Ukraine, cultural-historical heritage is often deliberately destroyed, while on

occupied territories it is repurposed to serve the narratives that deny the legitimacy of Ukraine as a distinct society and culture (Munawar and Symonds, 2023). Thus, heritage in modern conflict may be exploited by state and non-state actors for political, military, economic, and diplomatic purposes (Clack and Dunkley, 2023). It can be used as a symbol to claim territory, legitimize violence, deny the autonomy of a community, etc. Therefore, when we say that history and culture become “weaponized”, we mean that they are not merely used for political purposes, but that cultural-historical narratives are mobilized to legitimize aggression (Khislavski, 2022).

On the one hand, heritage should be understood not only as something tangible and material, but also as a set of values, memories, and identities through which communities understand themselves and their place in space and time. On the other hand, as numerous studies have shown, contemporary conflicts increasingly target cognition, identity, values, and collective memory (Meriläinen, 2026). For this reason, the digital dimension of cultural heritage weaponization is crucial. In the context of political and military conflict, cultural heritage is not only attacked or appropriated in physical space. It can also be targeted through various digital infrastructures, including collaboratively produced knowledge systems such as Wikipedia (Meriläinen, 2024).

This article addresses precisely this problem. It constitutes the first stage of a broader research project on the weaponization of cultural heritage and historical memory in English-language Wikipedia articles concerning the culture and history of Ukraine and Armenia. The wider project aims to develop an LLM-assisted analytical pipeline capable of identifying subtle manipulative strategies in Wikipedia revisions, both semantic and structural. The present article focuses on the manual validation of a “gold” standard subset of

¹The annotated dataset is publicly available (Hamdane et al., 2026).

revisions for which we propose a three-category annotation schema: *Potentially Manipulative*, *Not Problematic*, or *Ambiguous*.

This validation stage is important because we do not assume that every contested edit² on a politically sensitive page constitutes manipulation. A revision becomes potentially manipulative if the change alters the meaning, framing, interpretation, contextual understanding or readability of the text in a way that may be derogatory, manipulative or ideologically significant. The article examines the capacities and limits of LLMs and expert assessment in determining when a revision contributes to the strategic distortion, reframing, or erasure of cultural narratives. It provides a foundation for developing a taxonomy of manipulative strategies.

Ukraine and Armenia were selected because they make this problem visible in different but comparable ways. Ukraine is a central case of heritage weaponization because Russian military aggression has been accompanied by narratives that explicitly deny Ukrainian political and cultural distinctiveness (Khislavski, 2022). In this context, disputes over the Ukrainian past are part of a broader struggle over whether Ukraine is perceived as a political subject in its own right or as a derivative element of Russian imperial culture and history. Armenia is a second case in which contested heritage and war are deeply entangled. Destruction of Armenian memorial sites in the Nagorno-Karabakh/Artsakh conflict, as well as attempts to deny Armenian historical links and restrict access to those sites after the 2020 war, show how heritage becomes part of military conflict (Smith et al., 2023). These two cases differ in scale and international visibility, but they share a common feature: in both, cultural heritage and historical memory are not just background to conflict but active components of it.

Wikipedia plays a distinctive role in this process. It is a collaboratively produced and continuously revised knowledge infrastructure that functions both as a widely consulted source of information and as a battleground for competing narratives (Saez-Trumper, 2019; Aragón and Saez-Trumper, 2021). The article focuses on English Wikipedia because it reaches a global audience and often functions as an entry point for automated knowledge systems. Therefore, a distorted article on Wikipedia may circulate beyond the platform through search engines

and artificial intelligence (AI) systems.

The methodological contribution of this article is to treat manual validation as an epistemic procedure rather than a merely technical operation: when LLMs are used to identify potentially manipulative Wikipedia revisions, their outputs must be reviewed by experts who have a broader vision of the socio-political context. The remainder of the paper is organized as follows: Section 2 situates the study within research on heritage weaponization and perspectivist NLP; Section 3 presents the corpus, annotation task, and LLM classifier; Section 4 reports inter-annotator agreement, LLM performance, and error analysis; and Section 5 draws implications for annotation schema design in culturally sensitive NLP tasks.

2 Related Work

2.1 Cultural Heritage Weaponization

Research on heritage in conflict has moved from documenting destruction toward analyzing the political uses of heritage: as propaganda, as legitimation, and as a resource mobilized by state and non-state actors (Clack and Dunkley, 2023; Munawar and Symonds, 2023; Meskell, 2026). Work on cognitive warfare extends this to the informational domain, where the targets are memory, identity, and values rather than monuments (Meriläinen, 2026, 2024). On the documentation side, remote monitoring initiatives such as Caucasus Heritage Watch have tracked the condition of Armenian monuments in and around Nagorno-Karabakh (Smith et al., 2023), and digital memorialization practices have been analyzed as defensive counterparts to informational aggression (Graf, 2025).

2.2 Computational Approaches to Wikipedia Manipulation Detection

Research on Wikipedia vandalism has largely focused on detection, especially machine learning and feature analysis (Tramullas et al., 2016). But problematic interventions may be covert and difficult to distinguish from good-faith editing: undisclosed paid editing evades community scrutiny (Joshi et al., 2020), and automated adversarial edits can be contextually relevant, evading conventional detection while still affecting reader perception (Sharevski et al., 2020). Governance capture represents a distinct risk: a self-governing community can be taken over by a small group of administrators who introduce systemic bias

²Throughout this paper, a *contested item* is an edit for which the three annotators did not assign the same label during the initial independent annotation round.

and disinformation, as documented for the Croatian Wikipedia (Kharazian et al., 2024). In the Ukrainian context, Makovska (2025) likewise distinguishes knowledge manipulation from vandalism, detecting narrative-level interventions in Wikipedia edits. Recent computational research has examined cultural heritage weaponization directly. One study analyzed the users responsible for weaponized edits alongside Wikipedia governance and policy signals (Garambois et al., 2026). More broadly, Tamrazyan et al. (2026) conceptualize cultural heritage weaponization as a symbolic and epistemic intervention distinct from propaganda, disinformation, vandalism, and censorship. The study constructs a corpus of more than 70,000 salient revisions using LLM-generated silver annotations and expert-validated “gold” annotations, finding that weaponization rarely depends on explicit falsehoods and instead often emerges through accumulative, procedurally legitimate edits.

Wikipedia’s own policies and practices (editor vigilance, bots, assisted editing tools, etc.) constitute a comparatively developed defense system (McDowell and Vetter, 2020; Geiger and Ribes, 2010), yet the platform functions simultaneously as a perceived source of ground truth for fact-checking and as a battleground for competing narratives (Saez-Trumper, 2019; Aragón and Sáez-Trumper, 2021). AI is becoming an active agent in this ecosystem, raising questions about bias, authenticity, and authority in relation to history and memory (Tutuncuoglu, 2024). A consistent finding across this literature is a gap between blatant vandalism, which community tools catch quickly, and subtle, contextually plausible manipulation, which evades both automated and community detection, the gap in which culturally weaponized editing operates.

2.3 Pluralistic Annotation and Disagreement

Perspectivist NLP treats disagreement between annotators as a legitimate object of study rather than noise to be corrected (Plank, 2022): human label variation is often systematic, rooted in annotator background, task framing, and genuine ambiguity, and should be modeled rather than aggregated away. Surveys of learning from disagreement (Uma et al., 2021) and the CrowdTruth line of work (Aroyo and Welty, 2015) document how majority-vote consensus obscures whose judgment a “gold” standard encodes and discards legitimate minority interpretations. On the modeling side, LLMs have been

found effective as annotation *assistants* but unreliable as independent annotators (Gu et al., 2026). We accordingly use an LLM as a non-human comparison point rather than as a substitute annotator. Our study instantiates the conditions under which the perspectivist argument is strongest, interpretive judgments on politically sensitive material, and adds a resolution mechanism (structured team discussion) that, as we show, produces decisions not recoverable from votes alone.

2.4 Cultural Bias and Wikipedia

Beyond individual malicious edits, Wikipedia’s content reflects the composition and dynamics of its editing communities. Early work documented trolling as a distinct, socially motivated phenomenon beyond vandalism (Shachaf and Hara, 2010), and large-scale analysis of Wikipedia hoaxes showed that fabricated content can survive and circulate when it mimics legitimate form (Kumar et al., 2016). In contested geopolitical contexts, these dynamics acquire a systemic dimension: the Croatian Wikipedia case demonstrates how a capturable governance structure can encode a nationalist perspective across an entire language edition (Kharazian et al., 2024), and recent work on modeling source reliability across language editions (D’Ignazi et al., 2026) underscores that citation practices, which sources count as reliable and for whom, vary culturally rather than being neutral infrastructure. Our task sits at the intersection of these literatures: the manipulation we annotate is rarely hoax-like fabrication, but the culturally situated reweighting of narratives, sources, and framings.

3 Data and Methodology

3.1 Data Collection

Article curation. We constructed two country-specific article lists through a three-stage pipeline (Table 1). We first selected the seed articles for their relevance to prominent disputes involving cultural heritage, historical narratives, identity, or sovereignty in the regions studied (Appendix A). We expanded this initial set by following internal Wikipedia links to a depth of one. No keyword search or keyword matching was used: a zero-shot LLM³ instead assessed the resulting candidate arti-

³GPT-OSS-120b (OpenAI, 2025) was used for article-relevance filtering; the separate revision-classification experiment used Qwen.

cles against predefined topical inclusion and exclusion criteria. The inclusion criteria covered cultural heritage and identity, contested history, geopolitical disputes, religion, ethnic and demographic tensions, naming and terminology disputes, political figures, and governance or sovereignty. Articles retained by this filter then received a verification score on a 1–5 scale; we kept those scoring at least 3.

Measure	Ukraine		Armenia	
	<i>N</i>	%	<i>N</i>	%
<i>Article curation</i>				
Seed articles	5	—	8	—
After link expansion ($d = 1$)	2,720	—	4,128	—
After LLM relevance filtering	721	26.5	1,109	26.9
After verification (score ≥ 3)	503	69.8	851	76.7
<i>Verification-score distribution</i>				
Score 1–2	218	30.2	258	23.3
Score 3	246	34.1	205	18.5
Score 4–5	257	35.6	646	58.3
Mean score	3.21		3.48	
<i>Revision corpus</i>				
Total revisions	11.6M		12.9M	
Mean revisions per article	~23,000		~15,000	
Annotated sample	200		200	

Table 1: **Article curation and corpus statistics.** Percentages for LLM filtering indicate retention relative to the link-expanded set; percentages after verification indicate retention relative to the LLM-filtered set. Verification-score percentages are calculated over all LLM-filtered articles.

Revision fetching. For each curated article, we fetched the full revision history via the MediaWiki API, including editor username, edit comment, revision size, tags, and revision ID. Unified diffs were computed locally from successive revision texts. The resulting datasets contain approximately 11.6 million revisions across 503 Ukraine articles and 12.9 million revisions across 851 Armenia articles.

Notably, the revision density differs substantially between the two corpora: Ukraine’s articles average $\sim 23,000$ revisions each versus $\sim 15,000$ for Armenia. Despite having $1.7\times$ more articles after curation, Armenia’s corpus contains only $1.1\times$ the total revisions. This suggests that the higher-scoring Armenian articles are on average shorter and less edited than their Ukrainian counterparts, so the two datasets yield revision corpora of comparable scale despite the disparity in article counts.

3.2 Dataset Construction

From these full revision corpora, we constructed a human-annotatable sample of 200 edits per country (400 total), released as an open dataset (Hamdane et al., 2026). The sampling pipeline applied four

filtering criteria to ensure every edit was readable, self-contained, and carried potential semantic content:

1. **Talk page edits.** Edits containing Wikipedia user signatures (`[[User:...]]`) or UTC timestamps were removed, as they represent editor discussion rather than article content.
2. **Length.** Edits whose raw unified diff exceeded 5,000 characters were excluded to keep the annotation task tractable for human readers.
3. **Empty diffs.** Rows with missing or whitespace-only diff text were dropped.
4. **Encoding artifacts.** Diffs containing UTF-8 text misdecoded as Windows-1252 (mojibake) were repaired where possible and excluded otherwise.

Each of the 400 final edits includes the raw unified diff, page title, and when available, the editor’s comment and the article section. We also made available during annotation a preprocessed representation that strips wiki markup, resolves links to display text, and isolates added and removed passages for easier reading for humans.

The 200 edits per country were randomly sampled from the output of the verification pipeline (Section 3), which scores articles by weaponization relevance. Because the sample is drawn from articles already filtered for high cultural-heritage sensitivity, the prevalence of contested and potentially manipulative edits in our sample likely exceeds that of a random draw from Wikipedia’s full revision history. The annotation of contested items (those lacking unanimous agreement in the independent round) required approximately three hours of joint discussion per country (six hours in total), with straightforward cases resolved quickly and complex multifaceted edits demanding extended deliberation.

3.3 Annotation Task

Annotators assigned each edit one of three labels:

- **Potentially Manipulative (P):** the edit alters meaning, framing, or interpretation in ideologically significant ways.
- **Not Problematic (N):** the edit does not meaningfully affect interpretation.

- **Ambiguous (A):** impossible to classify without additional context (e.g., image-only edits).



Figure 1: **Example of a multifaceted edit.** The diff wrongly removes the mention of the Jewish Yishuv irregular units from the infobox but the same line also carried the flag decoration of the State of Israel, a state that did not exist until 1948. The team classified the removal as *Potentially Manipulative*.

The judgment targets the *effect* of a change on meaning, framing, contextual understanding, or readability. The LLM is also asked to judge the actual correctness of the change; something that is ideologically loaded but factually correct is NOT manipulative. We used *Ambiguous* on samples where the available context was insufficient to assess the change (e.g. visual edits, or references to external links). After an initial round of independent labeling, all items lacking unanimous agreement were discussed jointly and a team decision was recorded. The discussion was synchronous: annotators had access to each person’s initial independent labels as well as the LLM output from a preliminary classification run. The LLM output was shown by design to serve as context on convoluted edits, rather than as a tiebreaker. We are aware that this process might introduce circularity (see §6). The final decisions were reached by consensus after input from each annotator, with deference to the annotator with the strongest relevant linguistic or historical proficiency for contentious cases. Through this iterative process, the annotators refined their shared understanding of the task.

Further, the task proved challenging. Edits that one annotator saw as clearly manipulative sometimes appeared benign to another. Annotators’ individual understandings of what constitutes manipulation, shaped in part by their backgrounds and relationships to the histories and politics at stake, inevitably influenced their judgments even under a common guideline. The annotator team included one expert on Armenian history, one expert

on Ukrainian history, and a third annotator serving as a cross-country control. Determining what counts as manipulation is itself an interpretive act, not a mechanical classification.

The difficulty was compounded by *multifaceted* edits that combine several distinct changes within a single revision. Figure 1 illustrates a representative case. To handle such cases, we adopted a conservative labeling policy: any edit containing at least one manipulative change is labeled *Potentially Manipulative*. This minimizes false negatives for multidimensional edits, at the cost of flagging some benign changes for review.

3.4 LLM Classifier

We classified the same 400-edit sample with a single-stage few-shot LLM classifier to assess how well an automated pipeline aligns with human judgments. We chose the Qwen 3.6-35B-A3B model (Qwen Team, 2026)⁴, a 35-billion-parameter instruction-tuned model with mixture-of-experts (MoE) routing. Each call received the raw unified diff prefixed with available metadata (editor comment, article section, page title), and was prepended with 18 few-shot examples (5 *Potentially Manipulative*, 10 *Not Problematic*, 3 *Ambiguous*).

The system prompt defines the three label categories with decision rules and common patterns, instructs the model to judge the *effect* of the change rather than the editor’s intent, and includes a calibration toward *Not Problematic* in borderline cases. The 5/10/3 class balance in few-shot examples was designed to approximate the expected label distribution (manipulative edits are a minority) and to bias the model conservatively toward *Not Problematic* in borderline cases. The model returns a label followed by a two-sentence explanation⁵. The full prompt is reproduced in Appendix C, the inference configuration in Appendix D, and the few-shot examples in Appendix E.⁶

⁴We selected Qwen3.6-35B-A3B because its open-weight availability enabled self-hosted inference within our computational resources. The model serves as a case-study classifier; we do not claim that it is optimal for this task.

⁵The model returned a class label for all 400 edits; no refusals or missing classifications were observed.

⁶The prompt’s *Potentially Manipulative* category is intentionally broader than the conceptual definition of weaponization: it also covers pure vandalism and content disruption, since in a monitoring setting any harmful edit should be flagged for review.

3.5 Evaluation Metrics

To assess agreement among the three human annotators, we report Krippendorff’s α for nominal labels. Krippendorff’s α measures agreement among multiple annotators while correcting for agreement expected by chance, with $\alpha = 1$ indicating perfect agreement and $\alpha = 0$ indicating chance-level agreement. We calculate ternary agreement over the complete label set and binary agreement over annotations assigned either P or N (Ambiguous labels are excluded from the binary calculation). We additionally report pairwise Cohen’s κ to show how agreement varies across annotator pairs.

We then evaluate the LLM classifier against the discussion-derived team labels produced by the three-way consensus process described in Section 3.3. As our primary metric we report Cohen’s κ between the LLM predictions and the team’s consensus decision. Unlike raw accuracy, κ corrects for chance agreement and is sensitive to the label distribution, making it appropriate for a task where the three classes occur at substantially different rates.

We additionally report the F1 for the positive class, *Potentially Manipulative*. In a monitoring setting, the cost of missing a manipulative edit (false-negative) is higher than flagging a benign edit for review (false positive): an undetected framing change persists in the live article and may propagate through downstream knowledge systems, whereas a false alarm incurs only a human sanity check. F1 on the positive class captures this asymmetry better than overall accuracy or macro-averaged F1, both of which treat all errors as equally consequential. For the contested-subsets analysis (§4.2), we also report exact-match agreement (accuracy) with confidence intervals (CI). We also report macro-averaged F1 across the three classes.

4 Results

4.1 Why Consensus Was Necessary

The difficulty of the task is reflected in the initial round of independent annotation. On the Armenian subset (200 items, three annotators), 77 items (38.5%) lacked unanimous agreement: 75 split 2-1 and two split three ways. On the Ukrainian subset, 70 items (35%) lacked unanimous agreement: 65 split 2-1, five split three ways. In roughly two-thirds of cases, annotators converged immediately; in the remaining third, their judgments diverged

enough to require joint discussion.

Table 2 reports agreement coefficients computed from the full label matrix, presented here not as evaluation metrics but as evidence that the task is inherently interpretive.⁷

	Armenia	Ukraine
α ternary (P/N/A)	0.42	0.41
α binary (P/N)	0.66	0.55
κ Ann. 1-Ann. 2 (ternary)	0.55	0.65
κ Ann. 1-Ann. 3 (ternary)	0.48	0.35
κ Ann. 2-Ann. 3 (ternary)	0.28	0.29

Table 2: IAA as evidence of task difficulty ($n = 200$ per country). The binary–ternary gap shows that a large share of the difficulty is concentrated in the *Ambiguous* category; ternary agreement is nearly identical across the two conflict contexts.

The binary–ternary gap is the clearest signal: agreement is moderate when annotators are forced into P/N (Armenia $\alpha = 0.66$, Ukraine $\alpha = 0.55$) and drops sharply once “*I cannot decide without more context*” is a legitimate answer (Armenia $\alpha = 0.42$, Ukraine $\alpha = 0.41$). The ternary agreement is identical across the two conflict contexts, suggesting that the difficulty is a property of the annotation task rather than of either corpus.

Per-annotator label distributions diverge substantially, with Ambiguous usage rates ranging from 6% to 20% in Armenia and 3% to 13.5% in Ukraine. Critically, the Ambiguous label was originally used by annotators as a *fallback* when they were unsure of their judgment; only through the discussion process did the team converge on its current function as a substantive category for edits that cannot be classified from the available context. The Ukrainian subset reinforces this picture: 70 items (35%) were flagged for discussion. The same annotator-level structure in lone-dissenter counts emerges independently in both countries (Armenia: 8, 26, 41; Ukraine: 7, 13, 45), consistent with each annotator applying a distinct threshold for flagging edits as potentially manipulative rather than noise.

These agreement patterns motivated our decision to use discussion-derived consensus as the final label for every contested item, rather than majority vote. The IAA coefficients serve as context for the difficulty of the task, not as a performance target.

⁷The full matrix combines the three independent labels recorded for the contested items in the consensus-tracking file with the unanimous items.

4.2 LLM Classification Performance

We evaluate the LLM classifier against the team’s discussion-derived consensus labels, treating majority-vote agreement as a diagnostic counterfactual rather than a co-equal reference.

On the full 400-edit sample (both countries combined), the LLM achieves Cohen’s $\kappa = 0.69$ and an F1 score of 0.74 (precision = 0.69, recall = 0.80) on the positive class (*Potentially Manipulative*) against the team’s consensus labels. Macro-averaged F1 over the three classes is 0.79 (P: 0.74, N: 0.92, A: 0.72). For comparison, against the initial independent annotations (before discussion-based resolution, averaged across the three individual annotators’ labels), the same model achieves $\kappa = 0.56$ and F1 = 0.69. This gap of 13 points of κ , 5 points of F1, reflects the difference between raw annotator judgments and the considered labels produced through deliberation: the LLM aligns substantially better with the latter.

Of the 23 edits the team marked *Ambiguous*, the LLM assigned the same label to 14 (61%); the remaining 9 were classified as *Potentially Manipulative* (3) or *Not Problematic* (6), yielding an *Ambiguous*-class recall of 0.61 with precision of 0.88. This tendency to resolve rather than flag ambiguity mirrors the error patterns discussed below. On the contested subsets (the items where the three annotators lacked unanimous agreement, see §4.1) this pattern is even more pronounced. Of the 77 contested Armenian items, 75 split 2–1 and 2 were three-way ties; we restrict majority-vote calculations to the 2–1 splits where a genuine majority exists. On these 75 items, the LLM’s label matched the majority vote in 54 cases (72%). Against the team’s discussion-derived decision, it matched in 60 of 75 cases (80%).

This gap is the most consequential finding in our evaluation. The LLM agrees with the team’s *considered* judgment directionally more often than with the mechanical aggregate of first-pass labels—meaning that evaluating the model against majority-vote “gold” labels would substantially understate its alignment with expert consensus, and more fundamentally, that the two references embody different theories of what the ground truth is. On this subset, majority vote and team decision diverge in 16 of 75 cases (21%): for roughly one in five contested edits, aggregation alone would have produced a different “gold” label than deliberation did.

The Ukrainian contested subset shows the same

direction: of 70 contested items, 65 split 2–1 (5 three-way ties excluded). On these 65 items, the LLM matched the majority vote in 49 cases (75%) and the team’s discussion-derived decision in 54 cases (83%). With $n = 65$ –75 items per country, the agreement gaps between deliberation and majority vote are directional rather than precisely estimated. The team decision departs from the majority vote in 14 of 65 cases (22%), nearly identical to the Armenian rate (21%). That two independently annotated conflict contexts produce the same vote-deliberation divergence suggests the phenomenon is a property of the task, not of one corpus.

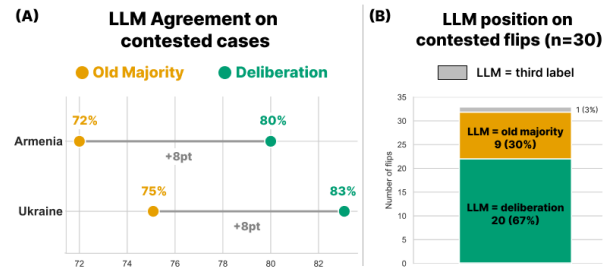


Figure 2: **LLM classification on contested items: agreement with majority vote vs. deliberation.**

Contested items are restricted to 2–1 splits; three-way ties (no majority) are excluded.

A: Agreement between the LLM and two alternative reference labels: the initial majority vote and the discussion-derived team decision, shown separately for Armenia and Ukraine. **B:** LLM outcomes for cases in which deliberation overturned the initial majority vote, distinguishing whether the model aligned with the deliberated decision, the initial majority, or a third label.

Across both countries, the LLM’s labels coincided with the post-deliberation label on 20 of the 30 items where deliberation overturned the initial majority vote (67%). Two factors shape this convergence: the discussion may have surfaced a judgment that the model also found plausible, and annotators could consult the model’s preliminary output while deliberating, so the two labels are not fully independent (§6).

However, agreement with the deliberated labels is 80–83%, well short of the near-perfect agreement we would expect if the team had simply copied the model. The team also left 49 model errors standing (§4.3). We therefore treat the deliberated labels as our best available reference. Figure 2 summarizes the agreement patterns and the model’s anticipation of deliberated decisions.

4.3 Error Analysis

A structural pattern emerges from the model’s errors against the team’s deliberation consensus. Treating *Potentially Manipulative* as the positive class against the other two labels, the model produces 26 false positives and 15 false negatives across the 400 items; in three-class terms, it disagrees with the deliberated label on 49 items (12.3%). The higher recall relative to precision (0.80 vs. 0.69 on the positive class) is consistent with the prompt’s conservative bias toward *Not Problematic* in borderline cases: the model misses fewer manipulative edits than it overflags.

4.3.1 False Positive Archetypes

Two patterns account for nearly all false positives:

Correctness-test failure. The prompt instructs the model that “*ideologically loaded but factually correct is NOT manipulative*”. The LLM systematically fails this clause because it struggles to verify factual correctness, so it proxies with topical sensitivity. Removing “alleged” from “alleged Armenian Genocide,” removing “wrongly” from a sentence about Jews being blamed for the Tsar’s assassination, and adding an incriminating quote to the *Operation Nemesis* article are all flagged as manipulative. The model sees sensitive content and cannot distinguish weaponization from its repair. The classifier is *symmetric* in a task where expert judgment is *asymmetric*: humans weigh who benefits from a framing change; the model, unable to verify correctness, defaults to flagging any edit touching sensitive semantics.

Toponym-rule overreach. The prompt’s “toponym changes... regardless of direction” clause fires on routine display and transliteration fixes. A link-display update from “Turkey” to “Türkiye” is flagged despite being a Wikipedia-mandated change. Yet the overreach is asymmetric: the same clause that catches routine updates also misses genuine toponym manipulation. A systematic update from “Kiev” to “Kyiv”, the canonical example of transliteration weaponization in the Ukraine context, is dismissed as “standard maintenance”; a city-name transliteration from “Mukačevo” to “Mukacheve” is similarly cleared as a “minor formatting adjustment.” The model follows categorical clauses literally and cannot distinguish routine maintenance from manipulative erasure.

4.3.2 False-negative Archetype: Procedural-Legitimacy Credulity

Every instance of false-negative reasoning invokes Wikipedia-legitimacy surface features rather than engaging with content. Aliyev’s claim of “liberated 16 villages” sourced to mod.gov.az is accepted as “properly attributed... standard factual update”; the erasure of the Armenian river name Yeraskh (Arax) is explained away as “corrects a character encoding issue”; a factually wrong date change (1931→1939 for the Christ the Saviour Cathedral demolition) is cleared as “minor maintenance” despite the model noticing the inaccuracy. This pattern confirms the project’s own prior finding that cultural weaponization “emerges through accumulative, procedurally legitimate edits”: the residual errors diagnose precisely where automation cannot replace deliberation; the model is fooled by exactly the signature the previous works identified.

4.3.3 Ambiguity-Error Archetypes

LLMs are known to struggle at recognizing the boundaries of their own knowledge (Yin et al., 2023). The model struggles to perceive that an edit’s interpretation depends on an image it cannot see or a template pointing to an off-page dispute. Two cases make this concrete. In *Culture of Azerbaijan*, an edit removed a Dispute about template⁸—the removal deletes a signal of contestation rather than content, and whether it is manipulative depends on materials no longer accessible. In *Flag of Armenia*, an edit removed an unsourced gallery of similar flags; the prompt explicitly instructs the model to classify visual-only edits as *Ambiguous*, yet the model, unable to perceive the images, dismissed it as “standard maintenance that does not alter the article’s narrative.” Both were recorded as *Ambiguous* by the team.

5 Discussion

5.1 *Ambiguous* as a First-Class Label

The binary–ternary α gap, the spread in per-annotator ambiguity rates, the evolution of the label’s meaning over the course of the task, and the existence of a multi-annotator ambiguous core together indicate that *Ambiguous* is not annotation failure but a substantive judgment. Schemas that omit such a category force annotators to launder un-

⁸Template:Dispute about flags contested and debated claims on Wikipedia. https://en.wikipedia.org/wiki/Template:Dispute_about

certainty into P or N, inflating apparent agreement and handing models a falsely confident training signal. For culturally sensitive tasks, we argue the label should be first-class: defined in guidelines, preserved in releases, and reported in evaluations rather than dropped.

5.2 Why Majority Vote Erases Minority Perspectives

Majority vote and deliberated team decisions diverged on 21% (Armenia) and 22% (Ukraine) of contested items where a genuine majority exists. A vote treats the three labels as interchangeable evidence; discussion does not; it lets an annotator articulate context the others missed, and the record shows the team sometimes siding with the initial minority. Aggregation by vote therefore does more than compress information: it systematically discards exactly those judgments that rest on perspective-specific knowledge. Our LLM result sharpens the point: the model tracked deliberated decisions, so a majority-vote benchmark would have misdescribed both the task and the model.

6 Conclusions & Future Work

This study provides a foundation for examining cultural heritage weaponization on Wikipedia through pluralistic human annotation and LLM-assisted analysis. Future research should extend the evaluation to additional conflicts and language editions to assess the generalizability of the findings, while developing a more detailed taxonomy of Wikipedia-specific weaponization strategies. Future work should also replicate the deliberation under blinded conditions to isolate any anchoring effect of model exposure, and extend the annotator pool and the range of evaluated models.

Tool-assisted approaches could further support both annotators and language models by retrieving cited sources, relevant revision histories, and other contextual evidence needed for claim verification. Although such grounding may reduce factual and contextual errors and improve annotation confidence, detecting culturally situated manipulation at scale will continue to require human expertise, explicit treatment of uncertainty, and deliberative judgment. In deployment, a model of this kind is best positioned as a conservative triage assistant that surfaces candidate edits for human review rather than as an autonomous judge.

Limitations

Agreement statistics are computed from the consensus-tracking exports. Per-annotator labels used for IAA coefficients and lone-dissenter counts are drawn from the initial independent annotation round; discussion-derived consensus labels are drawn from the final “gold” files for each country. The annotator pool is small (three experts) and not demographically balanced; asymmetries partly reflect individual annotation styles. The LLM comparison uses a single model and prompt configuration, with full-corpus evaluation against consensus labels ($\kappa = 0.69$, $F1 = 0.74$ on the positive class). The study covers English Wikipedia and two conflict contexts; generalization to other languages and conflicts is untested.

Beyond these design constraints, the errors documented in Section 4.3 reflect limitations of the evaluated model and prompt configuration. Clearer decision rules and revised examples may reduce some errors, particularly those involving the *Ambiguous* category. Other cases require evidence absent from the supplied input, such as images, cited sources, or broader revision context. Our experiment does not isolate the contributions of prompt design, model capability, and evidence availability.

The few-shot examples were informed by patterns observed in the evaluation data. Although topics, entities, and wording were changed to avoid verbatim duplication, this does not eliminate dependence between prompt development and evaluation. The reported scores therefore describe performance on this study sample rather than generalization to an untouched test set. Future evaluation should freeze the prompt and demonstrations before testing on independently sampled edits.

Finally, annotators had access to preliminary LLM predictions and explanations during deliberation, which may have anchored their judgments. The discussion-derived labels therefore do not constitute a reference constructed independently of model assistance. Higher agreement with these labels cannot establish that the evaluated model independently anticipated better expert judgments. Without a blinded deliberation condition, we cannot quantify the contribution of model exposure to the observed agreement.

Ethics Statement

This work analyzes public Wikipedia revision histories. We name no individual editors and report

no editor identities, although usernames are public. The analysis concerns patterns of manipulation, not the attribution of intent to individuals. The subject matter—contested heritage in ongoing and recent armed conflicts—requires care. We distinguish documented destruction and documented manipulation from interpretation throughout, and we treat the classification of an edit as manipulative as an expert judgment subject to open disagreement, not an automated verdict. The intended use of this pipeline is the protection of heritage-related knowledge integrity through transparent, reviewable flagging rather than automated enforcement. While designed for heritage-protection monitoring, automated flagging tools could in principle be misused for content suppression.

Acknowledgments

This research was supported by the 2026 CROSS Programme of EPFL and the University of Lausanne through the project “Narrative Wars: Detecting the Weaponisation of Cultural Heritage in the Digital Sphere”.

References

- Pablo Aragón and Diego Sáez-Trumper. 2021. [A preliminary approach to knowledge integrity risk assessment in wikipedia projects](#). arXiv preprint arXiv:2106.15940. Accepted at MIS2’21 (KDD workshop).
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Timothy Clack and Mark Dunkley, editors. 2023. *Cultural heritage in modern conflict: past, propaganda, parade*. Routledge Advances in Defence Studies. Routledge Taylor & Francis Group, London New York.
- Jacopo D’Ignazi, Andreas Kaltenbrunner, Yelena Mejova, Michele Tizzani, Kyriaki Kalimeri, Mariano G. Beiró, and Pablo Aragón. 2026. [Language-agnostic modeling of source reliability on wikipedia](#). *ACM Transactions on the Web*, 20(1):2:1–2:29.
- Maxime Garambois, Hamest Tamrazyan, and Emanuela Boros. 2026. A dataset of cultural heritage manipulation on english wikipedia in the russo-ukrainian context. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3183–3189.
- R. Stuart Geiger and David Ribes. 2010. [The work of sustaining order in wikipedia](#). In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM Conferences, pages 117–126.
- Sebastian Graf. 2025. *War, memorialisation, and digitality: mnemonic practices of Ukrainian virtual museums in the Russian war against Ukraine*. Department of History, Lund University, Lund.
- Feng Gu, Zongxia Li, Carlos R. Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2026. [Large language models are effective human annotation assistants, but not good independent annotators](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 71–89, San Diego, California, United States. Association for Computational Linguistics.
- Camil Hamdane, Hanna Perekhoda, Emanuela Boros, Hamest Tamrazyan, and Stefanie Prezioso. 2026. [Cultural weaponization on Wikipedia: A 400-edit multi-annotator annotation dataset for Ukraine and Armenia](#). Dataset, Zenodo. <https://doi.org/10.5281/zenodo.22898015>. Released at the Pluralistic AI & NLP: Diversity-aware, Sociotechnical, Responsible Alignment 2026 Workshop (PANDORA) @ EMNLP 2026, Budapest, Hungary.
- Nikesh Joshi, Francesca Spezzano, Mayson Green, and Elijah Hill. 2020. [Detecting undisclosed paid editing in wikipedia](#). In *Proceedings of The Web Conference 2020, WWW ’20*, pages 2899–2905, New York, NY, USA. Association for Computing Machinery.
- Zarine Kharazian, Kate Starbird, and Benjamin Mako Hill. 2024. [Governance capture in a self-governing community: A qualitative comparison of the croatian, serbian, bosnian, and serbo-croatian wikipedias](#). *Proceedings of the ACM on Human-Computer Interaction*, 8:1–26.
- Grigori Khislavski. 2022. [Weaponizing history: Russia’s war in ukraine and the role of historical narratives](#). *Journal of Applied History*, 4(1):102–125.
- Srijan Kumar, Robert West, and Jure Leskovec. 2016. [Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW ’16*, pages 591–602, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Viktoriia Makovska. 2025. [Vandalism or knowledge manipulation? Detecting narratives in Wikipedia edits](#). Master’s thesis, Ukrainian Catholic University, Lviv, Ukraine.
- Zachary J. McDowell and Matthew A. Vetter. 2020. [It takes a village to combat a fake news army: Wikipedia’s community and policies for information literacy](#). *Social Media + Society*, 6(3):2056305120937309.
- Niina Meriläinen. 2024. [The possible role of digital platforms in information operations](#). *European Conference on Social Media*, 11(1):137–143.

- Niina Meriläinen. 2026. [Weaponizing heritage in online cognitive warfare](#). *Journal of Information Warfare*, 25(2):65.
- Lynn Meskell. 2026. [Ruin warfare: How and why we weaponize the past](#). *Annual Review of Anthropology*, 55:29–43.
- Nour A. Munawar and James Symonds. 2023. [Empires of lies? the political uses of cultural heritage in war](#). *The Historic Environment: Policy & Practice*, 14(3):308–325. [_eprint: https://doi.org/10.1080/17567505.2023.2205193](https://doi.org/10.1080/17567505.2023.2205193).
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Barbara Plank. 2022. [The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Diego Saez-Trumper. 2019. [Online disinformation and the role of wikipedia](#).
- Pnina Shachaf and Noriko Hara. 2010. [Beyond vandalism: Wikipedia trolls](#). *Journal of Information Science*, 36(3):357–370.
- Filipo Sharevski, Peter Jachim, and Emma Pieroni. 2020. [WikipediaBot: Machine learning assisted adversarial manipulation of wikipedia articles](#). In *Proceedings of the 2020 Workshop on DYNAMIC and Novel Advances in Machine Learning and Intelligent Cyber Security*, pages 1–8, Lexington MA USA. ACM.
- Adam T. Smith, Lori Khatchadourian, and Ian Lindsay. 2023. [Monitoring heritage at risk: Caucasus heritage watch and the armenian monuments of nagorno-karabakh](#). In Yervand H. Grekyan and Arsen A. Bobokhyan, editors, *Systemizing the Past*, Papers in Near Eastern and Caucasian Archaeology Dedicated to Pavel S. Avetisyan on the Occasion of His 65th Birthday, pages 428–439. Archaeopress.
- Hamest Tamrazyan, Maxime Garambois, and Emanuela Boros. 2026. Linguistic mechanisms of cultural heritage weaponization in digital encyclopedic discourse: Evidence from Wikipedia revision corpora. *Scientific Journal of Dragomanov Ukrainian State University. Series 9. Current Trends in Language Development*, 31.
- Jesús Tramullas, Piedad Garrido-Picazo, and Ana I. Sánchez-Casabón. 2016. [Research on wikipedia vandalism: a brief literature review](#). In *Proceedings of the 4th Spanish Conference on Information Retrieval, CERI ’16*, pages 1–4, New York, NY, USA. Association for Computing Machinery.
- Bekir Tolga Tutuncuoglu. 2024. [Algorithmic memory: How AI is rewriting what societies choose to remember or forget](#).
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. [Do large language models know what they don’t know?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

A Seed Articles

The following seed articles were used for article curation (§3).

Ukraine: *Intangible Cultural Heritage of Ukraine; Culture of Ukraine; Holodomor; History of Christianity in Ukraine; Kyiv Pechersk Lavra.*

Armenia: *Nakhchivan; Artsakh; Ani; Armenian genocide; Republic of Artsakh; Culture of Armenia; Western Armenia; History of Armenia.*

B Per-Annotator Labels

Figure 3 and Table 3 report the initial independent round per annotator, reconstructed from the consensus-tracking exports for the contested items and from the final team labels for the unanimously agreed items; pairwise agreement coefficients are reported in Table 2. One caveat applies to all of these statistics: the meaning of the third label shifted during annotation. *Ambiguous* began as a fallback for annotators who could not decide, and only stabilised as a substantive category for edits that cannot be judged from the available context through the task itself (§4.1; §5.1). Agreement coefficients are therefore impacted by the team’s gradual convergence on the category’s meaning on top of genuine interpretive disagreement.

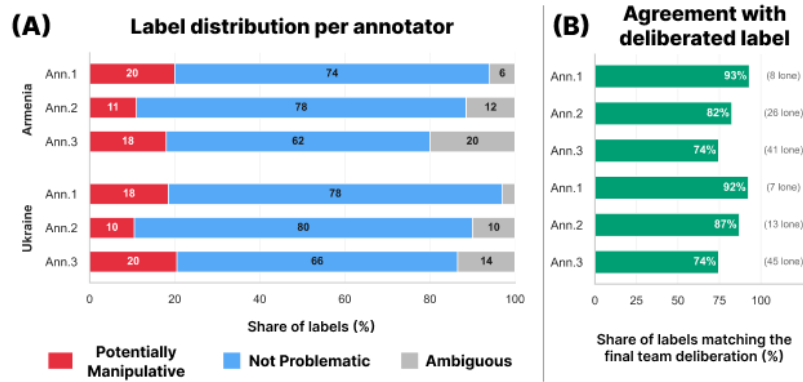


Figure 3: Initial-round labels per annotator.

Annotator 1 is the cross-country control, Annotator 2 the Ukraine specialist, and Annotator 3 the Armenia specialist **A:** share of *Potentially Manipulative* (P), *Not Problematic* (N), and *Ambiguous* (A) labels, per country ($n = 200$ each). **B:** share of each annotator’s initial labels that match the final team decision, with lone-dissent counts.

Annotator	P	N	A	Matches	Lone
<i>Armenia</i> ($n = 200$)					
Ann. 1	20.0	74.0	6.0	93.0	8
Ann. 2	11.0	77.5	11.5	82.5	26
Ann. 3	18.0	62.0	20.0	74.5	41
<i>Ukraine</i> ($n = 200$)					
Ann. 1	18.5	78.5	3.0	92.5	7
Ann. 2	10.5	79.5	10.0	87.0	13
Ann. 3	20.5	66.0	13.5	74.5	45

Table 3: Initial-round label shares (%) per annotator (P = *Potentially Manipulative*, N = *Not Problematic*, A = *Ambiguous*), share of labels matching the final team decision, and lone-dissent counts. Annotator 1 is the cross-country control, Annotator 2 the Ukraine specialist, and Annotator 3 the Armenia specialist.

C Classification Prompt

The following system prompt was used for all LLM classification runs reported in this paper.⁹

⁹The prompt text below has been formatted for readability; the model received the same instructions without the section headings.

You are an expert Wikipedia editor analyzing revision diffs for cultural manipulation.

Task

Classify each Wikipedia diff as **POTENTIALLY MANIPULATIVE**, **NOT PROBLEMATIC**, or **AMBIGUOUS**.

Decision Rule

Ask: *Does this change alter the meaning, framing, interpretation, contextual understanding, or readability of the text in a way that is derogatory, manipulative, or ideologically significant?*

POTENTIALLY MANIPULATIVE — The change shifts how readers understand, perceive, or interpret the subject. Includes:

- Swapping terms for ideologically loaded alternatives (Kiev→Kyiv, “annexation”→“reunification”, “genocide”→“alleged genocide”)
- Adding/removing text that introduces bias, one-sided emphasis, or emotional charge
- Removing inconvenient facts or citations (not bulk cleanup)
- Adding/removing quality/neutrality tags (`{{POV}}`, `{{disputed}}`) only if used to delegitimize content
- Commenting out content with `<!-- -->` to hide it
- Pure vandalism, content disruption or edits that degrade article quality and readability
- Toponym changes that systematically impose one naming convention over another, regardless of direction

NOT PROBLEMATIC — The change does not meaningfully affect interpretation or framing. Includes:

- Grammar, spelling, punctuation, or formatting fixes
- Neutral factual additions or corrections that don’t shift the broader narrative
- Technical or maintenance edits
- Talk page edits (editor discussion, not article content)
- Removing genuinely offensive or biased language (improves neutrality)

AMBIGUOUS — The diff involves changes not accessible through text alone, e.g. image changes or ambiguous external links. **Always** categorize visual-only edits as **AMBIGUOUS**.

Key Principle

Judge the **effect** of the change on meaning and framing, not the editor’s inferred intent.

Judge the actual correctness of the change; something that is ideologically loaded but factually correct is **NOT** manipulative. Use your best judgment.

Output Format

Respond with exactly:

Label: `<Potentially Manipulative | Not Problematic | Ambiguous>`. [Two sentences explaining your choice]

D Inference Configuration

All classifications used Qwen/Qwen3.6-35B-A3B self-hosted, with the following parameters: temperature = 0.2, random seed = 42, max_tokens at the API default. Each request classified a single edit, using the same instructions and 18 few-shot examples (as described in Section 3): five *Potentially Manipulative*, ten *Not Problematic*, and three *Ambiguous*.

E Few-Shot Examples

The prompt included 18 synthesized few-shot examples: five *Potentially Manipulative*, ten *Not Problematic*, and three *Ambiguous*. These examples were derived from patterns observed in the evaluation data and were designed to illustrate common errors, boundary cases, and annotation caveats. We changed the topics, entities, and wording so that no evaluation item was reproduced verbatim, thereby reducing direct overlap between the demonstrations and the evaluation set. Figures 4–6 show four of the five *Potentially Manipulative* examples, four of the ten *Not Problematic* examples, and all three *Ambiguous* examples, with the LLM’s explanations below each image. Full few-shot examples and reasonings are provided in the released dataset (Hamdane et al., 2026).



Figure 4: **Potentially Manipulative few-shot examples.** Four of five examples, in prompt order, with the LLM’s classification reasoning in italics below each image.

Treaty of Waitangi

↳ Fixed a typo in the History section

§ History

- However, pressure from missionaries and humanitarian groups, combined with the government's desire to regulate land purchases, led to the decision to negotiate a treaty with the Māori chiefs. Captain [[William Hobson]] was dispatched to New Zealand in 1839 with instructions to obtain sovereignty over part or all of the country.

+ However, pressure from missionaries and humanitarian groups, combined with the government's desire to regulate land purchases, led to the decision to negotiate a treaty with the Māori chiefs. Captain [[William Hobson]] was dispatched to New Zealand in 1839 with instructions to obtain sovereignty over part or all of the country.

The edit introduces a minor spelling error (“goverment” instead of “government”) without altering the meaning, framing, or factual content of the passage. Such typographical changes do not carry ideological or manipulative significance.

Panama Canal

↳ added section on canal operations

+ ==Lock operations==

+ The canal's locks use a gravity-fed system, with water flowing from [[Gatun Lake]] into the lock chambers through culverts. Each lock chamber is 33.53 meters wide and 320 meters long, with a usable depth of 12.6 meters. A typical transit requires approximately 200 million liters of fresh water.

The edit introduces a new section detailing the Panama Canal's lock system with neutral, factual information about its dimensions and operational requirements. This addition expands the article's scope without altering the existing narrative, introducing bias, or shifting the interpretive framing.

Egypt

↳ geography info should be in a geographic context

- Egypt is one of the most populous [[Arab world|Arab countries]]. The country's official language is [[Arabic]], which is spoken by the vast majority of the population.<ref name=cia-factbook>{{cite web|url=https://www.cia.gov/the-world-factbook/countries/egypt/ |title=Egypt |publisher=The World Factbook |accessdate=2022-03-15}}</ref>

+ The country's official language is [[Arabic]], an [[Afroasiatic languages|Afroasiatic language]], which is spoken by the vast majority of the population.<ref name=cia-factbook>{{cite web|url=https://www.cia.gov/the-world-factbook/countries/egypt/ |title=Egypt |publisher=The World Factbook |accessdate=2022-03-15}}</ref>

The edit removes a contested political classification of Egypt as an “Arab state” in the lede and replaces it with a straightforward geographic descriptor. This change improves the article's neutrality and factual precision without introducing bias or altering the core narrative.

Demographics of Brazil

↳ Standardise citation format ([[Wikipedia:Bots/Requests for approval/C...

- {{reflist|colwidth=30em

+ {{reflist|colwidth=30em}}

This edit is a routine technical maintenance change that standardises citation formatting by replacing a deprecated template parameter. It does not alter the article's content, meaning, or framing in any way.

Figure 5: Not Problematic few-shot examples.

Kashmir

↳ Updated map order to reflect current situation

```
- | image_name          = [[File:Kashmir map UN.png|240px]]<br>Map of the Kashmir region <br /><br />
{{Switcher|[[File:India Jammu and Kashmir map.svg|220px]]|Map showing Indian-administered Kashmir|[[File:Pakistan Azad
Kashmir map.svg|220px]]|Map showing Pakistan-administered Kashmir}}
+ | image_name          = [[File:Kashmir map UN.png|240px]]<br>Map of the Kashmir region <br /><br />
{{Switcher|[[File:Pakistan Azad Kashmir map.svg|220px]]|Map showing Pakistan-administered Kashmir|[[File:India Jammu
and Kashmir map.svg|220px]]|Map showing Indian-administered Kashmir}}
```

The edit swaps the order of two maps showing different territorial depictions of Jammu and Kashmir. Without being able to see and compare the maps, and without knowing which depiction each map shows, we cannot conclude whether this reordering affects the article's neutrality or framing. It is therefore categorized as ambiguous.

Democratic Republic of the Congo

↳ rv

```
+ [[Image:Stanley statue Kinshasa.jpg|thumb|upright|Statue of [[Henry Morton Stanley]], the explorer who facilitated
the establishment of the Congo Free State, in [[Kinshasa]]]]
```

The edit inserts an image of a colonial administrator with a statue. Since we cannot see the image, and since the caption's framing is ambiguous—the statue could be presented neutrally or as glorification—we cannot conclude without seeing the image itself whether this affects the article's framing.

Partition of India

```
- == External links==
- * {{Cite video |title= The Great Partition: The Making of India and Pakistan |people=Yasmin Khan (historian)
|date=March 15, 2010 |url=http://www.youtube.com/watch?v=ExamplePartitionLink }}
```

The edit removes an external link to a documentary on population displacement, along with the header of the now-empty section. Without being able to access and evaluate the documentary's content, we cannot assess whether the removal suppresses valid information or removes unreliable or propagandistic material. Therefore it is categorized as ambiguous.

Figure 6: **Ambiguous few-shot examples.** All three examples are shown in the order in which they appeared in the LLM prompt. The italic text below each image is the LLM's reasoning for the classification.