The Rashomon Wikipedia: A Data-Perspectivist Analysis of Divergent Historical Narratives

Claudiu Creanga[2,3], **Liviu P. Dinu**[1,3], **Anca Dinu**[3,4]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
[4] Faculty of Foreign Languages and Literatures,
University of Bucharest, Romania
claudiu.creanga@fmi.unibuc.ro, ldinu@fmi.unibuc.ro, anca.dinu@lls.unibuc.ro

Abstract

Wikipedia aims to provide a unified, neutral record of history, yet its independent language editions often function as distinct epistemic communities, creating divergent narratives around contested events. This paper investigates cross-lingual historiographical bias by analyzing Wikipedia articles across five languages (Romanian, Hungarian, Russian, Turkish, and English) focusing on three contentious events in Romanian history: the Battle of Posada (1330), the Soviet occupation of Bessarabia (1940), and the Night Attack at Târgoviște (1462). Using human annotators and Large Language Models (LLMs) to classify citation stance and quantify narrative evolution from 2005 to 2024, we identify a phenomenon of "citation isolation". In the case of the Battle of Posada, only 2 out of 119 citations were shared between language editions, with the Romanian edition exhibiting a 91% pro-national bias compared to the balanced Hungarian edition. Longitudinal analysis reveals that these narratives are volatile and responsive to contemporary geopolitics, evidenced by a significant shift in the Russian framing of Bessarabia in 2024. Finally, we propose a "Peace-Maker" pipeline to automate conflict reconciliation. We demonstrate that while standard prompting leads models to hallucinate consensus, "adversarial" prompting, which explicitly instructs the model to preserve and attribute disagreement, achieves near-perfect neutrality scores.

> History is that certainty produced at the point where the imperfections of memory meet the inadequacies of documentation.
>
> Julian Barnes

1 Introduction

Wikipedia is the first place many people turn to for historical information. It aims to provide a "Neutral Point of View" (NPOV), representing conflicting perspectives fairly and without bias. However, Wikipedia is not a single unified record. It is a federation of independent language editions, each written by a distinct community of editors who often rely on their own national historiography. When history is contested, such as in wars or territorial disputes, these communities can produce articles that describe the same event in fundamentally different ways.

This problem goes beyond simple translation differences. A reader of the Romanian Wikipedia might learn about a heroic defense of independence, while a reader of the Turkish or Russian Wikipedia reads about a failed rebellion or a justified annexation. While the NPOV policy exists in all editions, its application varies. Local editors often cite local sources, creating "citation universes" that rarely overlap. As a result, the bias is not just in the text, but in the choice of which facts to include and which authorities to trust.

In this paper, we study how these conflicting narratives form and evolve. We focus on three events that are central to Romanian history but contested by its neighbors: the Battle of Posada (1330) against Hungary, the Soviet occupation of Bessarabia (1940) involving Russia, and the Night Attack at Târgoviște (1462) against the Ottoman Empire. We analyze articles from five language editions (Romanian, Hungarian, Russian, Turkish, and English) to answer three questions:

1. How do citation choices reflect national bias? We find that editors almost exclusively cite authors from their own country. In our analysis of the Battle of Posada, only 2 out of 119 citations were shared between the Romanian, Hungarian, and English editions.

2. How do narratives evolve over time? By analyzing article versions from 2005 to 2024, we show that these narratives are not static. While the Romanian account of the Night Attack has moderated in recent years, the Russian account of Bessara-

Proceedings of NLPerspectives @ LREC 2026, pages 11–20
12 May 2026.

bia has been volatile, shifting significantly in 2024.

3. Can AI help reconcile these conflicts? We test whether LLMs can act as neutral arbiters. We find that standard prompting fails because models try to "fix" the conflict by choosing one side. However, an "adversarial" prompting strategy, which explicitly instructs the model to preserve the disagreement, can generate summaries that are rated as perfectly neutral.

It is important to note that "neutrality" in history is a contested concept. Often, there is no single objective truth between two national myths. Therefore, we do not define neutrality as finding a "middle ground" or a "correct" version of events. Instead, we define it operationally as *perspectival balance*: a neutral summary is one that accurately represents the existence and nature of the conflict itself, attributing claims to their respective traditions without endorsing one over the other.

2 Related Work

The challenge of addressing conflicting historical narratives on Wikipedia sits at the intersection of computational social science, natural language processing, and digital history. While early work focused on metadata-driven analysis of "edit wars", recent advances in LLMs have shifted attention toward semantic analysis of bias and automated conflict resolution.

2.1 Epistemic Communities & Cross-Lingual Divergence

Wikipedia is ostensibly a global project, but empirical research suggests it functions more as a federation of distinct "epistemic communities" (Samoilenko et al., 2016). Hecht and Gergle (2009) first quantified the "self-focus bias" inherent in these communities, showing that language editions disproportionately cover topics related to their own geography. Miquel-Ribé and Laniado (2018) expanded this to 40 languages, revealing a persistent "culture gap" where shared knowledge is the exception rather than the rule.

In the domain of history, this divergence often manifests as "citation isolation". Taylor et al. (2025) found that non-English Wikipedias cite millions of unique sources not found in English Wikipedia, effectively creating parallel knowledge bases. Baigutanova and Others (2023) further showed that sources deemed unreliable in one language often persist in others, highlighting the lack of a unified standard for verifiability. Our work provides a granular case study of this phenomenon: we show that for the Battle of Posada, the "Romanian epistemic community" and the "Hungarian epistemic community" rely on entirely disjoint sets of historical authorities, with only 2 shared sources out of 119.

2.2 Bias Detection: From Syntax to Semantics

Traditional approaches to bias detection in NLP have focused on linguistic cues. Recasens et al. (2013) created the foundational "NPOV corpus", identifying subjective intensifiers (e.g., "famous", "outrageous") as markers of bias. Pryzant et al. (2020) advanced this by using BERT-based models to automatically rewrite such sentences into neutral forms.

However, historical bias is often implicit and structural rather than purely stylistic. Rogers and Sendijarevic (2012) demonstrated this in their manual analysis of the Srebrenica massacre articles, where the bias lay in **which** facts were selected rather than **how** they were phrased. Recent work by Ghanbari Haez and Dragoni (2025) confirms that modern LLMs still struggle with this "narrative bias", often reproducing intersectional identity biases even when explicitly instructed to be neutral. Our methodology addresses this by moving beyond sentence-level style transfer to document-level narrative analysis, using LLMs to quantify the "stance" of citations and the evolution of historical framing over decades.

2.3 Perspectivism and Automated Reconciliation

The emerging "Perspectivist Data" paradigm (Perspectivist Data Manifesto, 2020) argues that disagreement in data should not always be aggregated away or treated as noise. Instead, valid conflicting perspectives should be preserved. This is particularly relevant for LLMs, which tend to hallucinate a single "consensus" reality when none exists. Köksal et al. (2023) and Li et al. (2023) have shown that LLMs exhibit strong "nationality bias", often aligning with the geopolitical views of their training data's dominant language.

To mitigate this, recent frameworks like "MoDS" (Moderating a Mixture of Document Speakers) (Balepur et al., 2025) and "Multi-Perspective Fusion" (Guan et al., 2025) have proposed treating documents as debating agents. Our "Peace-Maker" pipeline extends this line of inquiry. We show that

standard "summarization" prompts fail because they encourage the model to resolve conflict (a form of **hallucinated consensus**), whereas "adversarial" prompts that enforce the preservation of conflict, aligning with the perspectivist approach, achieve significantly higher neutrality scores.

3 Methodology

Our approach combines historical data mining with LLM-based analysis to quantify bias and generate neutral narratives. The pipeline consists of three stages: (1) multi-lingual data collection, (2) granular stance classification, and (3) adversarial narrative generation.

3.1 Data Collection

We targeted three historical events chosen for their contentious nature in Eastern European historiography:

1. **Battle of Posada (1330):** A foundational conflict between Wallachia and Hungary.
2. **Soviet Occupation of Bessarabia (1940):** A territorial dispute between Romania and the USSR/Russia.
3. **Night Attack at Târgoviște (1462):** A military encounter between Vlad the Impaler (Wallachia) and Mehmed II (Ottoman Empire).

For the **citation analysis** (Posada), we extracted the full content of the Romanian (ro), Hungarian (hu), and English (en) Wikipedia articles as of February 2026. We used a standard github library (Kurtovic, 2023) to parse the MediaWiki markup and extract all citations, including those in '<ref>' tags and bibliography sections.

For the **temporal analysis** (Bessarabia and Târgoviște), we used the Wikipedia Action API (Wikimedia Foundation, 2024) to fetch historical snapshots of the articles from January 1st of 2005, 2010, 2015, 2020, and 2024. This resulted in a dataset of 29 article versions across Romanian, Russian, Turkish, and English editions.

3.2 Stance Classification & Metrics

To quantify bias, we used LLMs and human annotators.

3.2.1 Citation Stance (Posada)

We classified 119 citations into four categories: *Pro-Romanian*, *Pro-Hungarian*, *Neutral*, and *Contested*. A major challenge was that many citations in the bibliography lacked context (e.g., just "Djuvara, p. 180"). To address this, we developed a **Context Enhancer** module. For each citation, the module searched the full article text for mentions of the author or title, extracting a ± 300 character window around the mention. This "enhanced context" was then fed to Google's Gemini 3.0 Pro Preview model (Gemini Team, Google, 2025) with the following prompt structure:

> "You are an expert historian analyzing the Battle of Posada (1330) between Wallachia and Hungary. Analyze the following text snippet from a Wikipedia article to determine the STANCE of the citation. The citation being analyzed is: "$citation - text$". The surrounding context is: "$context$". Determine if the citation is used to support a specific narrative:: Pro-RO (heroic defense), Pro-HU (treacherous trap), Neutral, or Contested".

All API calls were executed with a temperature of 0 to ensure deterministic reproducibility.

3.2.2 Human validation

To validate the LLM's classifications, we used two human annotators. One annotator reviewed a random 50% sample of the Romanian and English citations, while a second annotator reviewed the complete set of Hungarian citations. Both annotators assessed the same "enhanced context" provided to the LLM using the identical 4-category scheme (Pro-RO, Pro-HU, Neutral, Contested), blinded to the LLM's ratings. The human-LLM agreement was substantial, yielding Cohen's Kappa scores of 0.80 for Romanian, 0.75 for English, and 0.85 for Hungarian.

3.2.3 Granular Narrative Metrics (Bessarabia/Târgoviște)

For the temporal analysis, we defined event-specific metrics on a 0-100 scale. For example, for the Night Attack, we measured:

- **Vlad Heroism:** Extent to which Vlad is portrayed as a brave defender.
- **Ottoman Threat:** Emphasis on the magnitude of the invading force.
- **Cruelty:** Emphasis on impalement and brutal tactics.

We processed each article version through the LLM to score these metrics, allowing us to track the evolution of the narrative over 20 years. The full prompt used for this granular scoring is provided in Appendix A.

While the prompt requested categorical stances and confidence scores (0.0-1.0), we aggregated these into 0-100 scales for visualization by mapping the confidence of a "Pro-X" classification to a positive score and "Pro-Y" to a negative or lower score, normalized to a 0-100 range where 100 represents the maximum intensity of the national narrative.

3.3 The "Peace-Maker" Pipeline

To address the challenge of representing conflicting narratives, we developed the "Peace-Maker" pipeline.

1. **Claim Extraction:** We first extract key factual claims from each source text (e.g., "Romanian source claims 15,000 Ottoman casualties").

2. **Conflict Matching:** We use an LLM to identify pairs of conflicting claims (e.g., "RO: Attack was a victory" vs. "TR: Attack was a failed assassination").

3. **Adversarial Generation:** We tested four prompting strategies to generate a summary:
   - *Standard:* "Summarize these sources".
   - *Academic:* "Write as a peer-reviewed historian".
   - *Mediator:* "Write a diplomatic report acceptable to both sides".
   - *Adversarial:* "Preserve the conflict. Explicitly state 'Source A says X while Source B says Y'. Do not resolve the dispute".

4. **LLM-as-a-Judge:** We evaluated the generated summaries using a separate LLM instance to score them on *Neutrality* (0-1), *Conflict Preservation* (0-1), and *Source Attribution* (0-1).

4 Experiments & Results

We present results from three experiments: (1) citation stance analysis of the Battle of Posada, (2) temporal analysis of narratives surrounding Bessarabia and Târgoviște, and (3) the evaluation of the Peace-Maker pipeline.

4.1 Experiment 1: Citation Isolation (Posada)

We analyzed 119 citations across the Romanian (RO), Hungarian (HU), and English (EN) Wikipedia articles on the Battle of Posada (1330).

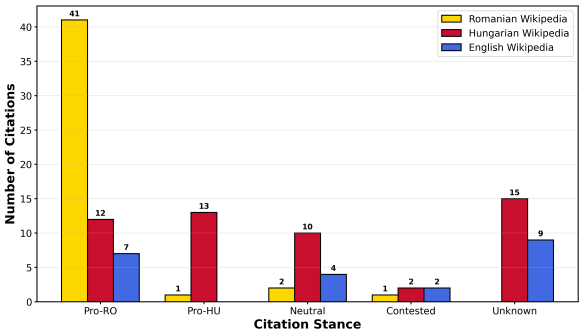

Figure 1: Raw count of citations by stance. Note the overwhelming volume of Pro-Romanian citations in the Romanian edition compared to the balanced distribution in Hungarian.

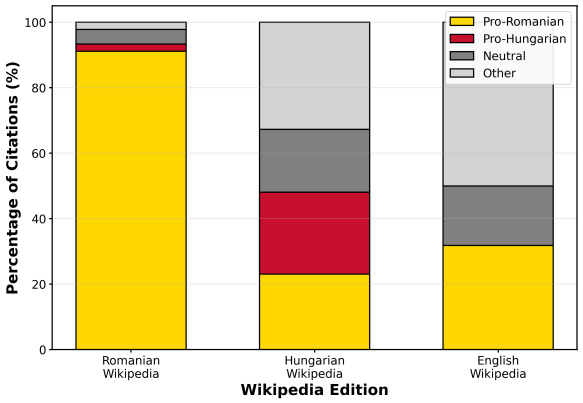

Figure 2: Stance distribution of citations in the Battle of Posada articles. Romanian Wikipedia is 91% Pro-Romanian, while Hungarian Wikipedia is balanced.

Bias Scores: The Romanian edition exhibited a strong national bias, with 91.1% of citations classified as *Pro-Romanian* and only 2.2% as *Pro-Hungarian* (Figure 2). In contrast, the Hungarian edition was remarkably balanced, with 23.1% *Pro-Romanian* and 25.0% *Pro-Hungarian* citations. The English edition, often assumed to be neutral, showed a moderate Pro-Romanian lean (31.8% Pro-RO vs 0% Pro-HU).

Figure 1 highlights the sheer volume disparity. It is important to note that raw citation volume is not a proxy for quality; a higher count could simply indicate a more detailed article. However, in this context, the volume reinforces the echo chamber. The Romanian article builds an "illusion of consensus" through the repetition of a large, monoperspectival corpus, whereas the Hungarian article achieves a balanced narrative with a smaller but more diverse set of references.

Citation Isolation: The most striking finding was the lack of overlap. Out of 119 unique ci-

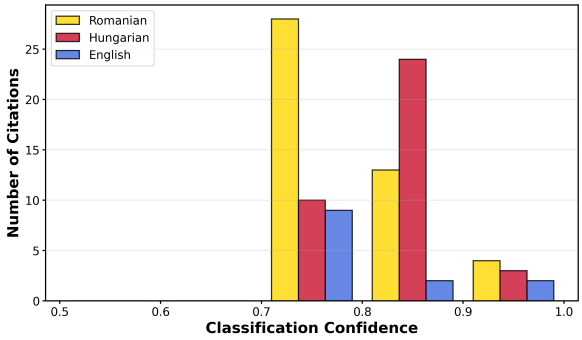

Figure 3: LLM Confidence Distribution. The high confidence scores across all stances indicate that the "Citation Isolation" is not due to model uncertainty but reflects clear, distinct narrative framing in the sources.

tations, only **2 sources** were shared across more than one language edition. This confirms that these Wikipedia communities operate in distinct "citation universes", constructing their narratives from entirely disjoint sets of historical authorities. It is important to note that this isolation is partly driven by linguistic accessibility: Romanian scholars naturally cite Romanian historians writing in Romanian, while Hungarian scholars cite Hungarian sources. However, the result is the same: readers of different language editions are presented with fundamentally different evidentiary bases. As shown in Figure 3, the LLM's classification confidence remained high across all categories, validating that these sources express clear, unambiguous stances rather than vague or subtle biases.

4.2 Experiment 2: Temporal Evolution

We tracked narrative metrics across 29 article versions from 2005 to 2024.

Bessarabia (1940): The Romanian article for "Basarabia" showed significant moderation, with its Pro-Romanian score dropping from a peak of 76 (2010) to 55 (2024) (Figure 4, Left). Conversely, the Russian article was highly volatile. After fluctuating between Pro-Soviet (30) and balanced (65) stances, the 2024 version of the "Accession" article spiked to a Pro-Romanian score of 72. We speculate that this may reflect a post-2022 shift where Russian editors are becoming more critical of Soviet expansionism, though further qualitative research is needed to confirm this causal link.

Figure 5 reveals the specific dimensions of this divergence. While both editions now acknowledge the Molotov-Ribbentrop Pact, they differ fundamentally on the consequences. The Romanian edition emphasizes "Deportations" (Score: 40) and "Victimhood" (Score: 80), whereas the Russian edition emphasizes "Soviet Justification" (Score: 40). Figure 6 tracks this "Deportation Gap" over time, showing that while Romanian mentions of atrocities have remained high, Russian mentions have been near-zero for two decades, only appearing slightly in 2024.

Night Attack (1462): The Romanian article peaked in nationalism in 2020 (Score: 85), portraying Vlad the Impaler as a "valiant defender", before moderating to 72 in 2024 (Figure 4, Right). The Turkish edition showed a gradual decline in Pro-Romanian sentiment (70 $\rightarrow$ 65) and consistently emphasized Vlad's cruelty (Score: 80) far more than the Romanian edition (Score: 60). The English edition remained remarkably stable at a score of 70 throughout the 14-year period.

The shape of these narratives is visualized in Figure 7. The Turkish narrative is defined by high "Cruelty" and "Ottoman Threat" scores but lower "Heroism". The Romanian narrative is the inverse. Figure 8 isolates the "Cruelty Gap", showing a persistent 15-20 point difference between Turkish and Romanian portrayals of Vlad's tactics.

4.3 Experiment 3: Peace-Maker LLM

We evaluated four prompting strategies for generating neutral summaries of the Night Attack, using conflicting claims extracted from Romanian and Turkish articles.

Prompt Strategy	Neutrality	Conflict Pres.
Standard	0.47	0.60
Mediator	0.95	1.00
Academic	0.97	1.00
Adversarial	**1.00**	**1.00**

Table 1: Performance of prompting strategies. Neutrality and Conflict Preservation are scored 0-1.

As shown in Table 1 and Figure 9, the *Standard* prompt failed (Neutrality: 0.47) because the LLM attempted to resolve the conflict, often choosing one side's version of events (e.g., regarding the attack's success). The *Adversarial* prompt, which explicitly instructed the model to *preserve* conflict, achieved perfect scores. Figure 10 breaks this down further, showing that the Adversarial prompt's success stems from its high "Source Attribution" score—it explicitly attributes disputed claims ("Romanian sources state X.."), avoiding the trap of hallucinated consensus.

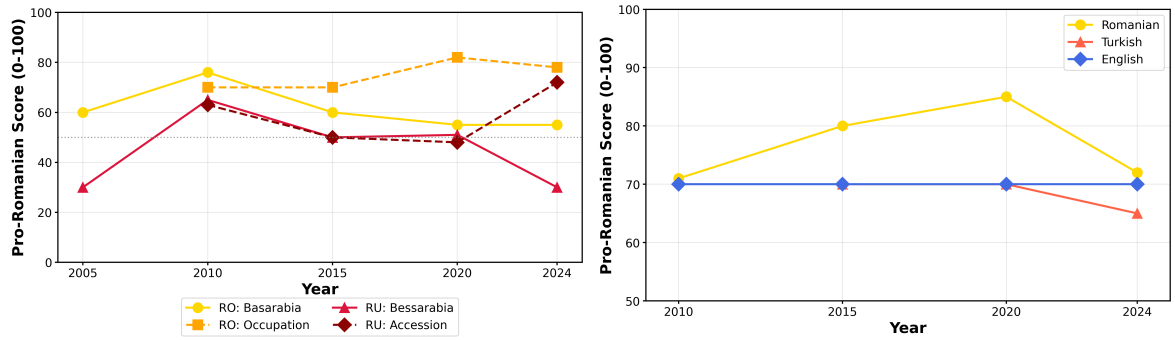

Figure 4: Evolution of the "Pro-National" score over time. Left: Bessarabia (RO vs RU). Right: Night Attack at Târgoviște (RO vs TR vs EN).

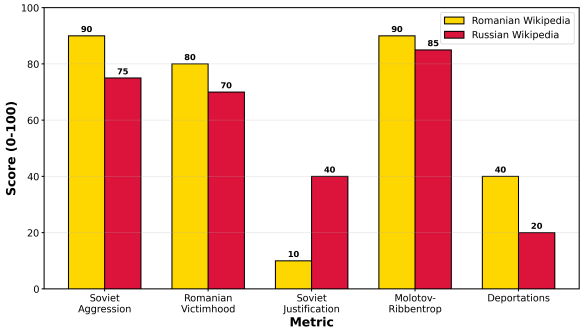

Figure 5: Detailed metric comparison for Bessarabia (2024). Note the divergence in "Soviet Justification" and "Deportations".

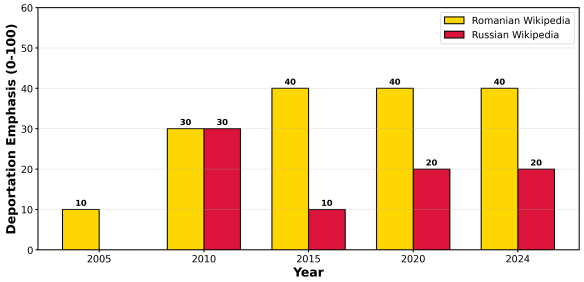

Figure 6: Deportation mentions over time. Romanian Wikipedia consistently highlights Soviet atrocities, while Russian Wikipedia largely omits them.

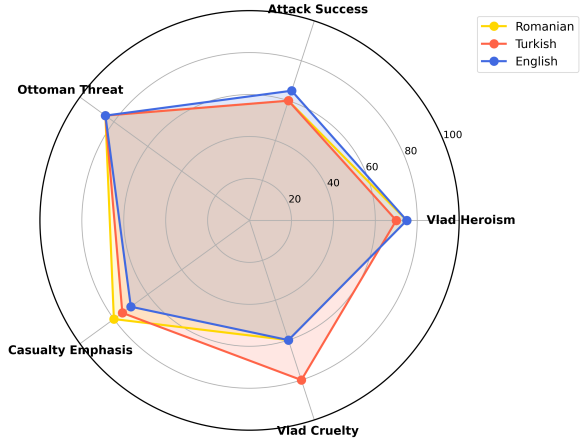

Figure 7: Radar chart of narrative metrics for the Night Attack (2024). The shapes illustrate distinct narrative priorities: Turkish (Cruelty-focused), Romanian (Heroism-focused), and English (Balanced).

5 Discussion

Our findings challenge the assumption that Wikipedia functions as a unified global encyclopedia. Instead, it operates as a federation of distinct epistemic communities, each validating knowledge through its own national lens.

5.1 The Illusion of Neutrality

The "Citation Isolation" we observed in the Posada case study shows the challenge of achieving neutrality. An article can adhere perfectly to NPOV guidelines, using neutral language and avoiding editorializing, while still presenting a heavily biased narrative simply by selecting sources from a single national tradition. This "bibliography bias" is invisible to standard NLP tools that focus on sentiment analysis, but it is the primary driver of narrative divergence in historical topics.

5.2 Narrative Evolution and Geopolitics

Our temporal analysis shows that these narratives are not static. The moderation of the Romanian "Night Attack" article (from 85 to 72) suggests that community maturity and international scrutiny can temper nationalism over time. However, the volatility of the Russian "Bessarabia" article demonstrates that Wikipedia is not immune to external geopolitical shocks. The sudden 2024 shift toward a Pro-Romanian stance likely reflects a broader realignment of Russian opposition discourse following the invasion of Ukraine, where Soviet imperial

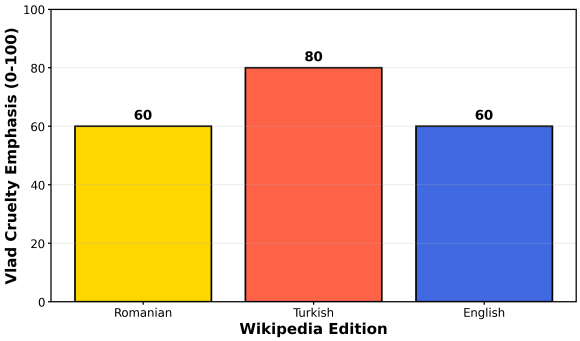

Figure 8: Comparison of "Vlad Cruelty" scores. Turkish Wikipedia consistently emphasizes Vlad's brutality much more than the Romanian edition.

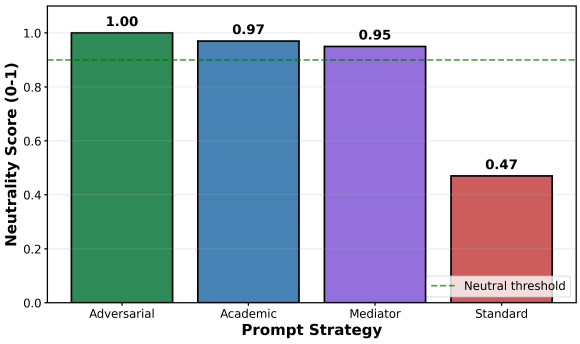

Figure 9: Neutrality scores (0-1) for different prompting strategies. The Adversarial prompt achieves perfect neutrality.

history is being critically re-examined.

Bias often manifests as silence. As Figure 6 illustrates, the Russian Wikipedia's historical omission of Soviet deportations is a form of narrative control. The sudden appearance of these mentions in 2024 signals a potential "thaw" in this historiographical freeze. Similarly, the "Cruelty Gap" in the Targoviste narrative (Figure 8) shows how national identity is constructed not just by what is celebrated (Heroism), but by what is minimized (Cruelty).

5.3 AI as a Mediator, Not a Judge

Standard LLM summarization fails because it mimics the human tendency to seek a single, coherent truth, effectively hallucinating a consensus where none exists. By explicitly prompting the model to be "adversarial" and preserve conflict, we force it to act as a mediator rather than a judge.

However, a distinction must be drawn between interpretive disagreements (e.g., whether a war was "justified") and factual contradictions (e.g., casualty counts). For the latter, "preserving conflict" does not imply that all claims are equally valid, but

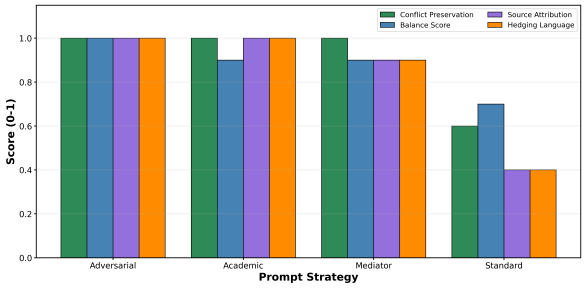

Figure 10: Detailed evaluation of generated summaries. The Adversarial prompt excels not just in Neutrality, but in Source Attribution and Conflict Preservation.

rather that the *disagreement itself* is a historical fact worth reporting. When Source A claims 15,000 casualties and Source B claims 1,500, a LLM that averages these numbers fails both. A LLM that reports the discrepancy preserves the epistemic integrity of the conflict. Future systems must balance this perspectivism with rigorous fact-checking to avoid "false balance" when one perspective is demonstrably pseudohistorical. This suggests that the future of AI in education and historiography lies not in generating "objective" answers, but in synthesizing and attributing diverse perspectives.

6 Conclusion and Future Work

This study demonstrates that Wikipedia functions not as a unified global archive, but as a federation of distinct epistemic communities defined by "citation isolation". Our analysis of the Battle of Posada revealed that the Romanian and Hungarian editions shared only 2 out of 119 citations, constructing mutually exclusive historical realities based on non-overlapping authorities. Longitudinally, we found these narratives to be volatile and responsive to geopolitics, evidenced by the 2024 shift in the Russian framing of Bessarabia.

Methodologically, we show that standard LLM prompting fails to address this by hallucinating consensus. Instead, we propose a "Peace-Maker" pipeline using "adversarial" prompting. By explicitly instructing models to preserve conflict rather than resolve it, we achieve "perspectival balance", generating summaries that accurately attribute disagreement without enforcing a false middle ground.

We propose several directions for future research:

- **Baseline Comparison:** Compare these contested events to uncontested historical topics to establish a baseline for citation overlap

and narrative divergence. This would verify whether the "citation isolation" we observed is specific to conflict or a general feature of Wikipedia's language editions.

- **Geographic and Domain Expansion:** While our case studies are well-chosen for their contentiousness, they are limited to Eastern European history. Future research should investigate if "Citation Isolation" holds true for global topics (e.g., WWII in the Pacific or Colonialism in the Americas) or scientific controversies.

Limitations

Our study relies on human annotators and LLMs. LLM-based classification, while calibrated, may still contain inherent biases. We focused on three specific events in Eastern European history; results may vary for other regions. Our validation methodology presents a potential circularity: annotators assigned by language (Romanian/English vs. Hungarian) were native speakers whose historiographical perspectives may reflect the national biases we sought to measure, rather than providing an independent benchmark. Annotator neutrality is hard to verify.

Ethics Statement

This study utilizes publicly available data under the Creative Commons Attribution-ShareAlike 4.0 license. To validate our LLM-based classifications on sensitive historical topics, we used two human annotators. Strict privacy protocols were maintained to preserve their anonymity. We emphasize that our computational metrics quantify "perspectival balance" and are not intended to adjudicate historical truth or resolve geopolitical disputes.

Acknowledgments

This research is supported by:

- the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416;

- a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV;

A Granular Narrative Scoring Prompt

The following prompt was used for temporal analysis of the Bessarabia and Târgoviște articles:

```
Analyze the following Wikipedia article excerpt about Bessarabia/Soviet occupation.
ARTICLE: {article}
LANGUAGE: {lang} (Romanian or Russian)
YEAR: {year}
TEXT:
{text}
—
Analyze the NARRATIVE STANCE of this article. Consider:
1. **Overall Stance**: Is the article Pro-Romanian, Pro-Russian, Neutral, or Mixed?
- Pro-Romanian: Portrays Soviet actions as occupation, theft, aggression
- Pro-Russian: Portrays Soviet actions as liberation, reunification, protection
- Neutral: Balanced presentation of facts without emotional framing
- Mixed: Contains elements of both perspectives
2. **Key Themes**: What are the main themes/topics discussed? (list 3-5)
3. **Emotional Tone**: What is the emotional framing?
- Accusatory (blaming the other side)
- Neutral (factual, academic)
- Defensive (justifying actions)
- Victimization (emphasizing suffering)
4. **Key Terminology**: What loaded terms are used? For example:
- "occupation" vs "liberation" vs "annexation" vs "reunification"
- "ultimatum" vs "agreement" vs "request"
- How is the USSR/Romania described?
Respond in this EXACT JSON format:
{
"overall_stance": "Pro-RO" or "Pro-RU" or "Neutral" or "Mixed",
"stance_confidence": 0.0 to 1.0,
"key_themes": ["theme1", "theme2", "theme3"],
"emotional_tone": "Accusatory" or "Neutral" or "Defensive" or "Victimization",
"terminology": {
"event_name": "what the event is called",
"soviet_actions": "how Soviet actions are described",
"key_loaded_terms": "any loaded/biased terms used"
},
"reasoning": "Brief explanation of why you classified it this way"
}
```

References

A. Baigutanova and Others. 2023. Reference reliability divergence in multilingual wikipedia. *arXiv preprint arXiv:2309.00196*.

Nishant Balepur, Alexa Siu, Nedim Lipka, Franck Dernoncourt, Tong Sun, Jordan Boyd-Graber, and Puneet Mathur. 2025. Mods: Moderating a mixture of document speakers to summarize debatable queries in document collections. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gemini Team, Google. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. Updated to include Gemini 3, November 2025.

Saba Ghanbari Haez and Mauro Dragoni. 2025. Neutral is not unbiased: Evaluating implicit and intersectional identity bias in llms through structured narrative scenarios. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Xin Guan, Pei-Hsin Lin, Zekun Wu, Ze Wang, Ruibo Zhang, Emre Kazim, and Adriano Koshiyama. 2025. Mpf: Aligning and debiasing language models post deployment via multi-perspective fusion. *Findings of IJCNLP 2025*.

Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 11–20.

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747.

Ben Kurtovic. 2023. mwparserfromhell: A parser for mediawiki wikicode. https://github.com/earwig/mwparserfromhell.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2023. Geopolitical bias in large language models. *arXiv preprint arXiv:2305.14610*.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in Physics*, 6:54.

Perspectivist Data Manifesto. 2020. The perspectivist data manifesto. https://pdai.info/. Accessed: 2026-02-09.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjectivity in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8669–8676.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Richard Rogers and Emina Sendijarevic. 2012. Neutral or national point of view? a comparison of srebrenica articles across wikipedia's language versions. In *Wikipedia Academy: Research and Free Knowledge*, Berlin. Wikipedia Academy.

Anna Samoilenko, Fariba Karimi, Daniel Edler, Martin Rosvall, and Markus Strohmaier. 2016. Linguistic neighbourhoods: explaining cultural borders on wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5:1–21.

Michael Taylor, Roisi Proven, and Carlos Areia. 2025. Evaluating the diversity of scientific discourse on twenty-one multilingual wikipedias using citation analysis. *arXiv preprint arXiv:2501.09666*.

Wikimedia Foundation. 2024. Mediawiki action api. https://www.mediawiki.org/wiki/API:Main_page. Accessed: 2025-02-09.