Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Whose story gets told? ChatGPT, collective memory, and ethical narratives in the age of Artificial Intelligence

Nuno Moniz
University of Notre Dame
Jasna Ćurković Nimac
jasna.nimac@unicath.hr
Catholic University of Croatia
Ivana Brstilo Lovrić
Catholic University of Croatia
Valentina Petrović
Catholic University of Croatia
Paula Zujić
Catholic University of Croatia

Article

Keywords: Collective memory, artificial intelligence, ethics, ChatGPT, historical narrative

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10456035/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This paper explores the ethical implications of large language models (LLMs), with a focus on ChatGPT's role in mediating and shaping collective memory in the digital age. Using a multilingual qualitative approach, the study conducted from June to August 2024 analyzes ChatGPT's responses to historically significant multinational conflicts to examine how AI-generated narratives may influence public understandings of the past. The questionnaire protocol was applied uniformly in the national language or languages associated with the conflict, as well as in control English, which served as a reference point due to its global reach and significance. While AI tools such as ChatGPT may contribute to the democratization of historical knowledge and improve access to diverse perspectives, they also risk distorting memory in the absence of ethical principles, human agency, and critical reasoning. The findings reveal how collective memory is shaped by algorithmic bias, factual asymmetries, and representational imbalances. Identified issues include incorrect information, inconsistent answers across languages, omission of key facts, biased portrayals, oversimplification, uneven detail, contextual mismatches, and narrative neutralization. These patterns raise concern about the reduction of complex histories into algorithmic logic. Ultimately, the study argues for ethically grounded use of AI in historical interpretation to prevent collective memory from becoming a product of algorithmic representation rather than critical reflection.

Introduction

Technological revolutions have consistently reshaped societies, redefining the means of production, communication, and cultural exchange. Castells (2000), in his theory of the network society and digital age, argues that the information revolution differs from earlier industrial revolutions by centering on information processing and communication technologies. These transformations alter not only economies but the very fabric of social life. In the preface to Castells's book, Croatian sociologist Katunarić (2000: 26) makes a far-reaching prediction that, as a result of this technological turn, cultural conflicts will increasingly be played out in virtual spaces. This aligns with postmodern perspectives that emphasize the collapse of boundaries between culture and society, driven by media and technology that simulate and (re)produce reality as global flows of images, information, and capital erode traditional, linear perceptions of time and space and collective and individual identity (Haralambos and Holborn, 2002).

Due to the connective turn taking place in the contemporary age, digital communication is becoming ever-present, and the sharing of information so widespread that it creates "shadow archives", a lasting yet invisible presence of the digital traces we leave behind (Hoskins, 2018: 105–106). The connective turn also affects how we understand memory and the role it plays. Paradoxically, digital connectedness simultaneously enables the preservation of the past and makes it more difficult, because the digital world can render the past always accessible while at the same time making it too easily accessible (Hoskins, 2018: 1–2). Yet digital memory, although unstable and often subject to change due to its adaptability, does not signify the end of collective memory. This does not imply the emergence of a new

ontology of memory that requires redefining or abandoning the traditional concept of collective memory; rather, it highlights the importance of analyzing these new transformations and employing new, different tools that can enable a better understanding of present conditions and help set guidelines for the future (Mandolessi, 2023: 1514).

The new interconnection between memory and technology is also evident in the storage and specific organization of data, where there is an increased reliance on digital platforms and artificial intelligence systems to access collective memory. These systems draw conclusions from collected data and generate responses for users (Mandolessi, 2023: 1521; Hammond, 2015), thereby becoming "gatekeepers of memory who shape how we understand and experience the past" (Makhortykh, 2021: 182). In this context, large language models (LLMs) such as ChatGPT do not merely store or transmit historical information. They act as mediators and co-creators of historical narratives, raising critical ethical questions about their influence on collective memory. This paper investigates how ChatGPT's responses to historical events illustrate AI's capacity to shape public recollections, intentionally or otherwise. The theoretical foundation combines insights from memory studies and digital ethics to examine how algorithmic design, training data, and technological infrastructure influence the credibility, inclusiveness, and neutrality of historical narratives. This framework supports an empirical case study of ChatGPT's responses to queries about key multinational conflicts, focusing on how memory is shaped, omitted, or biased in digital interactions.

1. Collective Memory in the Digital Age

Memory, both individual and collective, is essential for interpreting the world and forming judgments. Collective memory refers to the shared representations of a group's past, integrated into common narratives that foster social identity and cohesion (Misztal, 2021). It is not static but actively constructed through commemorative practices, symbolic frameworks, and increasingly, digital technologies. For Halbwachs, collective memory is socially framed because social groups determine what will be remembered, what is considered worth remembering, and how something will be remembered (Halbwachs, 1950; Misztal, 2021: 76). In explaining the concept of collective memory, Halbwachs rejects the possibility of individual memory existing independently of collective memory, since individual memory is embedded in elements of interaction with others, and individuals can acquire memories and recall them only through their belonging to a collective (Halbwachs, 1950, as cited in Handžić Ferović, 2020: 156).

As Olick and Robbins (1998) argue, collective memory is inherently subjective and normative, often contrasting with historical objectivity. Jenkins (2008) insists we must interrogate whose interests are being served by dominant historical narratives, while Shur-Ofry and Pessach (2020) highlight how memory can align with political agendas. Similarly, Novick (1999) emphasizes its instrumental use in legitimizing power. All these concerns are echoed by Derrida, who claimed that "*there is no political power without control of the archive*" (1996: 4). His point underscores the power dynamics inherent in memory governance. It is important to emphasize that memory is not autonomous but is always, in one

way or another, instrumentalized by different centres of power, whether state or state-like powers, or private actors and interests. In the context of generative artificial intelligence, it is particularly important to recognize that it is not a form of state power, but rather a product of industry and private interests. In such an environment, external influences on memory become more visible and tangible, while the power of the private sector acquires increasing significance. Power, moreover, spreads in a capillary manner through the interconnection of the AI systems analysed here with state policies, but also through private channels that are less visible and more difficult to verify. As a result, the very understanding of political power is changing. The role of state power is thereby relativized in relation to the growing technological oligarchy, which has become sufficiently powerful for us to speak of its partial control over the archive. In the digital age, these dynamics are intensified by algorithmic systems and technological infrastructures that shape what is remembered, stored, amplified, or excluded, raising pressing ethical concerns about representational fairness (Schwartz and Cook, 2002). However, these ethical questions cannot be separated from relations of power, as they depend on who governs digital infrastructures, determines the criteria of visibility and content selection, and controls the production and mediation of knowledge. In this article, we therefore understand power in a broader sense, as a relation that is not limited exclusively to state actors. Future analyses should consequently examine the governance of digital collective memory in greater detail from the perspective of state power, as well as from the perspective of private and non-state interests.

The digital turn has transformed memory from a curated and relatively stable process into a fluid, fragmented, and volatile phenomenon. According to Mandolessi (2023), digital memory is increasingly shaped by decentralization, real-time updating, and algorithmic mediation. Hoskins (2018) describes this new condition as a “post-scarcity culture,” characterized by an overabundance of information that complicates discernment and amplifies moral responsibility. In this context, Floridi reminds us that the easy availability of data increases our ethical obligation to verify it: “the more any bit of information is just an easy click away, the less we shall be forgiven for not checking it” (2014: 42–43).

This vast availability of digital information, however, is unevenly distributed. The spatial and temporal asymmetry of digital data leads to biases, older events or those in regions with weak digital infrastructure tend to be underrepresented. Wendy Chun’s (2008) concept of the *enduring ephemeral* captures this paradox: although data may appear permanent, digital memory is in fact mutable, unstable, and easily manipulated. Rigney (2016) similarly introduces the notion of *mnemonic regime shifts* to describe how dominant frameworks of remembrance change over time and technological context.

Within this uneven, unstable, and algorithmically mediated environment, artificial intelligence introduces an additional layer of complexity. Artificial intelligence systems generally do not produce new facts through direct observation of the world, rather they synthesize, reorganize, and mediate knowledge originally produced by journalists, researchers, scientists, public institutions, and other primary producers sources. As users increasingly rely on AI generated responses instead of consulting original sources, there is a risk that the visibility, revenues, and institutional incentives that support the production of original content will diminish. In the long term, such a dynamic could weaken the very

epistemic infrastructure on which AI systems themselves depend. Such an outcome, however, is not inevitable. Artificial intelligence may also function as an intermediary mechanism that directs users toward reliable sources, broadens access to expert knowledge, and facilitates the identification of relevant information under conditions of informational abundance. The important question, therefore, is not only whether artificial intelligence will replace sources of information, but also whether the value created by primary knowledge producers will be adequately recognized and sustained. Unless appropriate mechanisms of compensation and institutional support are established, the broader information ecosystem may gradually weaken. Preserving the resources and incentives required for the production, verification, and contextualization of new information is therefore one of the prerequisites for the long-term reliability and sustainability of digitally mediated knowledge.

As we mentioned above, memory is never neutral, and it is always mediated (Van Dijck, 2007; Garde-Hansen, 2011). Technologies do more than simply record the past; they construct, frame, and reinterpret it. Sturken (2008) speaks of “technologies of memory” that actively shape how the past is conceived. In the digital realm, this mediation is no longer the domain of experts alone. In the digital environment, users are no longer merely observers but active participants (produsages) who simultaneously consume, create, and share content, leading to information increasingly being shaped through collective, collaborative practices (Bruns, 2008). De Kosnik (2016) introduces the idea of *rogue archives*, alternative, often volunteer-driven digital repositories that preserve marginalized or suppressed histories, serving as tools of *counter-memory* (Misztal, 2021).

Further complicating the structure of digital memory is the tension between narrative and database logic, understood here as the organization and retrieval of information through discrete, searchable, and reconfigurable elements rather than through a coherent sequence of events. Manovich (2001) argues that the dominant form of cultural logic has shifted from narrative (which offers coherence and causality) to database (which is fragmented, modular, and reconfigurable). Although databases may appear anti-narrative, they also enable new, unexpected ways of understanding the past. LLMs produce only imitations of narrative, operating according to a horizontal logic that predicts words and sentences based on statistical patterns drawn from large databases, rather than from coherent causal reasoning or source-based interpretation. As Mandolessi (2023) and Kansteiner (2022) observe, this fragmentation can revitalize memory by revealing hidden patterns and alternative perspectives. This dynamic interplay between narrative and database reflects both the challenge and opportunity of digital memory systems. While they disrupt established conventions of remembering, they also provide new tools for historical exploration and participation. In this new memory ecology, the boundaries between documentation, mediation, and interpretation are increasingly blurred.

2. Ethics and AI

The integration of memory into technological systems signals a move toward algorithmic mnemotechnologies as AI tools that organize and manipulate memory through statistical and behavioral logic (Mandolessi, 2023). These systems shape what is remembered, forgotten, and emphasized. They

frequently reflect societal biases embedded in data or system design. Generative AI models like ChatGPT can reinforce or reproduce misinformation, selective narratives, or ideological framings depending on how they are trained. As Makhortykh (2021) notes, AI tools increasingly serve as memory gatekeepers, determining the informational contours of collective recollection. Shur-Ofry and Pessach (2020) propose the idea of memory fiduciaries as entities responsible for ethical stewardship of memory. This includes transparency, accuracy, and fairness in how historical content is curated and delivered.

Although algorithms that consider ethical dimensions are being developed (Kearns and Roth, 2020), they remain partial solutions. Human agency and critical interpretation remain essential, especially in historical contexts. Without these, AI risks becoming a tool of algorithmic mediation, flattening memory into neutralized, depoliticized, or manipulated narratives. Large-scale language models such as ChatGPT, which constitute the empirical focus of this study, exemplify the epistemological and ethical challenges of AI-mediated memory. These models are not inherently ethical or neutral; rather, they reflect the biases embedded in their training data, shaped by inconsistencies across datasets, representational imbalances, and the consequences of fine-tuning and optimization strategies. This duality between technical sophistication and epistemic fragility holds both promise and risk for the culture of memory.

Above mentioned underscores the indispensable role of human agency: critical reasoning, digital literacy, and reflective interpretation are essential when engaging with algorithmically generated content. Recent research reinforces these concerns. Buyl et al. (2023), in their study *Large Language Models Reflect the Ideology of Their Creators*, show that LLMs exhibit divergent ideological tendencies across languages and contexts. Their analysis of model-generated portrayals of controversial historical figures reveals systematic normative differences, suggesting that LLMs mirror the cultural and political assumptions of their creators. This challenges the common claim of AI neutrality and highlights the potential for political instrumentalization of generative systems (Buyl et al., 2023; Aksoy, 2025).

Similarly, Kansteiner points out that tools like ChatGPT “cannot differentiate between truthful and untruthful statements” (2022: 125) and tend to amplify structural biases found in their training material. While he remains cautiously optimistic about AI’s potential for historical research, he warns that such tools may reproduce propaganda or falsified narratives with the same technical polish as genuine content. To address these risks, Kansteiner emphasizes the necessity of human oversight at both input and output levels. As he puts it, ChatGPT “will never be able to tell us where the content of a given sentence originated, because each new word is the result of thousands and thousands of calculations of probability.

By transforming how memory is accessed and utilized, ChatGPT has significant implications for both the preservation and reinterpretation of history. Its conversational framework illustrates the transition to a prosumer model (Toffler, 1980), where users simultaneously consume and produce memory narratives. This debate directly frames the present research, which explores similar questions through a focused

examination of ChatGPT's handling of multinational conflicts across languages, with a specific interest in its potential role in shaping collective memory.

Methodology

The general aim of this study is to investigate how ChatGPT interprets multinational conflicts across different languages, with a particular focus on the accuracy and consistency of its responses, as well as the ethical implications these may have for society and collective memory. To achieve this, we employed a qualitative research design, conducted from June to August 2024, that treats ChatGPT both as a subject of analysis and as a research tool. Its selection is based on its status as a state-of-the-art large language model (LLM) that processes and interprets vast amounts of multilingual and multicultural data. ChatGPT offers a unique lens through which to examine how artificial intelligence technologies are reshaping traditional practices of collective memory formation in the digital age.

The specific version used during the study was ChatGPT-4o. Although the default version is freely available to a global audience, the paid version was employed in this research to ensure analytic consistency and extended prompt capacity (up to 80 prompts per session before system reset). This choice allowed for more comprehensive interaction with the model and ensured comparability across data collection points.

Unlike traditional memory institutions such as archives, museums, or media organizations which typically function as curated, hierarchical gatekeepers of historical knowledge, ChatGPT enables immediate, global, and decentralized access to historical content. This shift challenges established paradigms of memory-making and introduces new, democratized modes of engaging with history. A key feature of ChatGPT that informed its methodological value is its capacity for dynamic, dialogical interaction with users. This real-time co-creation of content allows researchers to explore the fluidity and participatory nature of digital memory construction.

In this study, ChatGPT was conceptualized as an actant (Latour, 2005), a digital entity occupying a liminal space between the animate and inanimate due to its interactive nature and algorithmic capacity for learning and response generation. This framing enabled a more nuanced understanding of its role in shaping historical narratives. It also emphasized the evolving role of user agency in AI-mediated memory processes, where memory is not merely retrieved but actively co-constructed. Alongside interpretive depth, ethical concerns were central to the study. ChatGPT's outputs were analyzed with attention to potential biases, omissions, and representational practices. As a widely used AI system increasingly embedded in educational and public discourse, understanding how it frames history is essential to assessing its influence on collective memory and societal understanding of the past.

3.1. Case Selection and Units of Analysis

The empirical component of the study focused on four historically significant and geopolitically diverse multinational conflicts: the war in the former Yugoslavia, the Israeli–Palestinian conflict, the Portuguese

colonial wars in Africa, and the Vietnam War. These conflicts were selected to reflect a variety of conflict types: ethnic conflicts, international conflict, anti-colonial struggle and postcolonial geopolitical tension and long-standing international disputes, to ensure cultural and temporal diversity in the analysis. The primary units of analysis were the responses generated by ChatGPT to a set of structured and standardized questions. These responses were thematically organized and analyzed in terms of content, form, and underlying assumptions. The structure allowed for comparative insights into how different conflicts are represented across languages and contexts.

3.2. Multilingual Framework

A core feature of the research design was its multilingual approach, developed to explore potential narrative, factual, and interpretive variation across different linguistic settings. For each conflict, questions were posed in a national language of the group involved into that conflict: Croatian and Serbian for the Yugoslav war, Arabic and Hebrew for the Israeli Palestinian conflict, Portuguese for the Portuguese colonial wars, and Vietnamese for the Vietnam War. These languages were chosen to simulate the perspective of a local, native-speaking user, testing how ChatGPT reflects and adapts to culturally embedded memory frames. In parallel, all cases were examined in English, which served as a universal reference point given its status as the dominant language of international academia and communication. English thus provided a comparative framework for analyzing differences between nationally grounded and globally generalized historical narratives.

3.3. Data Collection and Procedure

Data collection was conducted through structured interviews with ChatGPT, based on a research protocol specifically designed for this study. The protocol consisted of eight standardized questions, crafted to elicit responses that address key historical dimensions of each conflict. These included inquiries into the major opening and closing events of each conflict, the actors involved, and the primary victims, with particular attention given to the causes and consequences of the conflict. ChatGPT was also asked to describe the event that immediately followed the conflict's onset, and this question was repeated iteratively in order to trace the chronological development of the narrative and determine the point of response saturation defined as the moment when the generated answers became repetitive or ceased to introduce new information.

In addition, the protocol sought to uncover the bibliographic foundations upon which ChatGPT based its responses, with the aim of evaluating transparency and source reliability. This methodological structure allowed for the detection of how ChatGPT prioritizes, organizes, and justifies historical content, and how it engages with sources across different languages. The interview format enabled close observation of the model's depth of knowledge, coherence, and adaptability. Follow-up prompts and occasional control questions were incorporated to test consistency and to probe for clarification or elaboration where responses appeared vague or incomplete. The entire interaction was treated as a dynamic exchange, emphasizing the model's role in actively co-constructing the narrative content, rather than passively retrieving information.

3.4. Analytical Strategy

Collected responses were analyzed using a multi-phase thematic analysis. First, historical inaccuracies in the ChatGPT 4o generated responses were reviewed by historians specializing in the respective conflicts analyzed in this study[1]. Each expert assessed the responses related to the historical case within their field of expertise and identified factual errors or problematic interpretations. Then, the authors of this paper proceeded with open coding of ChatGPT's answers to identify key patterns, anomalies, and content gaps. These codes were then grouped into broader analytical categories including factual accuracy, representational bias, narrative omission, cultural framing, and cross-linguistic consistency. Subsequently, a comparative analysis was performed to detect consistencies and divergences between responses in different languages and for different conflicts. Special attention was given to identifying how ChatGPT integrates, distorts, or omits historical information depending on the linguistic and cultural framing of the input. Finally, the themes were interpreted through the lens of collective memory theory, enabling reflection on how digital, algorithm-driven memory systems align with or diverge from traditional forms of historical mediation. This process made it possible to assess ChatGPT's role not only as a disseminator of knowledge but as a potentially active agent in shaping how the past is remembered in the present.

Results and discussion

In line with the overarching aim of the study, this section focuses on an analytical examination of the results obtained. The research proceeds from the assumption that modern chatbots, such as ChatGPT 4o in our case, when addressing historically sensitive topics, generates narratives that reflect not only the available data but also its structural limitations. Accordingly, ten analytical categories are identified and systematized into 3 levels: individual response, interview and event level, with the addition of illustrative examples for discussion. Response-level concerns issues directly related to single answers from chatbots.

Interview-level scales this to a conversation (i.e. session, with multiple iterations). Finally, event-level includes categories that consider multiple sessions/conversations, mostly across different languages. Particular attention was given to recurring issues such as inaccuracies in dates, names, or events; inconsistencies in responses across language contexts; and omissions of significant details or events. Bias was another critical theme, with instances of preferential language or framing that appeared to favour one side of a conflict. Simplifications and uneven levels of detail were noted, alongside cultural or contextual misalignments in responses. Instances of neutralization, where ChatGPT avoided clear stances or presented overly balanced narratives that downplayed the complexity of conflicts, were also identified. The list of categories is described in Table 1, and illustrated in Fig. 1.

Table 1
Table of categories for analytical examination.

Level	Category	Description	Example
Response	Inaccuracy	Refers to wrong dates, names, and events, including fictional references.	The statement "*Immediately after the initiation of Operation Gordian Knot in October 1964, the next significant event was the Battle of Kandala*". Operation Gordian Knot took place in 1970, and there is no historical record of a Battle of Kandala.
	Omission of Key Information	Skipping important dates, events, or actors crucial to understanding the conflict.	Concerning the claim "Immediately after the Battle of Ia Drang, the United States continued to escalate its military involvement in Vietnam". It creates the impression of a direct connection between the Battle of Ia Drang and subsequent U.S. escalation, omitting the Johnson Administration's decisions made between 1964 and 1968. Moreover, deeper themes including massacres, the impact of chemical weapons on civilians, instances of anti-Asian racism, or the notion that the American forces could not defeat the Viet Cong or the Viet Minh are omitted from the narrative.
	Contextual Mismatch	Omission of essential cultural, political, or linguistic context necessary for a fuller understanding of an event.	The statement "Vukovar's fall became a symbol of the brutal nature of the war and a key turning point in the conflict in Croatia" does not include further explanation as to why this event was symbolic. Without a broader context, the significance of Vukovar's fall remains unclear.
Interview	Uneven Detail	Providing inconsistent amounts of detail in answers or narratives.	In the interview on the Vietnam War in English, the narrative proceeds operation by operation or battle by battle attempting to portray American military strategy as generally successful. For example: "Battle of Ia Drang, Operation Starlite, Operation Hump, Operation Harvest Moon, Operation Masher/White Wing, Operation Birmingham, Operation Georgia, Operation Indiana, Operation Oregon, Fall of Saigon and more". The Vietnamese-language interview frames the war mainly through major political turning points and post-war state consolidation, with relatively little attention to the operational–military detail that dominates the English interview.
	Neutralization	Use of general language that downplays or omits specific information about particular actors or circumstances.	For example, "The main and largest victims of the war were members of all ethnic communities in various conflicts", illustrates neutralization by employing a broad formulation that avoids identifying specific parties, the scale of events, or relevant contextual differences.

Level	Category	Description	Example
	Simplification	Ignoring the complexity and relevant aspects of a conflict.	In the case of the Portuguese Colonial Wars, ChatGPT 4o outputs merge distinct conflicts in Angola (MPLA), Mozambique (FRELIMO), and Guinea-Bissau (PAIGC) into a single, simplified storyline. This compression obscures important tactical differences, local resistance dynamics, and region-specific contexts, resulting in a flattened historical account that underrepresents the complexity of parallel anti-colonial struggles.
Event	Inconsistency	Refers to different answers to the same question in different languages that are not aligned with local/national interpretations.	To the question "What is the last most significant event related to the war in the former Yugoslavia in the 1990s?" In English, the answer is "signing of the Ohrid Framework Agreement (…)". In Croatian, the response is: "signing of the Dayton Peace Agreement". In Serbian, the answer is "the arrest and trial of Ratko Mladić (…)"
	Biased Portrayal	Favoring one side through choice of expressions and descriptions.	In the case of the Israeli-Palestinian conflict, biased representation is evident in how language-specific outputs privilege certain perspectives. English narratives emphasize global geopolitical dynamics and state-level actors, Hebrew outputs focus on Israeli security concerns and state perspectives, while Arabic outputs highlight humanitarian issues in Palestinian territories, often reducing complex social and political dynamics to simplified humanitarian narratives.

The categories of *Inaccuracy* and *Inconsistency* point to the absence of a stable temporal and factual structure, rather than a reflection of a unified historical knowledge base. The category *Omission of Key Info* demonstrates that models often produce narratives that appear historically plausible yet lack essential interpretive content, which reduces descriptive precision and affects the interpretation of historical events. The category *Biased Portrayal* shows that chatbots may privilege particular perspectives depending on language and does not function as a neutral intermediary in the transmission of historical knowledge. Furthermore, the next two categories point to structural tendencies in how models organize historical events and their complexity. *Simplification* is reflected in compressed accounts that merge distinct conflicts into a single narrative frame, while *Uneven Detail* reflects an asymmetrical narrative structure in which information and detail are distributed unevenly, both within the same event and across languages. In addition, *Contextual Mismatch* and *Neutralization* show that chatbots often produce semantically correct but contextually impoverished formulations that may depoliticize events, obscure actor responsibility within particular conflicts, or disrupt interpretive coherence across languages. Taken together, these patterns show that chatbots, and ChatGPT 4o specifically, does not passively reproduce historical content, but actively reconstructs it through processes of compression, selection, and linguistic framing. The implications of these dynamics raise

broader questions of historiographical responsibility, cross-linguistic accuracy, and the future of digitally mediated historical memory.

Discussion. The analysis of LLM-generated historical narratives indicates that ChatGPT 4o reconstructs historical events in ways shaped by the language of the prompt, the broader linguistic context, and the data on which the model was trained. Although the models are capable of producing coherent and formally structured historical descriptions, the typology developed in this study shows that these narratives systematically reproduce a range of epistemological limitations, biases, and interpretative distortions across languages and conflicts. The established typology of ten error categories enables a systematic understanding of how LLMs transform and at times misrepresent historical realities.

a. Use Case: War in the former Yugoslavia (1990s)

To better understand the impact and depth of the insights afforded by the proposed framework, we provide a real-world analysis of historical narratives concerning the war in the former Yugoslavia (1990s). In this paper, the term “war in the former Yugoslavia”, refers to the period from 1990 to 1999s. These war comprised a series of interconnected armed conflicts arising from the dissolution of the Socialist Federal Republic of Yugoslavia and the process through which individual Yugoslav republics gained independence. Slovenia, Croatia, Bosnia and Herzegovina, Serbia, Montenegro, Macedonia, and Kosovo were involved in the war in the former Yugoslavia.

In this analysis, we leverage the data already collected using the methodology described in Section 3., which was applied with ChatGPT 4o, using three separate languages: Croatian, Serbian and English. The results from the first two languages were then translated to English to provide an equal point of analysis and comparison. Using the information in Table 1. describing the categories and their respective examples, all of the responses by ChatGPT 4o were assessed as to their fit regarding the proposed categories, at response-, interview-, and event-level. Results are described in Table 2.

Table 2
Results of the application of the proposed methodology and categorization to the interviews with ChatGPT on the events related to the war in the former Yugoslavia. Results at Response-Level and Interview-Level refer to the percentage of responses where the category is present. Event-level results refer to aspects of the interview (such as a question or a topic) where the category is present.

		Croatian	Serbian	English
Response-Level	Inaccuracy	41.86%	57.14%	14.28%
	Omission of Key Information	16.28%	50%	42.86%
	Contextual Mismatch	4.65%	28.57%	21.43%
	No Issue	34.88%	7.14%	21.43%
Interview-Level	Uneven Detail	41.86%	92.86%	85.71%
	Neutralization	18.60%	78.57%	71.43%
	Simplification	16.28%	42.86%	35.71%
Event-Level	Inconsistency	2		
	Biased Portrayal	3		

Response-level analysis. Languages present a noticeable degree of disparity in the issues found. In Croatian, 28 of 43 phrases (65%) contained at least one issue, with *Inaccuracy* being the dominant problem (18 phrases, nearly half the entire dataset). Further analysis suggests a systematic misunderstanding or hallucination in the source model rather than isolated mistakes. Specifically, the most persistent were the incorrect U.S. recognition date (January 1992 instead of April 1992, appearing 5 times) and the wrong date for Bosnia's independence declaration (April 5 instead of March 3, 4 times). Organized by topics covered in this interview, we notice that the coverage of the Croatian War of Independence had the highest error rate (83%), driven heavily by chronological confusions around the JNA intervention, the Dubrovnik siege, and international recognition. The Dayton and post-peace phase was the topic with least inaccuracy, with only 1 of 7 phrases flagged. As for other response-level issues, omissions of key information and contextual gaps, these tended to cluster around the Brioni moratorium (systematically omitted in early phrases), the failure of the international community during the Bosnian War, and the downplaying of key actors like the JNA. In the Serbian interview, 13 out of 14 phrases (93%) showed at least one issue, indicating a high overall error rate across the dataset and multiple cases of more than one issue per response (43%). In this case, *Omission of Key Information* was the most prevalent issue, appearing in over half the phrases, suggesting that the text systematically underrepresents actors, causes, and consequences, particularly around atrocities (Ovčara, ethnic cleansing during Operation Storm), failed peace processes, and contested outcomes like Dayton. *Inaccuracy* is nearly as common an issue, with chronological errors being the most frequent subtype, notably, the reversal of Operation Storm and Dayton, and the misplacement of the ICTY's founding. *Contextual Mismatch* showed in events whose symbolic weight is asserted but not explained (Vukovar,

Dubrovnik, Montenegro's independence, Milošević's transfer to the ICTY). Finally, concerning the English interview, The dominant issue was also *Omission of Key Information* (43% of phrases), reflecting a pattern of presenting events in a simplified linear chain while leaving out crucial actors, events, and causal dynamics. *Inaccuracy* is the least common issue, although including a significant factual error Operation Deliberate Force's UN authorization status.

Interview-level. Issues are more prevalent at interview-level, as detailed below. Although present, the Croatian interview was the one with less presence of all categories. In terms of *Uneven Detail*, the narrative dedicates disproportionate attention to diplomatic and institutional milestones, particularly the Dayton Agreement and its implementation, which receive five or more separate entries, while entirely omitting major military events that are central to any historically credible account of the conflict (e.g., Battle of Vukovar, Operation Storm, Operation Flash, and the existence of the Republic of Serbian Krajina). *Neutralization* is present throughout in the form of language that systematically softens or obscures the nature of specific events and the responsibility of specific actors: the Srebrenica genocide referred to as "one of the darkest moments" and "this crime," ethnic cleansing is reframed as control of "territory they considered Serbian," and the JNA's direct role in aggression is repeatedly diluted through passive constructions such as "escalation of tensions." *Simplification* also manifests in the compression of several historically distinct conflicts into a single linear arc: the Ten-Day War in Slovenia is conflated with the Croatian War of Independence, the entire Bosnian War is reduced to a Sarajevo–Srebrenica–Dayton sequence, and the Croat-Bosniak War, the Washington Agreement, the concentration camps of 1992, and the successive failures of UNPROFOR's mandate are entirely absent. These patterns produce a narrative that is structurally coherent but historically thin, tending to foreground international institutional responses while suppressing the military, humanitarian, and legal dimensions of the conflicts it purports to describe. In the Serbian interview, all issues are much more present. The *Neutralization* issue is pervasive: the model systematically employs abstract, symmetrical, or procedural language that avoids naming perpetrators, attributing specific acts of violence, or conveying the scale of atrocities. This is compounded by *Uneven Detail*, the most prevalent issue in this interview: heavily destructive events (Kosovo War, Srebrenica, siege of Sarajevo) receive either no mention or a single sentence, while diplomatic instruments receive multiple paragraphs, although lacking substantive content. The *Simplification* category is particularly evident in the treatment of Operation Storm (omitting the Serb civilian exodus and contested legal status), the repeated flattening of Dayton into "peaceful coexistence," and the document's structural drift toward EU integration milestones as a substitute for reckoning with the wars' violence. Taken together, the interview produces a narrative of the Yugoslav War as a series of diplomatic events bracketed by tragedy, in which no specific party bears clear responsibility and institutional progress serves as closure. Finally, in the English interview, a similar degree of presence of issues to the Serbian interview is observed. In terms of *Uneven Detail*, the narrative allocates substantial depth to the Bosnia-Srebrenica-Dayton sequence while systematically underrepresenting the Croatian War of Independence (omitting Vukovar, Dubrovnik, and Operation Storm entirely), and providing only superficial coverage of the post-2008 Kosovo diplomatic situation. *Neutralization* manifests most consequentially in the repeated use of symmetrical framing, particularly

the characterization of the Bosnian War as an "ethnic conflict among" three equal parties, the consistent use of "warring factions" to avoid identifying primary aggressors, and the hedging of the Srebrenica genocide as something "widely regarded" as an atrocity rather than a judicially confirmed act of genocide under international law. *Simplification* is less frequent but includes a structurally significant factual error: the false claim that Operation Deliberate Force was conducted without UN Security Council authorization. This is a misattribution that conflates the 1995 Bosnia campaign with the 1999 Kosovo operation and misrepresents the legal basis of NATO's intervention. More broadly, the document imposes a linear, sequential narrative logic (Slovenia leading to Croatia leading to Bosnia leading to Kosovo) that obscures the overlapping timelines, independent causal dynamics, and distinct actor configurations that characterized each of the Yugoslav War as a separate conflict.

Event-level. A cross-language analysis of the interviews reveals systematic, language-conditioned divergences in both factual scope and normative framing that cannot be attributed to legitimate contextual variation. Regarding the *Inconsistency* category, the three language versions provide fundamentally different answers to the same question ("What is the last most significant event related to the war in the former Yugoslavia in the 1990s?") without any stated justification for the divergence. The English output concludes with the signing of the Ohrid Framework Agreement (2001, North Macedonia), a post-1990s peace deal ending a distinct Macedonian insurgency that is historically remote from the Bosnian and Croatian conflicts the question implies. The Croatian output identifies the Dayton Peace Agreement (1995) as the culminating event, while the Serbian output implicitly frames the arrest and prosecution of Ratko Mladić through the ICTY as the defining endpoint, extending its narrative all the way to 2022. Beyond the divergent answers, the three corpora operate on incompatible temporal scales: the Croatian version spans only six years (1990–1996) and omits Kosovo entirely; the English version covers thirteen years (1991–2004); and the Serbian version runs to twenty six (1991–2017), far exceeding the 1990s frame explicitly set by the question. We acknowledge that the scope of LLM generated responses depends partly on prompt formulation and that the broad temporal range may have resulted from the way the question was phrased. Nevertheless, this remains a relevant finding, as it demonstrates how the model extended the temporal framework of the 1990s to include the later consequences of the conflicts. These divergences (in the identity of the culminating event, in temporal scope, and in geographic coverage) are not complementary perspectives on the same history but incompatible answers to an identical question, constituting a clear case of inconsistency. As for the *Biased Portrayal* category, language-specific framing consistently favours one side's perspective through selective inclusion and selective omission. The Croatian outputs treat Operation Storm (the August 1995 Croatian Army offensive) exclusively as a legitimate restoration of sovereignty, omitting the approximately 200,000 Serb civilians displaced, the estimated 600 or more civilian killings documented by the Croatian Helsinki Committee, and the ICTY prosecutions of Croatian generals for unlawful attacks on civilian populations; this framing mirrors the official Croatian state commemorative narrative. The Serbian outputs present the inverse bias: Vukovar, Ovčara, and Dubrovnik are described as general symbols of wartime horror without attributing agency to the Serbian and JNA forces whose responsibility for those events was established by the ICTY, while Operation Storm is recast as a source of Serbian victimhood. Most starkly,

Srebrenica – a genocide recognised by the ICTY, the ICJ, and the UN General Assembly, in which over 8,000 Bosniak men and boys were killed under Mladić's command – is entirely absent from the Serbian corpus, while the English output treats it as the war's defining atrocity and the Croatian output acknowledges it briefly. In all three cases, the pattern of inclusion and omission tracks the state-level commemorative and media narratives associated with the respective language communities, producing outputs that privilege one national perspective over a balanced historical account.

Finally, narratives are shaped through language. At the same time, differences in representation should not be understood merely as a consequence of each linguistic community interpreting the past from its own perspective. The findings presented above indicate that these differences are systematic. That is, that they arise from distinct politics of memory and narrative frameworks rather than occurring by chance. The results reflect not only different historical viewpoints but also the policies, priorities, and design choices embedded in the systems that produce them.

The event-level analysis lends particular support to this conclusion, as it demonstrates that the observed inconsistencies in the selection of events, temporal and geographical scope, and the attribution of responsibility cannot be explained solely by contextual variation across languages. On the contrary, their recurrence and alignment with broader national and institutional narratives point to systematic, language-conditioned patterns in the construction of historical representations in AI generated content. In this sense, the analysis points to something broader than linguistic variation alone. The way in which the past is presented to users is mediated by algorithmic processes that privilege certain events, actors, and interpretations over others. This raises questions about how historical memory is reproduced and transmitted through artificial intelligence systems, as well as the extent to which these representations reflect broader political, cultural, and institutional narratives rather than neutral historical accounts.

Conclusion

This study offers a critical assessment of how large language models such as ChatGPT engage with historically sensitive content, demonstrating both their potential and their limitations in shaping collective memory in the digital age. By analyzing model-generated responses to complex multinational conflicts across different languages, the research reveals persistent patterns of factual distortion, omission, oversimplification, and ideological bias. These findings reaffirm that AI systems are not neutral observers of the past, but active mediators shaped by the assumptions, priorities, and blind spots embedded in their design and training.

The study also underscores the ethical implications of delegating memory work to algorithmic systems. While AI may expand access to historical knowledge and support the democratization of memory, this potential is undermined when critical human oversight is lacking. Without strong ethical frameworks and digital literacy, the risk increases that collective memory becomes a product of algorithmic mediation: sanitized, depoliticized, or even manipulated, rather than a space for critical reflection and contestation.

Furthermore, the methodology highlights the importance of linguistic and cultural diversity in AI evaluation, demonstrating how language not only mediates meaning but actively shapes historical interpretation. This insight points to the need for multilingual, context-sensitive approaches in future assessments of AI systems, particularly in memory-related domains. Ultimately, the research contributes to broader interdisciplinary debates on AI ethics, historical epistemology, and the digital humanities, particularly in relation to digital memory. It calls for the development of critical infrastructures that preserve the complexity of the past while safeguarding against its algorithmic distortion. In doing so, it opens a path for future studies to further explore how societies might engage with artificial intelligence not as passive consumers, but as active stewards of memory in the digital turn.

Declarations

Competing interests

The author declares no competing interests.

Author contributions

Author contributions have been removed for the purposes of double-anonymous peer review.

Ethics statement

Ethical approval was not required for this study because it did not involve human participants, personal data, or animal subjects. The research analysed responses generated by an artificial intelligence system.

Author Contribution

V.P. edited the manuscript, developed the general overview, and contributed to the theoretical section. J.Ć.N. wrote the introduction and contributed to the theoretical and interpretative framework. I.B.L. contributed to the research section and methodology. N.M. prepared the graphical representations and contributed to the discussion. P.Z. edited the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgement

The role of the experts was limited to the verification of historical accuracy; the analytical categories and interpretation presented in the article were developed by the authors. This work was carried out by Atalia Omer (University of Notre Dame), Peter Cajka (University of Notre Dame), Miguel Cardina (Centro de Estudos Sociais, University of Coimbra, Portugal) and Tomislav Anić (Catholic University of Croatia,

Croatia). Ethical expert verification was conducted within the framework of a broader research project under the supervision of the project leads, J. Ćurković Nimac and Nuno Moniz.

Data Availability

The data that support the findings of this study consist of ChatGPT-generated responses obtained through a structured multilingual prompt protocol. The data consist of ChatGPT4o generated responses for all historical cases and languages examined in project. As the manuscript focuses specifically on the case of the war in the former Yugoslavia, the relevant transcripts in Croatian, Serbian, and English have been uploaded directly to the submission system also stored on Google Drive. The transcripts for the other historical cases and languages are available through the Google Drive link provided in the accompanying excel document.The question protocol used for all cases is available through the same link.

References

1. Aksoy M (2025) Whose morality do they speak? Unraveling cultural bias in multilingual language models. Nat Lang Process J 12:100172
2. Bruns A (2008) Blogs, Wikipedia, Second Life, and beyond: from production to produsage. Peter Lang, New York
3. Buyl M et al (2024) Large language models reflect the ideology of their creators. arXiv
4. Castells M (2000) Uspon umreženog društva, svezak 1. Golden marketing, Zagreb
5. Chun WHK (2008) The enduring ephemeral, or the future is a memory. Crit Inq 35:148–171
6. De Kosnik A (2016) Rogue archives: Digital cultural memory and media fandom. MIT Press, Cambridge
7. Derrida J (1996) Archive fever: a Freudian impression (trans: Prenowitz E). University of Chicago Press, Chicago
8. Floridi L (2014) The 4th revolution: how the infosphere is reshaping human reality. Oxford University Press, Oxford
9. Garde-Hansen J (2011) Media and memory. Edinburgh University, Edinburgh
10. Halbwachs M, ([1926], Ditter FJ, Ditter VY (1950) Harper Colophon Books, London
11. Hammond K (2015) Practical artificial intelligence for dummies. Wiley, Hoboken
12. Handžić Ferović A (2020) Kulturna dimenzija sjećanja na genocid u Srebrenici. Hum: Časopis Filozofskog fakulteta Sveučilišta u Mostaru 15(23):153–168
13. Haralambos M, Holborn M (2002) Sociologija: teme i perspektive. Golden marketing, Zagreb
14. Hoskins A (2018) The restless past: an introduction to digital media and memory. In: Hoskins A (ed) Digital memory studies: media pasts in transition. Routledge, New York, pp 1–24
15. Jenkins K (2008) Promišljanje historije. Srednja Europa, Zagreb

16. Kansteiner W (2022) Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent? History Theory 61(4):119–133
17. Katunarić V (2000) Lica kulture. Naklada Jesenski i Turk, Zagreb
18. Kearns M, Roth A (2020) Ethical algorithm design. ACM SIGecom Exchanges 18(1):31–36
19. Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford
20. Makhortykh M (2021) Memoriae ex machina: how algorithms make us remember and forget. Georget J Int Affairs 22:180–185
21. Mandolessi S (2023) The digital turn in memory studies. Memory Stud 16:1513–1528
22. Manovich L (2001) The language of new media. MIT Press, Cambridge, MA
23. Misztal BA (2021) Teorije društvenog sjećanja (trans: Ćurković J). Kršćanska sadašnjost, Zagreb
24. Novick P (1999) The Holocaust in American life. Houghton Mifflin, Boston
25. Olick JK, Robbins J (1998) Social memory studies: From collective memory to the historical sociology of mnemonic practices. Ann Rev Sociol 24:105–140
26. Rigney A (2016) Cultural memory studies: mediation, narrative, and the aesthetic. In: Tota AL, Hagen T (eds) Routledge international handbook of memory studies. Routledge, Oxon, pp 65–76
27. Schwartz JM, Cook T (2002) Archives, records, and power: the making of modern memory. Arch Sci 2:1–19
28. Shur-Ofry M, Pessach G (2020) Robotic collective memory. Wash Univ Law Rev 97:975–1010
29. Sturken M (2008) Memory, consumerism and media: reflections on the emergence of the field. Memory Stud 1:73–78
30. Toffler A (1980) The third wave. Bantam Books, New York
31. Van Dijck J (2007) Mediated memories in the digital age. Stanford University Press, Stanford

Footnotes

1. The role of the experts was limited to the verification of historical accuracy; the analytical categories and interpretation presented in the article were developed by the authors. This work was carried out by Atalia Omer (University of Notre Dame), Peter Cajka (University of Notre Dame), Miguel Cardina (Centro de Estudos Sociais, University of Coimbra, Portugal) and Tomislav Anić (Catholic University of Croatia, Croatia). Ethical expert verification was conducted within the framework of a broader research project under the supervision of the project leads, J. Ćurković Nimac and Nuno Moniz.

Figures

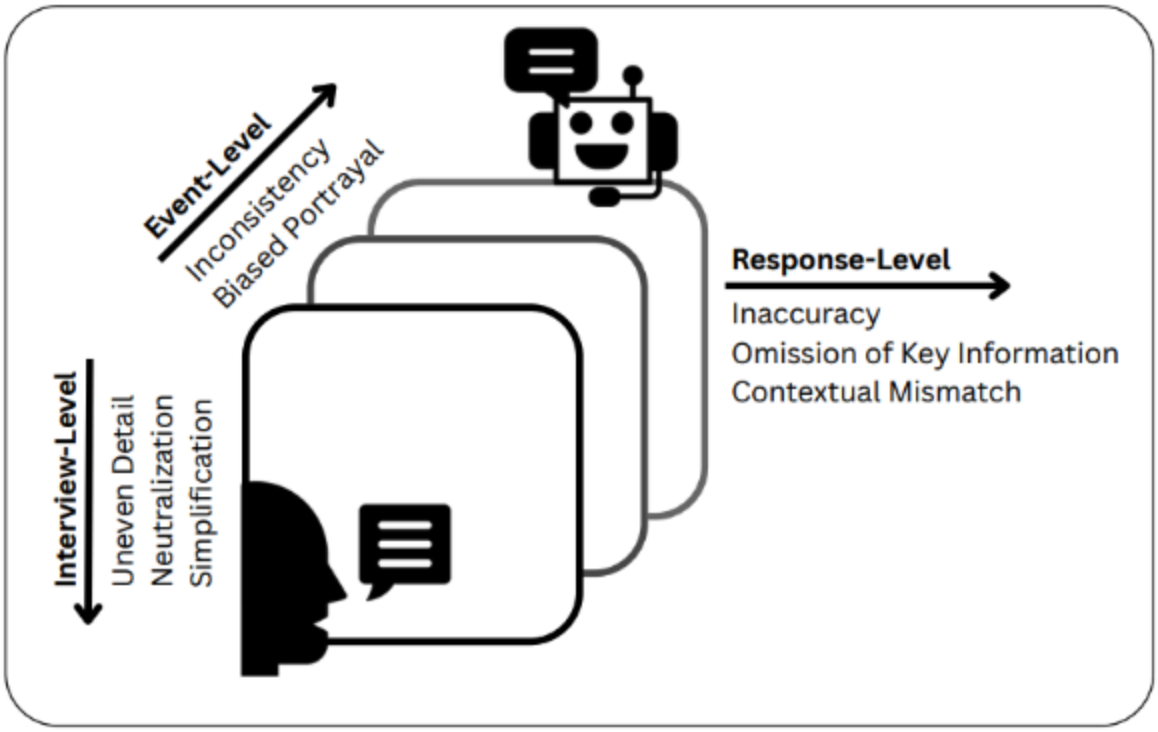


Figure 1

Categorization of Responses

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Datalinkgoogledisk.docx