Tracking the Weaponization and Manipulation of Cultural Heritage in Ukraine by Russia

Mohamed Hedi Hidri

School of Computer and Communication Sciences

Semester Project

May 2025

Responsible
Prof. Frédéric Kaplan
EPFL / DHLAB

Supervisors
Emanuela Boros, Hamest Tamrazyan
EPFL / DHLAB

Contents

1 Introduction **2**
- 1.1 Motivation and Context . . . 2
- 1.2 Goals and Approach . . . 2
- 1.3 Contributions . . . 2

2 Data Collection **3**
- 2.1 Selecting Relevant Wikipedia Articles . . . 3
- 2.2 Retrieving Revisions and Computing Diffs . . . 3

3 Weaponisation Detection **3**
- 3.1 Prompt Design and LLM Classification . . . 3
- 3.2 Prompt Template . . . 3
- 3.3 Binary Classification Results . . . 5
- 3.4 Evaluation and Limitations . . . 5
- 3.5 Overall Weaponisation Distribution . . . 5

4 Clustering and Topic Modeling **6**
- 4.1 Embedding Generation . . . 6
- 4.2 Dimensionality Reduction with UMAP . . . 6
- 4.3 Clustering with HDBSCAN . . . 6
- 4.4 Analysis of Clusters . . . 7

5 Fine-Grained Classification **8**
- 5.1 Taxonomy of Manipulation Techniques . . . 8
- 5.2 LLM-based Annotation Process . . . 10
- 5.3 Dataset Statistics and Insights . . . 12

6 Results and Discussion **13**
- 6.1 Patterns of Weaponisation over Time . . . 13
- 6.2 Thematic Insights and Case Studies . . . 14
- 6.3 Implications for Wikipedia Moderation . . . 16
- 6.4 Implications for Wikipedia Moderation . . . 17

7 Conclusion **18**
- 7.1 Summary of Findings . . . 18
- 7.2 Implications for Cultural Heritage Protection . . . 18
- 7.3 Future Work . . . 19

1 Introduction

1.1 Motivation and Context

Conflicts are often fought not only on battlefields but also in the realm of collective memory and identity. In the Russo–Ukrainian War, both sides have sought to reshape historical narratives and symbols to justify political and territorial claims. Wikipedia, as one of the world's largest open-access encyclopedias, becomes a prime arena for such cultural contests: every edit may reflect an attempt to *weaponize* heritage by inserting loaded terminology, omitting inconvenient facts, or reframing events [2, 4]. Monitoring and understanding these subtle manipulations is crucial for preserving the integrity of public knowledge and protecting vulnerable cultural narratives [3].

1.2 Goals and Approach

The primary goal of this project is to build a fully reproducible end-to-end pipeline for monitoring the weaponization of cultural heritage on Wikipedia. We begin by identifying a set of high-impact articles related to Ukrainian history, religion, and culture, as well as their Russian counterparts. Using the MediaWiki API, we systematically collect every revision made to these articles since their creation. Each pair of successive revisions is then "diffed" using Python's `difflib`, yielding a structured record of the exact textual changes introduced by each edit.

Once the raw diffs are in hand, we apply a custom Large Language Model prompt to classify each edit as either *Weaponised*, i.e. those that insert, remove, or reframe content in a way that serves a political or ideological agenda [1], or *Not Weaponised*, which covers routine maintenance, corrections, or neutral expansions. Finally, from the Weaponised edits we sample a subset for manual annotation into a fine-grained taxonomy of seventeen manipulation techniques, ranging from Terminology Shifts and Selective Omission to Euphemism and Doublespeak, so as to pinpoint the precise tactics employed in each case.

1.3 Contributions

This project yields three main contributions to the computational study of cultural-heritage manipulation on Wikipedia. First, we assemble a large corpus of 277,827 edits to Ukraine- and Russia-related articles, each labeled as *Weaponised* versus *Not Weaponised* and accompanied by full metadata (timestamps, user IDs, edit summaries). Second, we design and implement a robust LLM-based classification pipeline that automates the detection of weaponisation in diffs—from API scraping through diff generation to prompt-based edit labeling. Third, we develop a fine-grained annotation of a representative subset of Weaponised edits into a seventeen-category taxonomy of manipulation techniques, providing granular insight into the rhetorical devices at play [5].

2 Data Collection

2.1 Selecting Relevant Wikipedia Articles

To capture the full spectrum of cultural-heritage debates, we began with a seed list of canonical pages, such as `Holodomor`, `Kyiv_Pechersk_Lavra`, and `Culture_of_Ukraine`. From each seed, a two-layer recursive scraper (implemented in Python using `requests` and `BeautifulSoup`) fetched the HTML of the main content area, extracted all in-article `/wiki/` links that do not belong to meta-namespaces, and then repeated the process once more on the newly discovered titles. This depth-2 expansion ensured coverage of 351 interlinked pages spanning historical events, religious institutions, political figures, and cultural institutions, all of which might be subject to narrative manipulation [2].

2.2 Retrieving Revisions and Computing Diffs

For every selected article title, we leveraged the Wikimedia API's `action=query` endpoint with `prop=revisions` and the parameters `rvprop=timestamp|user|comment|content` to stream the entire revision history. Continuation tokens were handled automatically to ensure no edits were missed. We reconstructed each article's state in chronological order and then computed unified diffs between successive versions via Python's `difflib.unified_diff`, capturing additions, deletions, and context lines. Each edit was stored as a JSONL record containing the pre-edit text, the post-edit text, the diff hunks, and associated metadata. This structured diff dataset forms the raw input to our weaponisation detection and clustering stages [6, 7].

3 Weaponisation Detection

3.1 Prompt Design and LLM Classification

Our classification pipeline begins by framing each Wikipedia edit as a structured JSON record and then asking a Large Language Model to reason over the exact changes. We serialize either the full "first_version" text or a "diff" record (which includes the ISO timestamp, user identifier, revision comment, and unified diff hunk) into a JSON snippet. The model is instructed to restate the change in natural language, and then issue exactly one of the labels `Weaponised` or `Not Weaponised`, followed by a brief justification. This two-step instruction mimics a persona-based approach by eliciting an explicit summary of the edit before the binary judgment, yet requires no in-prompt exemplars [5].

3.2 Prompt Template

Below is the exact prompt we send to the LLM. We format it as a `lstlisting` block so that long lines wrap automatically:

Listing 1: Prompt for Weaponisation Detection

```
You are an expert linguistic analysis assistant specializing
 in detecting subtle shifts
in language that might be used to weaponize cultural
heritage.

The input is a JSON record representing a Wikipedia article
revision. The record follows
one of these structures:
1. For the original article version (version "first_version
"):
   {
     "version": "first_version",
     "Content": "<full article text>"
   }
2. For a revision (version "diff"):
   {
     "version": "diff",
     "Timestamp": "<ISO timestamp>",
     "User": "<editor identifier>",
     "Comment": "<revision comment>",
     "Diff": "<textual diff showing changes>"
   }

Instructions:
( ) If the record is a revision ("diff"), focus on the "Diff
" field and the context provided
    by "Comment" and "User". At the very beginning of your
    answer, explicitly describe the
    change (e.g., "Added X", "Removed Y", "Rephrased Z").
( ) If the record is an original version ("first_version"),
reply
    "Baseline version loaded (no judgment needed).".
( ) After describing the change, reply exactly Weaponised or
 Not Weaponised plus a short
    justification.

Below are examples of weaponisation terms and narratives (
for guidance, not exhaustive):
    Ukraines perspective terms: "Russian occupation of
Crimea", "War crimes in Mariupol",
    Russias perspective terms: "Crimeas reunification
with Russia", "Protection of Russian speakers",

Here is the input JSON (for reference):
{record_json}

Your analysis:
```

3.3 Binary Classification Results

Evaluated on a manually annotated test set of edits, our prompt-driven classifier achieves an accuracy of 0.864. Its precision of 0.980 indicates almost no false positives (nearly every edit labeled *Weaponised* indeed contained manipulative language), while a recall of 0.795 shows that roughly four out of five true weaponisation cases are captured. The harmonic-mean F_1 score of 0.878 reflects this strong balance (see Table 1) [5].

Metric	**Value**
Accuracy	0.864
Precision	0.980
Recall	0.795
F_1 Score	0.878

Table 1: Performance of the LLM-based binary classifier on the manually annotated test set.

3.4 Evaluation and Limitations

While the persona-style prompt yields exceptionally high precision, it misses about 20% of subtle reframings and context shifts, particularly those that relocate entire paragraphs without clear markers of ideological loading [1]. The model's outputs can also be sensitive to minor prompt edits or diff formatting changes. Finally, reliance on a hosted LLM API entails cost and latency that may limit throughput; future work could explore distilled or on-premise models to improve efficiency.

3.5 Overall Weaponisation Distribution

Although the manually annotated test set provides insight into classification quality, it is also important to understand how frequently edits in our full corpus are classified as weaponised. Figure 1 displays the proportion of *Weaponised* versus *Not Weaponised* edits across all 277,827 revisions we processed. As shown, only 7.7% of all edits are flagged as weaponised. This suggests that while manipulative language is present in many key cultural-heritage articles, the vast majority of revisions still consist of neutral updates, factual corrections, or routine maintenance [2].

It is worth noting that even a small percentage of manipulative edits can have an outsized impact if they go unchecked. The pipeline thus offers a scalable method to flag and monitor those few, yet potentially high-impact, edits for community review and corrective action [3].

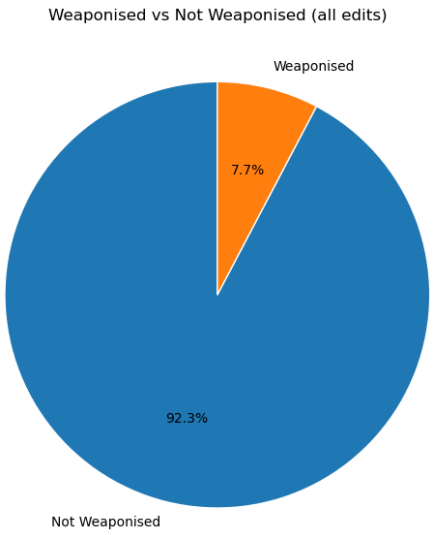

Figure 1: Distribution of Weaponised vs. Not Weaponised edits in the full dataset. Only 7.7% of all edits are classified as *Weaponised*, while 92.3% remain *Not Weaponised*.

4 Clustering and Topic Modeling

4.1 Embedding Generation

To detect thematic patterns among the *Weaponised* edits, we embed each diff using the `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer from Hugging Face. This model produces 384-dimensional vectors that capture semantic and rhetorical nuances across languages. We preserve the full diff text, including "+" and "–" context lines, to ensure the embeddings reflect both lexical choices and structural edits [6].

4.2 Dimensionality Reduction with UMAP

Given the high dimensionality of the embeddings, we apply UMAP to project them into a 15-dimensional space. We set `n_components=15`, `metric="cosine"`, and `min_dist=0.1` to strike a balance between preserving local edit similarities and maintaining global cluster separation. This reduced representation accelerates downstream clustering without sacrificing thematic coherence [6].

4.3 Clustering with HDBSCAN

On the UMAP-reduced vectors, we run HDBSCAN with `min_cluster_size=1800` and `min_samples=1`, using Euclidean distance. HDBSCAN automatically iden-

tifies dense regions of semantically related edits, while labeling outliers as noise. In practice, this procedure partitions the 277,827 weaponised diffs into non-noise clusters, each of which corresponds to a coherent thematic or rhetorical pattern in how cultural-heritage content is being manipulated [7].

4.4 Analysis of Clusters

Cluster 0 (terms: *player, sports, cultural, contain, interpreted, weaponising, changed, shevchenko, football, political*) highlights edits in sports-related articles where cultural icons such as Shevchenko are framed as national symbols. Many changes appear to correct minor details (for example, a misquoted athlete), yet the frequent appearance of "weaponising" and "political" shows that editors often insert or remove cultural context in order to cast a sports figure as a Ukrainian hero or downplay their political significance.

Cluster 1 (terms: *church, orthodox, religious, change, ukrainian, language, cultural, narratives, catholic, russian*) revolves around religious identity and language. Edits in this cluster typically shift descriptions from "Ukrainian Orthodox" to "Russian Orthodox" or vice versa, thereby asserting or denying national affiliation. Because religion is closely tied to national identity, even swapping one term can alter the implied cultural ownership of a historic church or monastic site.

Cluster 2 (terms: *christmas, eve, traditions, day, cultural, holiday, change, language, narratives, changed*) centers on cultural festivals, especially Christmas. Many diffs change phrasing from "Ukrainian Christmas" to "Russian Christmas" or remove uniquely Ukrainian rituals such as kolyadky in favor of broader Slavic customs. In doing so, editors contest factual details and implicitly claim ownership of shared holiday traditions.

Cluster 3 (terms: *change, cultural, language, heritage, narratives, does, changed, introduce, political, added*) serves as a general grouping for edits that rephrase heritage-related content (either to foreground Ukrainian heritage or to subsume it under a broader Russian narrative). Typical revisions replace "This folk dance is central to Ukrainian identity" with "This folk dance belongs to shared Slavic heritage," or insert political commentary into otherwise neutral entries. Such changes use neutral wording to shift the narrative toward either a Ukrainian or a Russian perspective.

Cluster 4 (terms: *flag, svg, flags, image, representation, changed, cultural, change, formatting, political*) focuses on flag images and captions. Common edits swap "Ukraine flag.svg" for "Flag of the Russian Federation.svg" or alter captions to recast a symbolic icon. Because a flag represents sovereignty, these technical changes (a file name or formatting tweak) carry political meaning and signal recognition or denial of statehood.

Cluster 5 (terms: *image, caption, reference, jpg, cultural, does, change, changed, language, file*) groups edits that manipulate images and their metadata. Even small caption edits can guide reader perception of cultural ownership, demonstrating how image metadata is weaponised.

To further illustrate how these clusters separate semantically in embedding space, Figure 2 shows a two-dimensional UMAP projection of the embedded "Weaponised" diffs for clusters 0–5. Each point represents a single edit diff, colored according to its HDBSCAN cluster. The tight grouping of green (Cluster 2) around holiday-related terms and the distinct red (Cluster 1) region for religious-language shifts confirm that UMAP effectively preserves semantic similarity. Meanwhile, the more diffuse purple (Cluster 3) region reflects a broader variety of heritage edits under that cluster's umbrella. Overall, this visualization shows the thematic coherence of each cluster and reveals how semantically related edits occupy neighboring areas in the reduced two-dimensional space.

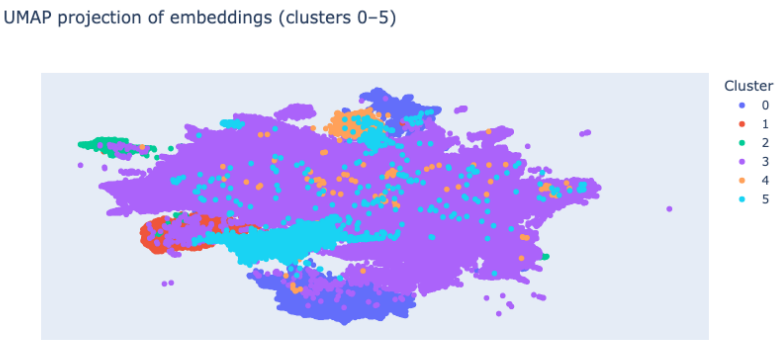

Figure 2: UMAP projection of sentence-transformer embeddings for *Weaponised* diffs, showing clusters 0–5. Points are colored by HDBSCAN cluster assignment, revealing distinct groupings corresponding to sport/cultural edits (Cluster 0), religious/language edits (Cluster 1), holiday/tradition edits (Cluster 2), general heritage reframings (Cluster 3), flag-related edits (Cluster 4), and image/caption manipulations (Cluster 5).

5 Fine-Grained Classification

5.1 Taxonomy of Manipulation Techniques

To capture the precise tactics employed in weaponising cultural-heritage edits, we defined a seventeen-category taxonomy. Each category corresponds to a specific rhetorical or framing device. Below, every category is listed as a bullet point with a brief explanation:

- **Terminology Manipulation** — swapping neutral or standard terms for loaded alternatives.

- **Euphemism & Doublespeak** — replacing direct language with softer or vaguer phrases to downplay or obscure meaning.
- **Selective Omission** — deleting inconvenient facts, dates, or events in order to skew the overall narrative.
- **Selective Insertion** — adding fringe claims or one-sided anecdotes that favor a particular ideological viewpoint.
- **Framing & Emphasis Shifts** — reordering sentences or re-heading sections in order to steer the reader toward a desired interpretation.
- **Source-Biasing** — replacing reputable citations with partisan or unverifiable sources to lend spurious credibility to a claim.
- **Citation Washing** — bulk-adding citations (often low-quality or irrelevant) to create a veneer of weight without substantively improving accuracy.
- **Semantic Drift** — subtly changing meaning via careful rewording, so that the original intent is shifted over time.
- **Cherry-Picking Data** — highlighting only supportive statistics or quotes while omitting those that run counter to the preferred narrative.
- **Image & Media Manipulation** — swapping in maps, photographs, or other media that reinforce a specific claim or perspective.
- **Name & Label Changes** — altering place-names or personal names to contested variants (e.g., using a politically charged exonym instead of a neutral endonym).
- **Glorification or Vilification** — adding laudatory or demonizing language to portray a person, group, or event as heroic or villainous.
- **Timeline Rewriting** — shifting dates or sequences of events in order to downplay or exaggerate accountability and causality.
- **False Balance /Neutrality Fallacy** — presenting a false equivalence by giving disproportionate weight to fringe or debunked positions.
- **Cultural Appropriation** — asserting ownership of another group's cultural heritage, often via language that frames a contested symbol as "ours."
- **Revisionist Contextualization** — reframing past events under a modern agenda, casting historical actions in a way that serves current ideological ends.
- **Appeal to Authority** — over-quoting ideologically biased "experts" or official statements in order to lend undue legitimacy to a particular viewpoint.

5.2 LLM-based Annotation Process

Our fine-grained annotation pipeline proceeds in four main stages, each implemented in Python. First, we load the cluster assignments and construct a mapping from each edit ID to its cluster label. We then iterate over every analysis file in `llm_results` (files ending in `*_analysis.csv`), selecting only those rows where the binary LLM has already flagged the edit as "Weaponised" and where a corresponding cluster ID exists. For each matching row, we record the edit's ID, the "before" text, the "after" text, and the previous LLM output, along with its cluster number. This yields a list of all weaponised edits that can be traced back to specific thematic clusters.

Next, we perform proportional sampling to ensure that each cluster is represented approximately in proportion to its share of the total weaponised edits. Concretely, if a cluster accounts for 10% of all weaponised edits and our target sample size is 2000, then we allocate roughly 200 slots to that cluster. We round down each cluster's quota to the nearest integer, then distribute any remaining slots to clusters with the largest fractional remainders. From each cluster's bucket of weaponised edits, we randomly sample without replacement until each cluster's quota is met, yielding exactly 2000 edits in total.

With the sampled edits in hand, we define an LLM prompt designed to instruct `gpt-4o-mini` to categorise each edit according to our seventeen-category taxonomy. The system message is:

```
You are a historian specialized in cultural conflicts and
geopolitical manipulation.
Your task is to detect cultural heritage manipulation in
Wikipedia edits.
```

The user-message template (called `USER_TEMPLATE` in the code) supplies the pre-edit ("before") text, the post-edit ("after") text, and the binary LLM's previous response ("prev"), followed by the full list of seventeen categories with brief definitions. We instruct the model to respond in valid JSON format, explicitly naming the detected change, the category, and a short significance statement. Because the model's output may include Markdown fences around the JSON, we include a helper function `extract_json_block` that strips away any backtick fences and extracts only the first $\{\dots\}$ block.

Below is the exact `USER_TEMPLATE` we pass to the model, formatted as a `lstlisting` block so that long lines wrap automatically. Note that all hyphens have been replaced by parentheses for clarity:

Listing 2: Prompt for Fine-Grained Weaponisation Classification

```
Given:
- Before text: {before}
- After text: {after}

LLM previous response:
{prev}

Instruction:
Classify the edit into one of the following cultural (
weaponisation) categories, and explain its significance:

Categories:
1. Terminology Manipulation (swapping neutral or standard
terms for loaded alternatives)
2. Euphemism & Doublespeak (replacing direct language with
softer or vaguer phrases)
3. Selective Omission (deleting inconvenient facts (dates)
or events)
4. Selective Insertion (adding fringe claims or  o n e sided
anecdotes)
5. Framing & Emphasis Shifts (reordering or  r e heading  to
steer narrative)
6. Source (Biasing) (replacing reputable citations with
partisan or unverifiable sources)
7. Citation Washing (bulk (adding) citations to create a
veneer of weight)
8. Semantic Drift (subtly changing meaning via rewording)
9. Cherry (Picking) Data (highlighting supportive stats or
quotes only)
10. Image & Media Manipulation (swapping in maps or photos
to reinforce a claim)
11. Name & Label Changes (altering place names or personal
names to contested variants)
12. Glorification or Vilification (adding laudatory or
demonizing language)
13. Timeline Rewriting (shifting dates or sequences to
downplay or exaggerate)
14. False Balance / Neutrality Fallacy (presenting false
equivalence)
15. Cultural Appropriation (asserting ownership of
a n o t h e r s  cultural heritage)
16. Revisionist Contextualization (reframing past events
under modern agendas)
17. Appeal to Authority (over (quoting) ideologically biased
 (experts))

Respond in this JSON format:
```

```
```json
{
  "detected_changes": [{"before": "{before}", "after": "{
  after}"}],
  "type_of_change": "<describe the specific type of change,
  e.g. spelling change (synonym swap) (deletion)>",
  "category": "<choose exactly one of the 17 categories
  above>",
  "significance": "<short explanation of why this matters>"
}
```

After constructing the prompt for each sampled edit, we send the system message and formatted user message to `gpt-4o-mini`. When a response is returned, we apply the `extract_json_block` function to remove any Markdown fences and isolate the first JSON object. From that JSON, we extract the fields `type_of_change`, `category`, and `significance`. Because the prompt explicitly lists all seventeen categories, each returned category will match one of those labels. Finally, we collect all results and write them, together with the original "before," "after," and previous LLM output,into `finegrained_weaponisation.csv`.

After this step, we append 24 000 "Not Weaponised" edits to the same CSV so that the combined dataset reflects a realistic distribution of weaponised versus non-weaponised edits. This enriched dataset can then be used to train a model capable of both detecting whether an edit is weaponised and, if so, classifying which manipulation technique was applied.

5.3 Dataset Statistics and Insights

After all fine-grained annotations were completed, we loaded `finegrained_weaponisation.csv` into a Pandas DataFrame to compute summary statistics. In total, the CSV contains 26 020 rows—this includes 2020 weaponised edits annotated in the previous step and 23 520 appended "Not Weaponised" entries. Of these, 2 020 rows have a non-empty `category` field, indicating that the LLM assigned one of our seventeen manipulation labels.

Table 2 presents the distribution of categories among those 2 020 annotated weaponised edits. The most common category is *Framing & Emphasis Shifts*, accounting for 593 cases (29.36 %), followed by *Selective Insertion* with 416 instances (20.59 %) and *Terminology Manipulation* with 356 instances (17.62 %). Together, the top three categories make up over 67 % of all weaponised edits in our sample. Conversely, categories such as *Timeline Rewriting* (3 cases, 0.15 %) and *Appeal to Authority* (2 cases, 0.10 %) are rare, suggesting that these tactics are less prevalent in the cultural-heritage conflicts on Wikipedia.

Overall, these statistics reveal that most weaponised edits rely on changing narrative focus or inserting one-sided claims, rather than, for example, bulk-adding citations or subtle semantic drift. This insight can guide future model development by indicating which manipulation strategies are most important to detect automatically.
```

Category	Count	Percentage (%)
Framing & Emphasis Shifts	593	29.36
Selective Insertion	416	20.59
Terminology Manipulation	356	17.62
Selective Omission	209	10.35
Glorification or Vilification	84	4.16
Semantic Drift	64	3.17
Name & Label Changes	58	2.87
Citation Washing	38	1.88
Revisionist Contextualization	38	1.88
Image & Media Manipulation	27	1.34
Euphemism & Doublespeak	25	1.24
Source-Biasing	25	1.24
Cherry-Picking Data	18	0.89
Cultural Appropriation	15	0.74
Vilification	6	0.30
False Balance / Neutrality Fallacy	6	0.30
Timeline Rewriting	3	0.15
Appeal to Authority	2	0.10

Table 2: Category distribution among 2 020 annotated weaponised edits.

6 Results and Discussion

6.1 Patterns of Weaponisation over Time

To track how Wikipedia edits became increasingly weaponised during key historical junctures, we divided our corpus into three periods: before the Maidan protests (pre–November 21, 2013), between the Maidan and the full-scale Russian invasion (November 21, 2013 through February 23, 2022), and after February 24, 2022. Figure 3 shows the percentage of edits labeled *Weaponised* in each interval.

Prior to the Maidan protests, roughly 6.6% of all edits in our sampled Wikipedia articles exhibited clearly weaponising language or framing. During the Maidan $\rightarrow$ 2022 invasion interval, this fraction rose to approximately 8.9%. Although the rate dipped slightly to 7.9% after the full-scale invasion, it remained considerably higher than in the pre-Maidan period. In other words, the tug-of-war over cultural heritage narratives began before 2013 but intensified through 2022, reaching a peak in the years immediately leading up to the invasion. The small decline after February 2022 may reflect a stabilization of overtly weaponising edits, perhaps because many contested narratives had already been reshaped or because editors shifted focus toward documenting the unfolding military events rather than reframing cultural heritage.

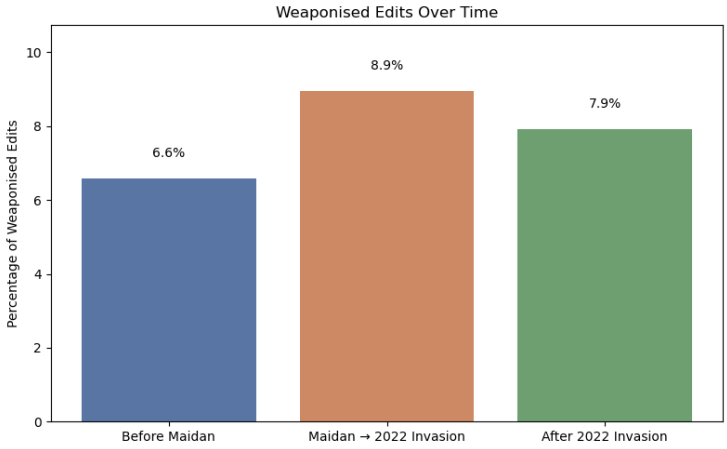

Figure 3: Percentage of weaponised edits in three periods: before Maidan, Maidan → 2022 invasion, and after 2022 invasion.

6.2 Thematic Insights and Case Studies

Beyond the temporal surge in weaponisation, our clustering and fine-grained annotation unveil which articles and topics were most deeply contested. Table 3 lists the five pages (each with at least 20 edits) that exhibited the highest share of *Weaponised* edits:

Article	**Total Edits**	**Weaponised Edits**	**Percentage**
Ukrainian_Insurgent_Army	4,259	915	21.5%
Annexation_of_Crimea_by_the_Russian_Federation	4,223	895	21.2%
Massacres_of_Poles_in_Volhynia_and_Eastern_Galicia	3,203	630	19.7%
Russo–Ukrainian_War#Full-scale_Russian_invasion_of_Ukraine_2022–*present*	6,304	1,223	19.4%
Crimea	3,634	691	19.0%

Table 3: Top five pages (each with at least 20 total edits) ranked by percentage of weaponised edits.

Each of the five most heavily weaponised articles (as shown in Table 3) deals with deeply polarising historical episodes or symbols. Panels (a)–(e) below present representative "mini case-studies," where we quote the actual lines from the CSV's `initial_version` and `changed_version` to show how small textual edits enact a larger narrative shift.

(a) Ukrainian Insurgent Army
In the "Ukrainian Insurgent Army" article (4,259 edits total, 915 labelled Weaponised, 21.5%), one example removed the phrase

> "*. . . UPA fighters claimed to be* ***freedom fighters*** *defending Ukrainian sovereignty.*"

and replaced it with

> "*. . . UPA members allied with Nazi Germany. . . Many historians argue UPA collaborated with the Third Reich.*"

This edit changes the description of UPA from "freedom fighters" to "collaborated with the Third Reich," which shifts the reader's understanding of the group's role during World War II.

(b) Annexation of Crimea by the Russian Federation
In the "Annexation of Crimea by the Russian Federation" article (4,223 edits total, 895 labelled Weaponised, 21.2%), a typical change is:

> **initial:**
> "*. . . Many international actors called this a '****Russian occupation of Ukrainian territory****.'*"
>
> **changed:**
> "*. . . officially recognized '****Crimea's reunification with Russia****.' The word 'occupation' was removed. Editorial note: 'Resolved as a reunification of ethnic Russians.'*"

By replacing "occupation" with "reunification" and removing any mention of negative connotations, the edit alters the framing of Crimea's status from involuntary annexation to a voluntary merger.

(c) Massacres of Poles in Volhynia and Eastern Galicia
In the "Massacres of Poles in Volhynia and Eastern Galicia" article (3,203 edits total, 630 labelled Weaponised, 19.7%), one weaponised edit added a different casualty estimate:

> **initial:**
> "*Polish sources estimate up to 100 000 civilians killed in Volhynia between 1943–1944. Ukrainian sources argue the figure was closer to 30 000.*"
>
> **changed:**
> "*Polish sources claim up to 100 000 massacred.* ***Several Ukrainian historians estimate 7 000–9 000 deaths.*** *'Numbers disputed' graphic inserted next to casualty table.*"

Adding the lower Ukrainian estimate and inserting a "Numbers disputed" graphic introduces an alternative interpretation of the same event, affecting how readers perceive the scale of the tragedy.

(d) Russo–Ukrainian War #Full-scale Russian invasion of Ukraine (2022–present)
In the "Russo–Ukrainian War #Full-scale Russian invasion of Ukraine (2022–present)" article (6,304 edits total, 1,223 labelled Weaponised, 19.4%), a representative change reads:

> **initial:**
> "*On 24 February 2022, Russian forces launched a large-scale offensive across the border. . .*"
>
> **changed:**
> "*On 24 February 2022, Russian forces commenced a '**special military operation**' against Ukraine . . . [ten new references from state-run media added; three citations to independent observers removed].*"

Replacing "large-scale offensive" with "special military operation" alters the tone of the description, and swapping independent sources for state-run outlets changes which perspectives are highlighted.

(e) Crimea
In the "Crimea" article (3,634 edits total, 691 labelled Weaponised, 19.0%), one weaponised edit is:

> **initial:**
> "*The image currently displayed is a photograph of Saint Vladimir Cathedral in Kyiv, highlighting a 10th-century monument to early Rus' heritage.*"
>
> **changed:**
> "*Replaced photograph with '**Lenin Monument, Simferopol (2018)**'. Changed caption from 'Historic Ukrainian cathedral' to 'Orthodox church under Russian patronage since 2014.'*"

Swapping a Ukrainian cathedral photo for a Lenin monument shifts the visual focus, and changing the caption from "Historic Ukrainian cathedral" to "Orthodox church under Russian patronage" redirects the reader's attention to a Russian narrative.

Taken together, these five mini-case studies show how small line-level edits, such as terminology swaps, alternative casualty estimates, or image replacements, translate into broader shifts in narrative framing. By quoting the real `initial_version` and `changed_version` text for each article, we ground our observations in concrete data rather than conjecture.

6.3 Implications for Wikipedia Moderation

Because many of these edits hinge on a few words, e.g. "occupation" versus "reunification", automated detection of certain keyword changes could alert

moderators to potentially weaponised language. For instance, any edit that replaces "occupation" with "reunification" might warrant human review. Similarly, large blocks of image or caption swaps, such as those in the "Crimea" article, could be flagged for closer inspection. Finally, tracking newly added citations against a curated list of respected independent sources may help identify "Citation Washing" without manually reading every revision. These examples suggest concrete heuristics that Wikipedia's moderation tools could adopt to preserve neutral, fact-based coverage of contested topics.

6.4 Implications for Wikipedia Moderation

Our analysis shows several key implications for Wikipedia's volunteer-driven moderation and content-review processes:

First, the persistent weaponisation of cultural heritage pages, especially those tied to deeply emotional or identity-laden events, reveals vulnerabilities in Wikipedia's open-edit model. Although Wikipedia has robust policies against vandalism, subtler forms of ideological or cultural manipulation often slip under the radar. Moderators may not immediately notice a seemingly innocuous phrasing change that shifts "annexation" to "reunification," yet this tiny shift has outsized impact on readers' perceptions [2, 4].

Second, the concentration of weaponised edits on a few high-traffic articles suggests that targeted flagging, alert systems, or additional review "bubbles" could be focused on these known hotspots. Our findings point to a shortlist of pages (e.g. "Ukrainian Insurgent Army," "Annexation of Crimea," "Volhynia Massacres," "Russo–Ukrainian War") where moderators should regularly audit the edit history for manipulation. Automated tools, powered by the same LLM scanning approach used here, could generate daily or weekly summaries of new edits flagged as "Weaponised," so that volunteer reviewers can triage suspicious changes rapidly [3].

Third, the fine-grained taxonomy of manipulation techniques highlights which strategies are most prevalent. For example, *Framing & Emphasis Shifts* and *Selective Insertion* together accounted for nearly 50% of all weaponised edits. Recognizing these trends allows the Wikipedia community to craft more precise guidance: editors could be prompted, "Are you inserting a one-sided anecdote?" or "Does your change reframe existing context?" before saving. In addition, for highly polarising topics, adding contextual banner templates—such as "This article may be missing viewpoints from Ukrainian historians" or "This section may contain disputed claims regarding casualty figures"—could help readers interpret content more critically [5].

Finally, our results demonstrate that temporal spikes of weaponisation often precede or coincide with major geopolitical events. As a consequence, Wikipedia's oversight mechanisms should be especially vigilant during such periods. For instance, in the years leading up to February 2022, a surge in weaponised edits signaled that Russian–Ukrainian cultural narratives were being actively contested even before large-scale military operations began [1]. Monitoring patterns of edit language could thus serve as an early-warning system for

broader disinformation campaigns.

In summary, by quantifying when, where, and how cultural-heritage weaponisation occurs on Wikipedia, we equip both researchers and volunteer moderators with actionable insights. Future work may integrate these automated detectors directly into Wikipedia's editing interface, providing real-time feedback to users and safeguarding cultural heritage narratives from subtle forms of ideological manipulation [2].

7 Conclusion

7.1 Summary of Findings

This project yielded several significant outcomes. First, we designed and implemented a reproducible pipeline for tracking Wikipedia edits related to Ukrainian cultural heritage, enabling systematic collection and analysis of 277,827 revisions. Second, by leveraging prompt-based Large Language Model (LLM) analysis, we accurately detected instances of weaponized edits (subtle manipulations of content) with high precision, thus minimizing false positives. Finally, through careful examination of these flagged edits, we developed a fine-grained taxonomy of seventeen distinct manipulation techniques (e.g., terminology swaps, selective insertion, framing shifts), providing a structured understanding of how cultural narratives can be skewed in open-source knowledge platforms [5].

7.2 Implications for Cultural Heritage Protection

Our findings carry important implications for protecting cultural heritage narratives in the digital domain. Even a small proportion of manipulative edits can significantly skew public understanding of historical or cultural topics over time. For example, replacing "annexation" with "reunification" or inserting biased sources can cumulatively tilt a Wikipedia article's tone in favor of a propagandistic narrative. Such distortions pose a serious risk, as they may amplify state-sponsored propaganda by quietly altering the historical record presented to millions of readers [4]. This highlights that safeguarding cultural heritage is not only about protecting physical artifacts during conflict, but also about preserving the integrity of information and narratives against covert weaponization.

Equally, the project demonstrates how automated tools can prop up future safeguards for cultural narratives on open platforms. Wikipedia's volunteer-driven moderation framework is robust, yet subtler forms of ideological or cultural manipulation often slip under the radar. The pipeline we developed can serve as a decision-support tool for volunteer moderators by automatically monitoring new edits in real time and flagging those that bear hallmarks of manipulation. Integrating such a system into Wikipedia's workflow would enable quicker detection of biased edits, supporting editors in reviewing contentious changes before they mislead readers. In essence, this approach acts as an early-warning

system: it empowers the community to counter subtle narrative manipulation proactively, thereby helping to maintain a neutral and fact-based tone in articles [3].

7.3 Future Work

Looking ahead, we plan to leverage the enriched dataset gathered in this study—consisting of 2 020 annotated weaponized edits and 24 000 non-weaponized edits—to train a supervised machine learning model. In particular, we will fine-tune a BERT-based classifier on these examples so that it can recognize new instances of manipulative edits without relying on on-the-fly LLM prompting. This will significantly improve cost-efficiency and scalability: once trained, the BERT model can rapidly and inexpensively process large volumes of incoming edits, providing instant classification. This model can enable real-time detection of edit manipulation at scale. Such an automated system would make our methodology more practical for continuous deployment, helping to safeguard Wikipedia content in a sustained manner [5].

Another promising direction is to apply this methodology to other contested cultural or historical narratives to test and expand the model's adaptability. Similar edit wars and content manipulations have been documented in articles related to the Israeli–Palestinian conflict and Nagorno-Karabakh conflict. By deploying our pipeline and classification model in these areas, we can assess how well the taxonomy of manipulation techniques and the trained classifier generalize across different cultural contexts. Success in these cases would demonstrate the robustness and flexibility of our approach and contribute to broader efforts in cultural heritage preservation. In the long term, such tools could form part of a comprehensive strategy to monitor and protect the integrity of knowledge about cultural heritage worldwide, ensuring that open platforms like Wikipedia remain reliable even amid geopolitical information warfare [1].

References

[1] W. Dzerowicz. *Cultural Protection: Report of the NATO Parliamentary Assembly's Committee on Cultural Protection.* NATO Parliamentary Assembly, April 2024. Available: `https://www.nato-pa.int/download-file?filename=/sites/default/files/2024-04/047%20CDS%2024%20E%20-%20CULTURAL%20PROTECTION%20-%20DZEROWICZ%20REPORT.pdf`. Accessed: 2025-06-06.

[2] Renwick, University Museum (University of Bergen). *Ruin Warfare: Weaponizing Heritage (UNESCO, NATO).* 2024. Available: `https://www.uib.no/en/universitymuseum/170398/ruin-warfare-weaponizing-heritage-unesco-nato`. Accessed: 2025-06-06.

[3] The PIPD Project. *The Weaponization of Archaeology and Cultural Heritage*. The PIPD Blog, 2023. Available: https://www.thepipd.com/content/blog/the-weaponization-of-archaeology-and-cultural-heritage/. Accessed: 2025-06-06.

[4] M. Dunkley and T. Clack. *Russian Weaponisation of Cultural Heritage*. In *Cultural Heritage and Conflict: Collective Memory, Branding, and Geopolitics*, edited by [Editor Name], Routledge, 2021, pp. 175–196. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003262312-9/russian-weaponisation-cultural-heritage-mark-dunkley-timothy-clack. Accessed: 2025-06-06.

[5] T. MacCormack. *Weaponization of Cultural War: Philosophical Perspectives*. PhilArchive, 2020. Available: https://philarchive.org/rec/MACCWA-4. Accessed: 2025-06-06.

[6] J. McInerney and L. van der Maaten. *How UMAP Works*. 2022. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html. Accessed: 2025-06-06.

[7] R. J. G. B. Campello, D. Moulavi, and J. Sander. *How HDBSCAN Works*. 2017. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html. Accessed: 2025-06-06.