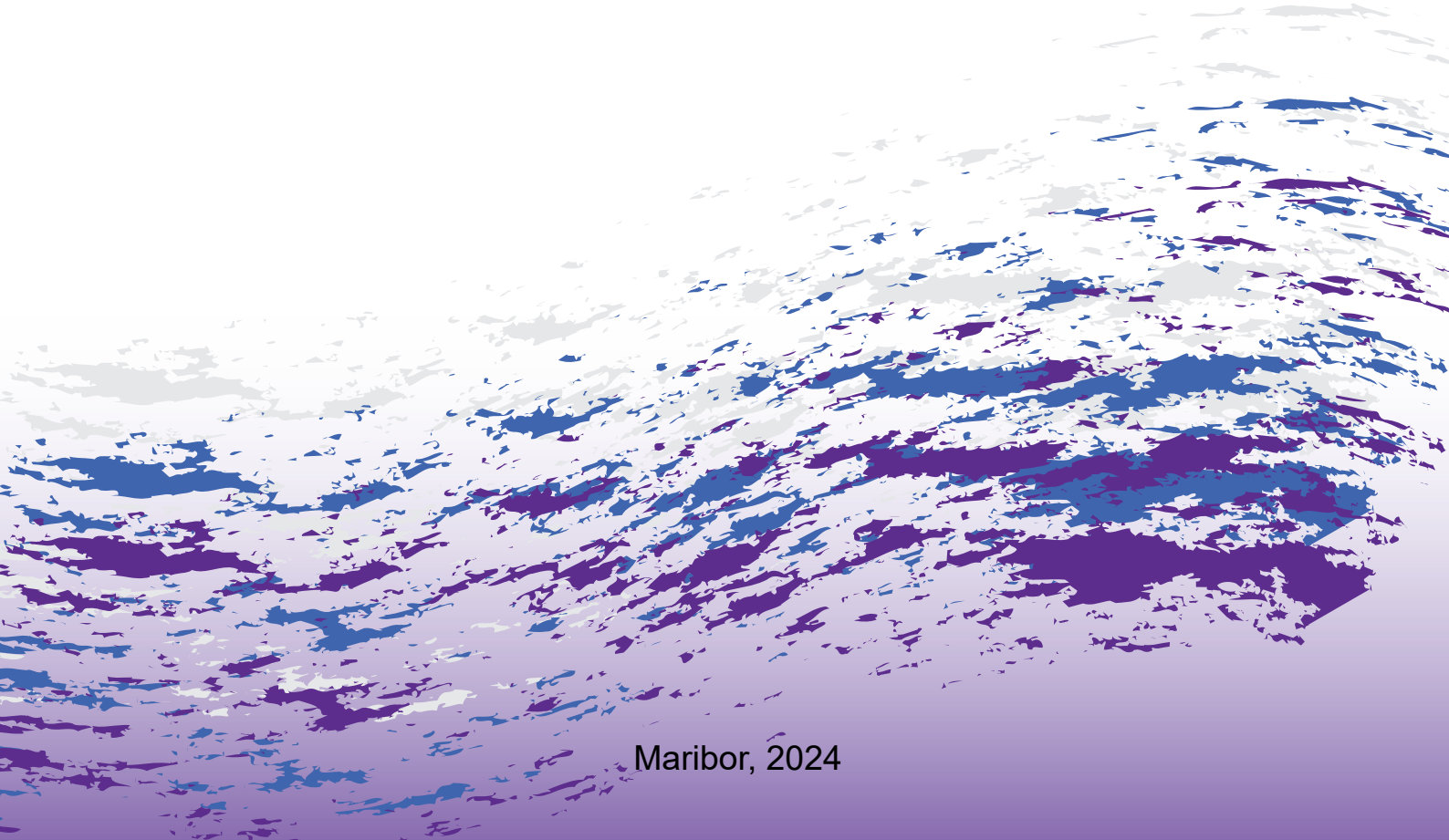
PAM Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

ISSN 2591-0884
https://doi.org/10.54356/MA

Letnik 7 (2024), št. 2 / Year 7 (2024), No. 2



Maribor, 2024

Prejeto / Received: 10. 10. 2024

1.01 Izvirni znanstveni članek 1.01 Scientific article

https://doi.org/10.54356/MA/2024/BIUB3010

EVALUATING CHATBOT ASSISTANCE IN HISTORICAL DOCUMENT ANALYSIS

asist. dr. David HAZEMALI
University of Maribor, Faculty of Arts, Slovenia
david.hazemali@um.si

asist. Janez OSOJNIK
University of Maribor, Faculty of Arts, Slovenia
janez.osojnik1@um.si

izr. prof. dr. Tomaž ONIČ
University of Maribor, Faculty of Arts, Slovenia
tomaz.onic@um.si

asist. dr. Tadej TODOROVIĆ
University of Maribor, Faculty of Arts, Slovenia
tadej.todorovic@um.si

asist. dr. Mladen BOROVIČ
University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia
mladen.borovic@um.si

Abstract:

The article explores the potential of PDFGear Copilot, a chatbot-based PDF editing tool, in assisting with the analysis of historical documents. We evaluated the chatbot's performance on a document relating to the Slovenian War of Independence. We included 25 factual and 5 interpretative questions to address its formal characteristics and content details, assess its capacity for in-depth interpretation and contextualized critical analysis, and evaluate the chatbot's language use and robustness. The chatbot exhibited some ability to answer factual questions, even though its performance varied. It demonstrated proficiency in navigating document structure, named entity recognition, and extracting basic document information. However, performance declined significantly in tasks such as document type identification, content details, and tasks requiring deeper text analysis. For interpretative questions, the chatbot's performance was notably inadequate, failing to link cause-and-effect relationships and provide the depth and nuance required for historical inquiries.

Keywords:

chatbots; historical document analysis; Yugoslav Wars; generative artificial intelligence; large language models

Izvleček:

Vrednotenje pogovornega sistema kot pomočnika pri analizi zgodovinskih dokumentov

Članek preučuje potencial pogovornega sistema PDFGear Copilot kot pomočnika pri analizi zgodovinskih dokumentov. Učinkovitost pogovornega sistema smo preizkusili na dokumentu, povezanim s slovensko osamosvojitveno vojno. Zmožnost poglobljene interpretacije, kritične analize ter jezikovno robustnost pogovornega sistema smo presojali na podlagi njegovih odgovorov na 25 faktografskih in 5 interpretativnih vprašanj, povezanih z lastnostmi in vsebino dokumenta. Pri odgovarjanju na faktografska vprašanja je bil pogovorni sistem delno uspešen. Ni pa imel težav s prepoznavanjem strukture dokumenta, razpoznavanjem imenskih entitet in

pridobivanjem osnovnih informacij o dokumentu. Slabše se je odrezal pri opredelitvi vrste dokumenta, navajanju vsebinskih podrobnosti in pri nalogah, ki zahtevajo poglobljeno analizo besedila. Pri interpretativnih vprašanjih je njegova uspešnost opazno upadla, saj ni bil kos vzročno-posledičnim odnosom ter ni izkazal potrebne poglobljenosti in natančnosti, potrebni za zgodovinopisne raziskave.

Ključne besede:

pogovorni sistemi; analiza zgodovinskih dokumentov; vojne na območju nekdanje Jugoslavije; generativna umetna inteligenca; veliki jezikovni modeli

1 INTRODUCTION[1]

The cornerstone of historical research as well as many other fields in the humanities lies in the meticulous analysis of primary sources, often textual documents, through critical thinking, page-by-page reading, and interpretation. Traditionally, historians relied on physical access to archives for this crucial task, but this has changed considerably in the digital era. Digitisation and the rise of applied computational research methods, including data mining, topic modelling, natural language processing (NLP), and optical character and handwritten text recognition (OCR), have significantly broadened access to historical documents and facilitated the research process. However, while helpful on the one hand, access to large digital datasets now offers a new challenge: how to increase efficiency in analysing the often readily accessible extensive data, which is available in diverse formats, languages and metadata, while ensuring (historical) accuracy. This necessitates new tools and perhaps even changes to the established historical research practices.

Artificial Intelligence (AI) now presents a promising new frontier. Large language model (LLM)-based chatbots, which have begun to permeate nearly every aspect of human activity (see, for example, Ray, 2023; Deng & Lin, 2023; Dwivedi et al., 2023; and Chiarello et al., 2024), raise intriguing questions about their underexplored potential to augment historians' toolkit. However, a critical evaluation of these capabilities and their limitations is necessary to determine their effectiveness as research assistants. By carefully examining both the potential and limitations for historical research, historians may be able to unlock new levels of efficiency and insight (Kansteiner, 2022).

This article presents a proof-of-concept case study, exploring the potential of PDFGear Copilot, a chatbot-based PDF editing tool powered by the GPT-3.5 LLM, to assist in historical research. We investigated its previously untested capabilities in analysing non-English digitised documents. A multimethod approach, combining qualitative and quantitative analysis, was adopted to assess the tool's performance. 30 questions were designed, encompassing both factual inquiries and more demanding interpretative tasks, in an attempt to answer the following research questions:

- To what extent can the chatbot accurately identify, extract, and summarise factual information about the historical document and its contents?
- How successfully can chatbot handle complex historical questions that require interpretation and contextualisation across languages?
- How successfully does the chatbot handle (deliberate) language mistakes (typos, misspellings, and grammatical errors), domain-specific terminology, and complex sentences in Slovenian?

[1] *The authors wish to acknowledge the NOO project "ZELEN.KOM (3330-22-3514 (PP5))", financed by the Ministry of Higher Education, Science and Innovation of the Republic of Slovenia.*

To ensure the robustness of the findings, this experiment was repeated ten times on the same historical document. After analysing the chatbot's answers, a detailed evaluation of its capabilities and limitations was conducted, and its potential for enhancing research efficiency highlighted. This proof-of-concept case study contributes to the understanding of the role of chatbots in historical research, particularly in the context of analysing digitised non-English historical documents. Making the chatbot a potential new methodological tool for historians is central to this innovative approach.

2 RELATED WORK

Modern AI chatbots, ChatGPT in particular, have already been recognised as capable of assisting historians with the often-time-consuming tasks of scientific writing, such as helping organise materials, conducting comprehensive literature reviews, managing citations effectively, proofreading or even drafting and generating passages, paragraphs, summaries, reports etc. (Stokel-Walker, 2023; Else, 2023; and Altmaë, Sola-Leyva & Salumets, 2023). However, continuous active testing within the broader scientific community reveals a crucial caveat: while chatbots are demonstrably able to execute these tasks, the outputs often contain a concerning mix of true and entirely fabricated information or "hallucinations" (Buholayka, Zouabi & Tadinada, 2023; and Salvagno, Taccone & Gerli, 2023). Lozić and Štular (2023) explored AI chatbots' capabilities in scientific writing specifically for the humanities disciplines. Their study employed an interdisciplinary approach, testing six chatbots (including ChatGPT) with two prompts requiring 500-word complex historical analyses. Human experts evaluated the generated content for factual accuracy ("quantitative accuracy") and scientific contribution ("qualitative precision"). While some chatbots achieved near-factual accuracy, all struggled to generate original scientific insights (Tirado-Olivares et al., 2023; and Carretero & Gartner, 2024).

LLMs and chatbots present exciting possibilities for data curation and archival work. Historians can use this technology to expedite the digitisation of vast collections of historical documents, perform automatic indexing and classification for improved searchability, and potentially assist archivists with data curation tasks (Spina, 2023; and Lorenz & Konečný, 2023). The impact of AI extends beyond text-based sources. Oral historical archives, once time-consuming to analyse, could become more accessible with AI technology that captures natural language (Pessanha & Salah, 2021). This is particularly relevant for family historians, who often rely on in-person interviews with elderly individuals, possibly facing geographical or technical limitations. Studies suggest that chatbots designed to gather family histories show promise in overcoming these challenges (Drumm ad Tran, 2023). Museums can also utilise AI to create innovative and engaging experiences for visitors (Varitimiadis et al., 2021; and Trichopoulos et al., 2023). AI-powered virtual personas and voice assistants could act as personalised guides, providing historical context and insights tailored to individual interests, and even revolutionise how we approach "communicative remembering" in memory studies (Pentzold, Lohmeier, & Birkner, 2023; Henrickson, 2023; and Sisto, 2020).

Historical research could gain significant advantages from LLMs and chatbots, particularly in analysing historical documents. These often suffer from errors in transcriptions (frequently owing to OCR) that hinders their usability. Researchers are exploring the potential of LLMs to address this challenge. One study evaluated fourteen LLMs on correcting errors across various historical texts (Boros et al., 2024). The results suggest LLMs are currently not ideal for this task, while a different study, focusing on historical newspapers, achieved promising results (Thomas, Gaizauskas, & Lu, 2021), which could prove useful for studies like the one by Mezeg & Žigon (2023) that uses a

manual approach for analysing historical newspapers, or that by Vuković Vojnovič (2023) conducting manual research into a non-historical corpus. Historians can utilise LLMs and chatbots to analyse single digitised documents or massive datasets for information extraction, while simultaneously uncovering patterns and trends that might be difficult or time-consuming to identify with established (often non-digital) methods or due to language barriers (Lozić & Šular, 2023; Gonzalez Garcia & Weilbach, 2023; and Siu, 2023). In a recent study, Gonzalez Garcia and Weilbach explored how different LLMs affect the performance of chatbots for assistance in historical research. In the scope of the study, an open-source personalised chatbot named KleioGPT was developed and integrated for this purpose with various LLMs. The results showed that ChatGPT outperformed KleioGPT in question-answering tasks, achieving 92.5% accuracy compared to KleioGPT's best performance of 87.5%. Let us say that KleioGPT was using only the data from a corpus of 86 (mostly) peer-reviewed books. This makes it rather specialised for querying information from that corpus. ChatGPT, however, only answered the questions without being fed data from the corpus. The authors concluded that while custom chatbot solutions offer advantages in privacy and customisation for academic research, commercial chatbots currently perform better. AI can thus accelerate the research process by helping historians sift through vast amounts of data to extract relevant information (Gonzalez Garcia & Weilbach, 2023).

The exciting possibilities of AI tools that historians could extensively benefit from in historical document analysis venture beyond mere information extraction towards interpreting the documents, contextualising the contents, identifying linguistic and semantic patterns, and validating historical theories. At this point, however, these remain only aspirations, since the AI tools lack reliability as they struggle to differentiate truth from falsehood and the ability to "reason" (Kansteiner, 2022).

According to Ali et al. (2024), these challenges can be grouped into several areas: user limitations, technological limitations, operational limitations, ethical limitations, and environmental limitations. User limitations include difficulties with complex queries, limited contextual understanding, and a lack of creative response generation. Chatbots are often heavily reliant on training data, which can restrict their ability to engage in open-ended exploration or to provide truly original insights. Additionally, some users might perceive machine interaction as impersonal and will thus miss the special touch of human conversation (Currie, 2023; Floridi, 2023; Stevenson et al., 2022; Diederich et al., 2022; and Kasneci et al., 2023). On the technological side, data privacy and security are key concerns when dealing with historical documents, as some data might be sensitive. Ensuring the security of such data and protecting user privacy while utilising AI tools requires careful consideration (Kasneci et al., 2023; and Dwivedi et al, 2023). Operationally, personalising chatbot interactions to individual historians' research needs can be complex. Furthermore, training and maintaining these advanced language models can be expensive, requiring significant computational resources (Kasneci et al., 2023). Ethically, the potential for bias in training data, used for chatbots, is a critical concern. It is crucial to ensure that historical documents, used to train these models, represent a diverse range of perspectives to avoid perpetuating historical biases in the AI's responses (Ali et al., 2024; Makhortykh et al., 2023; and Abadie, Chowdhury, & Mangla, 2024). Finally, the environmental impact of training and running large language models is an ongoing challenge, as they require significant energy consumption. In his influential study, Kansteiner lists several key limitations in using ChatGPT, specifically in historical research. Firstly, the lack of source attribution due to its vast and opaque body of training data raises concerns about the verifiability and reliability of historical information it provides. Secondly, ChatGPT operates on the principle of "sequential plausibility," meaning it prioritises generating grammatically correct and natural-sounding responses over factual accuracy. Finally, its filters, designed to prevent the spread of

misinformation, might inadvertently limit historians' ability to explore and discuss controversial topics objectively (Kansteiner, 2022).

This section reveals that the application of modern chatbots for historical document analysis is still in its rudimentary stages. Furthermore, there is a scarcity of research on their effectiveness as research assistants, particularly in the document analysis phase. Our analysis attempts to bridge this gap.

3 METHODOLOGY

This section introduces the selected source document, its analysis, reference literature, the selected chatbot and software, and the experiment setup details.

3.1 The established methods for historical document analysis

To assess chatbot potential for historical research, this section first outlines the established approach to analysing historical documents as it is currently performed by historians. This approach is a multi-layered, non-linear process that varies depending on specific research questions, document type, and the individual historian's *modus operandi*. It requires critical thinking, meticulous attention to detail, and deep understanding of the historical context to allow the historian to accurately navigate the nuances of the source and draw meaningful conclusions.

The process typically begins with the historian's careful examination of the text, minding the language, author's perspective, and historical context, identifying key themes, arguments, and potential biases. Simultaneously, the source's authenticity, reliability, and origin are critically evaluated by examining the author's credentials, potential biases, and the historical context of the text's creation, as well as its physical condition and alterations. The historian then contextualises the document by analysing the relevant period, events, and social, political and cultural factors that may have influenced the author and the content. They compare the information from the analysed document with other archival records and other types of sources (such as published research, newspaper articles, diaries, memoirs, and interviews) to verify or challenge the information presented through triangulation. Finally, a historian interprets the document's significance in light of research questions and historical knowledge, navigating a delicate balance between striving for objectivity and factual accuracy, while acknowledging the inherent subjectivity and context-dependence of historical understanding.

3.2 Source document selected for the experiment

In this experiment, we analysed the "Unauthorised Magnetogram of the Conversation between the President of the Presidency of the Republic of Slovenia Milan Kučan and the President of the Republic of Croatia Dr. Franjo Tuđman with the Delegation of the European Community, Zagreb, 28/29 June 1991 from 23.30 to 01.15" [original title: Neavtoriziran magnetogram razgovora predsednika predsedstva Republike Slovenije Milana Kučana in predsednika Republike Hrvaške dr. Franja Tuđmana z delegacijo Evropske skupnosti, Zagreb 28./29. 6. 1991 od 23,30 do 01,15] (Repe, 2004, 108–117). This published document is a primary source, a record of a diplomatic exchange between Slovenian and Croatian officials and members of the European Economic Community (EEC) delegation (Troika) in Zagreb in 1991 in the wake of the Slovenian War of Independence, also referred to as The Ten-Day War, the first of the ensuing Yugoslav wars that followed the disintegration of Yugoslavia. Despite Slovenian

and Croatian plebiscites for independence (Osojnik, 2022, 2023) and their request for EEC recognition, the EEC had agreed to delay the recognition of their independence. An additional reason for selecting this document for the experiment is that three decades have passed since the bloody conflict in the heart of the Balkans (Pirjevec, 1995; Repe, 2002; Pesek, 2007; Hribernik, 2018; Zajc, 2020; and Repe, 2022). What is more, the document has been used in academic research (e.g., Repe, 2002; Kádár et al., 2024).

The 10-page document originates from the last part of a three-volume collection of published sources *Sources on Democratization and Independence of Slovenia* (original title: *Viri o demokratizaciji in osamosvojitvi Slovenije*), edited by the eminent Slovenian historian Božo Repe (Repe, 2001, 2002a, 2002b, 2003, 2004, 2015, 2022; and Repe & Kerec, 2017). Each paragraph consists of a single speaker's contribution. It is worth noting that the record contains ambiguous parts and lacks perfect coherence or editing. The document is predominantly in Slovene, with some sections in Croatian and Serbo-Croatian. Additionally, four footnotes are included.

3.3 Reference literature

To establish ground truth answers for evaluation, we consulted relevant scholarly literature; primarily Repe’s series of published sources. This volume also features introductions by Repe to each thematic section that contextualise the included documents. In the Zagreb talks, Repe's preface (2004, 89-91) proved to be an invaluable reference point for our ground truth answers.

3.4 Chatbots and selected software

Chatbots are software applications that simulate human conversation through text or voice. They use natural language processing to understand and respond to user inputs, making the interaction conversational. Since 2023, the wide adoption of chatbots such as ChatGPT, Gemini, and Claude has brought this way of interaction closer to the public. Currently, chatbots are used in various fields, where they are proving to be a useful addition to the work process. Modern chatbots use a variety of components in their workflow, with LLMs being the key component for generating responses. Figure 1 shows some of the possible components within the workflow of a modern chatbot. It is worth mentioning that not all components are included in every modern chatbot. This depends on the implementation details of a specific chatbot, which are quite vague and unknown for the popular commercial chatbots such as ChatGPT, Gemini, and Claude. As an alternative, open source chatbots exist, but their workflow is simpler with most of them simply using the large language model component and nothing else. Consequently, open source chatbots seem to perform worse than their commercial counterparts.

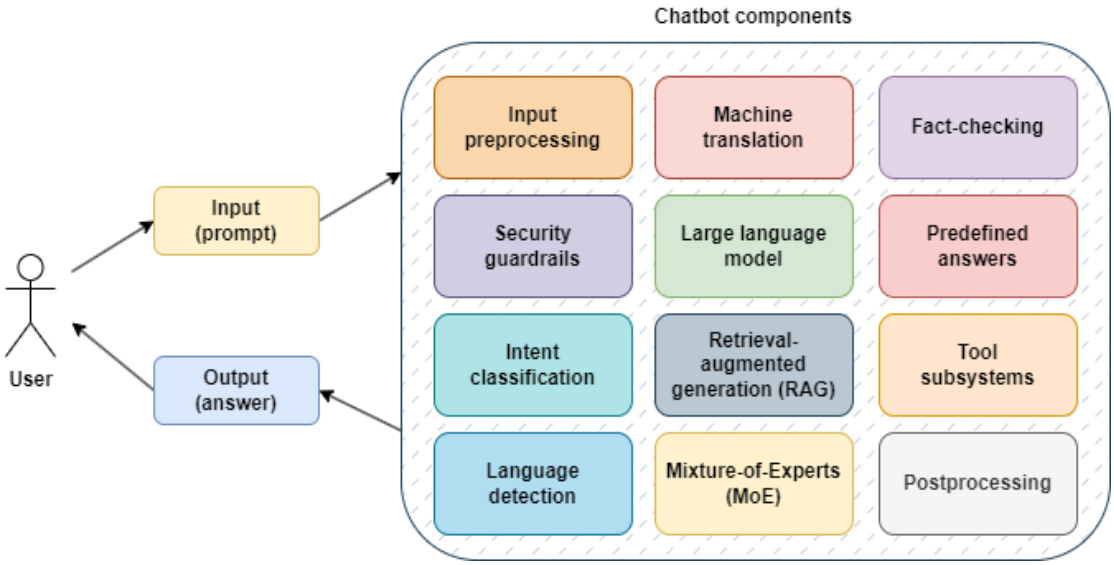


Figure 1: Various components in a general modern chatbot architecture.

For the experiment, we chose PDFGear Copilot, which is an AI-powered chatbot built into a PDF editor. It uses natural language processing to understand requests and can perform various tasks on PDF documents. The spotlight functionality of this software is the ability to interact with PDF documents in natural language. Instead of reading through all the text in the document, users can ask a question in the user interface and the software will generate an answer. The user experience is similar to that in popular chatbots such as ChatGPT, Gemini, and Claude. The choice of software was based on this functionality, since we wanted to test the usability and effectiveness for the established process a historian performs during document analysis of this type. This software uses the OpenAI's GPT-3.5 large language model. The software is free and can be installed locally, which makes it easily accessible. Additionally, one study (Gonzalez Garcia & Weilbach, 2023) showed promising results when using commercial chatbot solutions. These were the main reasons for selecting it in our experiment setting.

3.5 Experimental setup and details

The controlled experiment took place in a designated room with a laptop computer. The experiment supervisor prepared the room from the technical point of view and pre-prompted the chatbot, as shown in Figure 2. A separate group of 4 experts (historians and linguists) studied the reference document and formulated lists of 25 factual and 5 interpretative questions that simulate the established process of document analysis used by historians. Every experiment participant (10 academics from the Faculty of Arts, University of Maribor) who tested the software was alone in the room with an internet connection. They used the lists of questions prepared by the expert group and prompted the chatbot with them. Zero-shot prompting was used, which means that the prompts did not include any examples of expected answers. The questions, which were of different taxonomic levels, were asked in the order shown in the Appendix (for the readers' convenience, the questions were translated into English). All questions were related exclusively to the source document and its content. They were asked in Slovenian. After collecting the responses to their questions, the evaluators (2 historians with expertise in the topic and 2 linguists) analysed and evaluated the responses referring to the predefined ground truth answers. The document's historical context and potential

interpretations were derived from the relevant peer-reviewed scholarly literature, in our case from Repe's accompanying text on the Zagreb talks (see subsection 3.3. Reference literature). For factual questions, the answers were either readily extracted or derived directly from the document content.

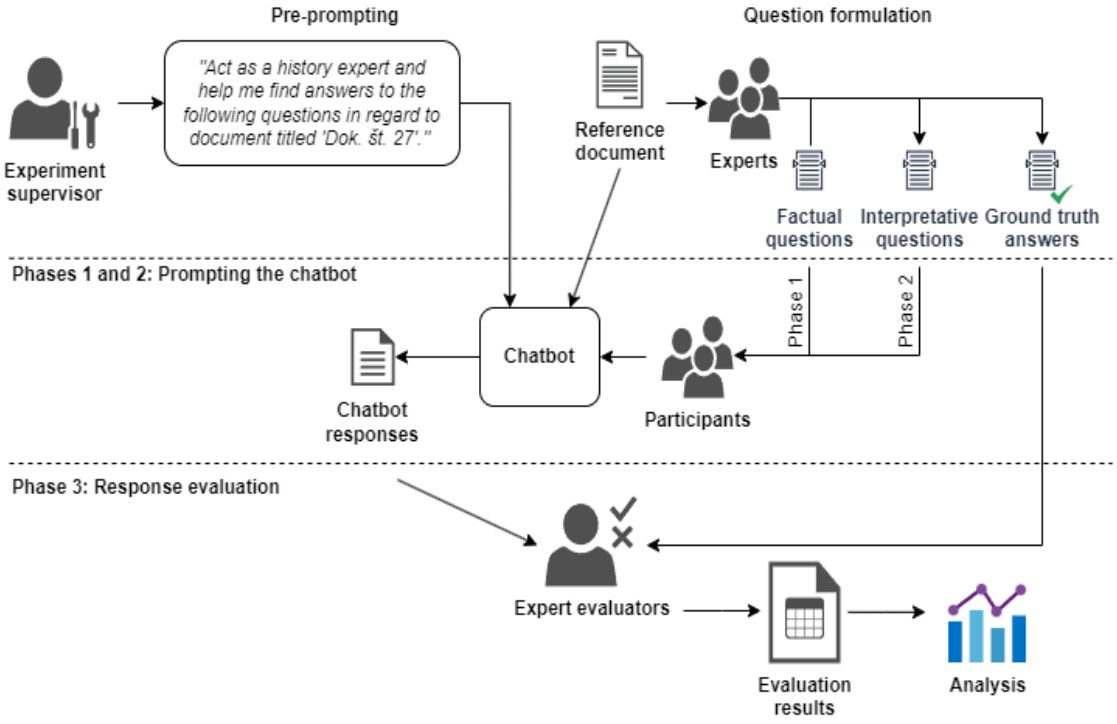


Figure 2: Design of the experiment with phases, activities and involved parties.

A) Factual Questions (FQ)

To assess the chatbot's performance on factual tasks, we designed a set of 25 questions of varying difficulty levels across five categories: document characteristics, basic meeting details, participants and roles, specific terms and mentions, and language robustness (further details in the Appendix). Notably, a small number of questions overlapped between the first four categories and the language robustness category.

Category 1: Document Characteristics (FQs 1-7)

These questions focus on basic information about the document and its (formal) characteristics. This includes its title, type, date of creation, author, number of pages, character count, and number of footnotes. Assessing the chatbot's ability to answer these questions provides an insight into its proficiency at extracting basic metadata and navigating the structure of the document.

Category 2: Basic Meeting Details (FQs 8-10)

These questions address the details about the meeting described in the document: its duration, location, and the number of active participants. Evaluating the chatbot's performance in this category helps us understand its ability to extract specific and/or accurate details related to the event recounted in the document.

Category 3: Participants and Roles (FQs 11-13, 18)

This category focuses on identifying participants of the meeting and their roles. Questions refer to the (full) names of individuals and their positions at the time of the meeting (such as president, foreign minister). Analysing a chatbot's effectiveness in answering these questions sheds light on its ability to recognise certain named entities and the relationships among them.

Category 4: Specific Terms and Mentions (FQs 14-17, 19, 21-24)

These questions test the chatbot's ability to identify and locate specific terms, names, dates, and concepts within the document. Evaluating performance in this category reveals the chatbot's skill retrieving facts and “comprehending” specific details in the text, as well as potentially attempting some level of content analysis.

Category 5: Languages and Language Robustness (FQs 20 and 25)

This category addresses the issue of language robustness, that is, the chatbot’s ability to provide meaningful and correct answers despite (intentional) typos and grammar mistakes, its ability to recognise data in various forms (such as month written with the number vs. the name of the month), and its ability to distinguish between passages in languages other than English. FQs 20 and 25 were designed to test this particular issue, while language related issues were also monitored in all other questions, particularly questions 17, 19, and 21.

B) Interpretative Questions (IQ)

In addition to factual questions, the chatbot was asked 5 longer interpretative questions. The aim of the first question (IQ 1) was to determine if the chatbot is able to provide a more extended, substantial answer located within a single paragraph of the document. In IQ 2, we first tested whether the chatbot is able to identify the correct speaker in a meeting based on the definition of their role, and consequently analyse only one paragraph of a document. Next, we tested whether the chatbot was able to extract the main idea from a slightly extended exposition. If the answers to the first two questions were only in one paragraph of the document, the third question (IQ 3) tested whether the chatbot was able to provide a longer, more substantial interpretative answer contained in several paragraphs. IQ 4 was more complex and combined aspects of the second and third questions. On the one hand, we wanted to test if the chatbot was able to give an answer that was contained in several paragraphs and/or paragraph passages in a document – the answer was given by several speakers –, on the other hand, to determine whether the chatbot was able to form a concise answer. The fifth question (IQ 5), which was the most complex of the interpretative questions, was aimed at testing whether the chatbot was able to take into account the whole document and make a value judgment, while constructing arguments for the reasoning on which the judgment was based.

The chatbot's answers were evaluated independently by the two expert evaluators (historians), and then compared. Since interpretations in the historiography – as well as in the humanities in general – are not unambiguous, their evaluations occasionally differed, e.g., one found the chatbot's answer acceptable and the other unacceptable. Therefore, this rating scale was set for evaluating the questions: a) the chatbot's answer was considered acceptable if it was defined as acceptable by both expert evaluators; b) the chatbot's answer was considered unacceptable if both expert evaluators defined it as unacceptable; c) any chatbot's answer that was not defined by both evaluators as either acceptable or unacceptable was noted as conditionally acceptable.

4 RESULTS

In this section, we report the results of our experiment. The interpretative questions are presented separately from the factual ones, while each of the two clusters is presented in a descending order – starting with those where the chatbot performed well and progressing to those where it was less successful.

4.1 Factual Questions

4.1.1 Category 1: Document Characteristics (FQs 1-7)

The chatbot achieved an average accuracy of 41.4% (29 correct answers out of 70) for the 7 document characteristics questions in 10 experiment runs. Quantitatively analysing the chatbot's performance on a question-by-question basis revealed a mixed picture. The chatbot excelled at recognising the absence of author information (FQ 4), achieving a perfect accuracy score of 100%. It achieved high accuracy (80%) in determining the number of pages (FQ 5) and moderate-to-high accuracy in identifying the document's creation date (70%) (FQ 3). For the remaining questions in this category, the chatbot exhibited low accuracy. It achieved an accuracy of 40% in identifying the document type (FQ 2). It failed (0% accuracy) to distinguish between the document's and the volume's titles (FQ 1), as well as in providing the character count (FQ 6) and the number of footnotes (FQ 7).

Table 1: Results limited to factual questions in Category 1. Acceptable answers are marked with ✓, unacceptable answers are marked with ✗.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Acc
FQ1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ2	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	40%
FQ3	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	70%
FQ4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%
FQ5	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	80%
FQ6	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ7	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%

4.1.2 Category 2 Basic Meeting Details (FQs 8-10)

The chatbot achieved an average accuracy of 26.7% (8 correct answers out of 30) for the 3 questions about the meeting details in 10 experiment runs. The chatbot performed poorly in answering questions related to basic meeting details. It was moderately successful (50% accuracy) in determining the number of participants who actively spoke at the meeting (FQ 10). It demonstrated some ability to identify the meeting location (FQ 9) with 30% accuracy (3 correct answers out of 10 experiment runs). However, the greatest challenge for the chatbot in this category seemed to be the extraction of information about the meeting duration (FQ 8). It failed to answer this question correctly in all attempts (0% accuracy).

Table 2: Results limited to factual questions in Category 2. Acceptable answers are marked with ✓, unacceptable answers are marked with ✗.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Acc
FQ8	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ9	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	30%
FQ10	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	50%

4.1.3 Category 3: Participants and Roles (FQs 11-13, 18)

In this category, the chatbot achieved an impressive 72.5% accuracy in 10 experiment runs (29 correct answers out of 40), demonstrating strong ability to identify key individuals and their roles in the meeting. The chatbot achieved a perfect score (100% accuracy) in identifying the Slovenian President of the presidency (FQ 11). It demonstrated a similarly strong ability (100% accuracy) to recognise and link Franjo Tuđman with his role (FQ 18), even when the spelling of his name differed from the correct one (Tudjman, Tudjaman). The chatbot's accuracy dropped to 70% in identifying the Dutch Foreign Minister (Hans) van den Broek from a footnote reference (FQ 13). The most significant hurdle in this category was the identification of members of the European Ministerial Troika, a specific group of representatives (FQ 12). Here, the chatbot only achieved 20% accuracy.

Table 3: Results limited to factual questions in Category 3. Acceptable answers are marked with ✓, unacceptable answers are marked with ✗.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Acc
FQ11	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%
FQ12	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	20%
FQ13	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	70%
FQ18	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%

4.1.4 Category 4: Specific Terms and Mentions (FQs 14-17, 19, 21-24)

The chatbot achieved an average accuracy of 27.7% (25 correct answers out of 90) for the 9 questions on specific terms and mentions in 10 experiment runs. It achieved high accuracy of 90% in identifying the page number of the location of an acronym (FQ 14). Similarly high accuracy (80%) was shown in identifying (the absence of) "ethnic conflicts" in FQ 23. The accuracy was moderate (60%) in FQ 19, which asked to distinguish a specific subgroup (voters for independence) from a larger group (all participants) within a complex sentence lacking an explicit comparison. The chatbot demonstrated some ability to handle complex information extraction tasks, particularly those involving multiple categories or groups of people within a sentence. It demonstrated difficulties in identifying full names for missing acronyms. It achieved low accuracy (20%) in FQ 22, where it had to link a full name ("Jugoslovanska ljudska armada") to its acronym ("JLA"), which is not explicitly mentioned in the text. Counting tasks also proved challenging. The chatbot failed (0% accuracy) in both attempts to determine the frequency of specific last names (FQs 15 and 16). Processing complex sentence structures proved similarly challenging. The chatbot was unable (0% accuracy) to identify the language spoken by a participant based on context clues (FQ 17). The chatbot's ability to recognise information across languages was equally limited; in FQ 21, it failed (0% accuracy) to identify a date ("17. 8.") written in Croatian ("17. kolovoza"). Finally, the chatbot proved incapable of tackling the most complex question in this category by failing to identify an ambassador ("Zimmerman") whose name was spelled

with a single instead of a double "m" and when it appeared at a line break, with the first part ("Zimer-") in one line and the last part ("manu") in the following one, when the question used a synonym for "ambassador."

Table 4: Results limited to factual questions in Category 4. Acceptable answers are marked with ✓, unacceptable answers are marked with ✗.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Acc
FQ14	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	90%
FQ15	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ16	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ17	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ19	✗	✗	✓	✗	✓	✓	✓	✗	✓	✓	60%
FQ21	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ22	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	20%
FQ23	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	80%
FQ24	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%

4.1.5 Category 5: Language robustness: resistance to intentional typos, grammar mistakes, symbols and language switching (FQs 20 & 25)

Table 5: Results limited to language robustness questions in Category 5. Acceptable answers are marked with ✓, unacceptable answers are marked with ✗.

FQ20	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
FQ25	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	20%

The chatbot performed poorly on language robustness questions. With only 20% accuracy in 10 experiment runs, it managed 2 correct answers in FQ 25 (about how many languages the document contains) and none in FQ 20 (the chatbot's ability to link the Slovenian synonyms "referendum" and "plebiscit"). The chatbot's other answers suggest it also made occasional language and grammar mistakes in Slovenian.

4.2 Interpretative Questions

Table 6: Cumulative results for interpretative questions. Unacceptable answers are marked with ✗, conditionally acceptable answers are marked with ~.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Acc
IQ1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
IQ2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
IQ3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
IQ4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0%
IQ5	~	✗	~	~	✗	✗	✗	~	~	~	60%*

The chatbot performed poorly on the interpretive questions, as seen in Table 6. With only 12% accuracy, it managed 6 conditionally acceptable answers out of 50 for the 5 interpretative questions in 10 experiment runs. Surprisingly, the chatbot's performance peaked in IQ 5, the broadest taxonomic category in our test. We assessed its ability to make value judgments and construct arguments for the reasoning on which the judgment is based. It managed 6 conditionally acceptable (60% accuracy) answers, which is why accuracy is marked with an asterisk (*) in Table 6. In these instances, the chatbot

correctly concluded that a definitive, unambiguous judgment of the meeting's success (see Appendix) cannot be made solely on the basis of this document. However, as with more demanding factual questions, the reasoning behind these value judgments remained unsatisfactory. It primarily reiterated irrelevant information from the document, failing to provide a clear justification. The chatbot was unsuccessful in IQs 1-4 (0% accuracy). It was unable to provide concise or extended, substantial answers from either a single paragraph or when referencing multiple paragraphs and/or paragraph passages.

5 DISCUSSION

This section discusses the results, separately examining the implications for factual and interpretative questions. Section 5 is structured parallelly to Section 4 (Results).

5.1 Factual Questions

5.1.1 Category 1: Document Characteristics (FQs 1-7)

The chatbot's performance on document characteristics tasks was uneven. Its highest accuracy was achieved in FQ 4, determining that, in all 10 experiment runs (100% accuracy), the document lacks information about its author. This suggests the chatbot is able to navigate the document structure, recognise missing data and the specific named entity. It correctly identified the number of pages (FQ 5) in 8 out of 10 experiment runs (80% accuracy). Together with similar accuracy (70%) in determining the date of the document creation (FQ 3), these two results indicate moderate-to-high capability in identifying basic (meta)data.

Low accuracy in other questions in this category reveals the chatbot's significant limitations. With only 4 out of 10 correct answers, its ability to identify the document type (FQ 2) was inconsistent. The failure to distinguish between the document's title and the title of the book (FQ 1) highlights a potential limitation in comprehending basic document structure. Similarly, the challenges to determine character count (FQ 6) and number of footnotes (FQ 7) show the chatbot's complete inability to recognise and count these elements, possibly due to document format, question complexity, or limitations in the chatbot's natural language processing abilities.

While the chatbot excelled at identifying missing information and performed adequately for basic details like page count and creation date, it struggled with more nuanced tasks like document type identification. It completely failed to distinguish between document and volume titles, as well as to count the characters or footnotes. These findings suggest that while chatbots may offer some assistance in retrieving basic document characteristics, their limitations in handling even moderately complex metadata tasks call for caution.

5.1.2 Category 2 Basic Meeting Details (FQs 8-10)

The chatbot's performance in Category 2, which dealt with basic meeting details, was poor. Its ability to determine the number of active participants with 50% accuracy (FQ 10) suggests some capability in understanding conversational clues. In this document, participant names are explicitly mentioned before their contributions (dialogue). The chatbot's moderate accuracy could be a result of its training on data sets that include features like speaker identification or dialogue segmentation.

The chatbot's 30% accuracy in identifying meeting location (FQ 9) suggests limitations in its ability to consistently recognise location information within the text. In several experiment runs, it incorrectly claimed the information was missing, even when the latter was clearly stated in the document. For example, in one of the experiment runs, the chatbot responded that the length of the meeting cannot be determined, since this information is missing, while the location was explicitly mentioned ("Zagreb") seven times on the first page of the document (Repe, 2004, 108). This points to potential limitations in recognising location-related information, even in clear and repetitive formats. While a variety of ways to express location (proper nouns, descriptive phrases etc.) posed some challenges, the repeated failures in this instance suggest the need for further training on meeting-specific contexts. Another experiment run indicated a potential issue with "hallucinations", where the chatbot refers to the document as a "Confidentiality Statement" despite no such statement (nor words like "confidentiality") exist in the document. The same can be said about another experiment run, where the chatbot discussed "thematically mismatched PDF pages".

The failure (0% accuracy) to answer questions regarding meeting duration (FQ 8) exposes a significant challenge in extracting specific time-related details from documents. This limitation could stem from difficulties in recognising expressions of time (such as "2 hours" or "10:00 AM"). Even when presented with clear start and end times – as in the case of our document: "23.30 do 01.15" (Repe, 2004, 108) –, the chatbot incorrectly claimed in one of the experiment runs that the information was missing. This inconsistency suggests limitations beyond mere recognition of time formats.

The chatbot's performance on questions addressing basic meeting details was substandard. While it achieved partial success in identifying the number of active participants, its accuracy in answering questions referring to both meeting location and particularly meeting duration was unsatisfactory.

5.1.3 Category 3: Participants and Roles (FQs 11-13, 18)

The chatbot was more successful in this Category, tackling questions about the participants and their roles in the meeting. It consistently identified the Slovenian President of the Presidency with perfect accuracy (FQ 11). This suggests a strong ability to recognise and extract titles, government positions, and specific named entities. The training data likely included a significant amount of information about government structures and titles, allowing the chatbot to effectively connect "predsednik predsedstva Republike Slovenije" (President of the Presidency of the Republic of Slovenia) with the head of state. Furthermore, the chatbot demonstrated proficiency in handling name variations, achieving 100% accuracy in linking "Franjo Tuđman" with his role as President of the Republic of Croatia (FQ 18) despite various inconsistencies in name spelling throughout the document. This highlights its ability to overcome a common challenge in natural language processing.

The chatbot's proficiency in this category also extends to footnotes. It exhibited a moderate-to-high accuracy (70%) in identifying the Dutch Foreign Minister (FQ 13), which indicates some ability in recognising information about nationalities, government positions, and named entities. An even more impressive finding is that in 7 successful experiment runs it used the accurate spelling "van den Broek", which appears in the footnote (Repe, 2004, 108) together with his role/position (and not the misspelt version "van den Brook", used in the main body of the text (Repe, 2004, 108 and 117) together with his role/position). The chatbot's relative success with this question suggests that it is capable of processing footnotes but not counting them. Moreover, two observations were noted here. First, in one experiment run, the chatbot listed the Dutch Foreign

Minister twice when naming the members of the "Troika" (FQ 12) but failed to do so in FQ 13. This suggests limitations in the chatbot's ability to maintain and utilise information across tasks within a single experiment run. Second, in a FQ 13 run referring to the Dutch Foreign Minister, the chatbot responded ambiguously. It claimed it was unable to return his name owing to missing information about the date, despite having access to the list of participants, which it explicitly stated.

In its attempt to identify members of the European Ministerial Troika (FQ 12), even though this information is not explicitly listed in the document, the chatbot performed poorly (20% accuracy). In one experiment run, the chatbot listed Dutch Foreign Minister van den Broek twice as a member of the Troika, not realising the absurdity of this claim, since the "Troika" consisted of three different members. In several other experiment runs, the chatbot listed various false compositions, which suggests that it struggles with tasks that require combining numerical information with specific concepts.

Ultimately, the chatbot's performance in Category 3 showcased its strengths in named-entity recognition, handling variations in named-entity spellings, and to some extent, extracting and processing information from footnotes. However, it faced challenges in FQs 12 and 13, suggesting limitations in identifying group members.

5.1.4 Category 4: Specific Terms and Mentions (FQs 14-17, 19, 21-24)

The chatbot's performance in locating, identifying, and extracting specific terms within the document was poor. High accuracy (90%) was displayed in identifying the page number of the acronym "KEVS" (Question 14). This suggests a strong capability for extracting specific factual terms within the text.

To evaluate the chatbot's ability to handle spelling errors with distinct terminological meanings, we deliberately introduced such a misspelling in FQ 23, mimicking real-world scenarios where researchers may not always type perfectly. We replaced the word "ethnic" [in Slovenian "etnični"] (conflicts) (Repe, 2004, 110) with "ethical" [in Slovenian "etični"] (conflicts) to observe if it would recognise the error. While the chatbot successfully (80% accuracy) identified the absence of "ethical conflicts," it failed to recognise and potentially correct the misspelling, in contrast to its success in the "Brook/Broek" instance in FQ 17. This indicates that while the chatbot can handle some imperfections in user input, it struggles with misspellings that alter meaning.

The chatbot showed some ability to handle complex information extraction tasks. It achieved a moderate accuracy of 60% in determining the percentage of plebiscite participants who voted "for an independent and sovereign Slovenia" (FQ 19). With this question, we tested if it can distinguish a specific subgroup ("voters for independence") from a larger one ("all plebiscite participants") without an explicit comparison between them. So the chatbot can process and extract information from complex sentences that involve multiple categories or groups of people. Interestingly, analysis of its responses showed that one of the chatbot's incorrect responses included a piece of specific statistical data, non-existent in the document (88.5% of participants voted for independence). While this could be another example of "hallucination", this percentage exists in several peer-reviewed sources, even on Wikipedia (Pesek, 2007, 274; Čepič, 2005, 1297; and Jambrek, 2021, 221). This raises intriguing questions about the chatbot's pre-existing knowledge, and highlights potential complexities in interpreting chatbot outputs, especially when they appear to include information not explicitly present in the analysed text.

The chatbot's success was low (20%) in connecting missing acronyms with named entities; however, FQ 22 was designed to be particularly challenging. Instead of a less

complex task where the chatbot needs to identify the full meaning of an acronym, it required the chatbot to determine the existence of the full name "Jugoslovanska ljudska armada", which appears once on page 110 (Repe, 2004), based on its acronym "JLA", which is not present in the document at all. The low accuracy in this task exposes a limitation in the chatbot's ability to make inferences based on missing information.

The chatbot encountered even greater difficulties when tasks required it to count named entities. In both runs, it failed to determine the frequency of a specific last name in the document. All occurrences of "Poos" (FQ 15) are in the text's main body, while "Van den Broek" also appears in footnotes. In the former case, the correct answer was 8, but the chatbot's responses in 10 experiment runs ranged from a vague answer ("multiple times") to incorrect ones (2, 3 or 5 occurrences). In the case of "Van den Broek ", the correct answer was 5, but the chatbot's responses in 10 experiment runs, again, ranged from a vague answer ("at least twice") to incorrect ones (0, 1 or 2 occurrences). This highlights a limitation in its ability to handle counting tasks in the main body of the text as well as in the footnotes.

Similarly, the chatbot was unable to identify the language used by a participant, based on contextual clues ("sledi angleški prevod", FQ 17) (Repe, 2004, 108), which highlights another limitation. While the question ("In which language did Van den Brook speak at the meeting?") also included an intentional misspelling of the participant's last name ("Brook" instead of "Broek"), the chatbot used the correct spelling in its responses in all 10 runs. This suggests some ability to correct minor errors in the questions.

Recognising information across languages (date in Croatian, FQ 21) also proved a problem. The chatbot was asked to identify a date ("17. 8.") that appears twice in the document (Repe, 2004, 111) but in Croatian ("17. kolovoza"), yet without success. This suggests the chatbot's struggle to recognise the information presented in various languages. Interestingly, in the experiment run for Participant 2, the chatbot replied in English and then continued doing so for all the following questions in that run.

The chatbot also failed in complex information extraction tasks, for example, when asked to identify the presence of ambassador "Zimmerman" in the text. FQ 24 contained several layers of complexity: the accurately spelled name in the question ("Zimmerman") contained only one "m" in the document ("Zimerman"). Furthermore, the question used the term "veleposlanik" (Slovenian for "ambassador"), while the document used the synonym "ambasador" (a borrowing, but also used in Slovenian). The third hurdle was the appearance of the ambassador's last name exactly at the line break, with the first part ("Zimer-") in one line and the last part ("manu") in the following one (the final "u" is the dative (3rd nominal case) suffix in Slovenian, an equivalent of "to Zimerman"). This complex question reflects real-world scenarios where information might be presented in non-standard formats or with minor inconsistencies. This answer shows the chatbot's limited ability to extract multi-line information while simultaneously handling other language issues.

The chatbot's overall performance in identifying specific terms and details within the document was poor. While it excelled at extracting page location of an acronym, noting the absence of specific terms like "ethical conflicts", and performed moderately well in distinguishing a specific subgroup from a larger group, it faced challenges with other, more complex tasks. Extracting full names from acronyms, counting occurrences, and handling data across languages proved difficult. The most significant challenge was its inability to handle a question that combined information spread across lines, slight spelling variations, and synonyms. These findings suggest limitations beyond basic factual retrieval, particularly when dealing with intricate tasks or nuanced information.

5.1.5 Category 5: Language robustness: resistance to intentional typos, grammar mistakes, symbols and language switching (FQs 20 & 25)

The chatbot's performance on language robustness questions was poor. In none of the answers to FQ 20 (about the percentage of voters who supported Slovenian independence on the referendum) did it recognise the synonymic relation between the Slovenian expressions "referendum", used in the question, and "plebiscit", used in the document to provide the relevant percentage. The Croatian voters' percentage, however, was introduced with the expression "referendum", which most likely misled the chatbot to confuse the two bits of data. The chatbot was similarly unsuccessful in recognising the languages in which the document was written: despite the 2 correct answers in experiment 10 runs (FQ 25) in which it correctly identified the number of languages used in the document, it failed to list all three, and it misidentified one out of three in each case.

Despite a different primary focus, other questions also contributed to our insight into the chatbot's performance. Intentional misspellings in FQs 17 and 23 were handled with partial success. While the chatbot successfully accommodated the incorrect spelling of "Brook" by returning the correct version of the name in his answers (without the explicit mention of the typo), the "et(n)ični" (ethical/ethnic, FQ 23) issue was not recognised as a spelling mistake in the question. The chatbot was, therefore, unable to contextualise the question and make an assumption if a certain question is a sensible one to ask.

Several other (types of) language mistakes can be found in the chatbot's answers, most notably those having to do with punctuation (occasional missing commas), grammar (incorrect use of grammatical case or number), syntax (inaccurate word order), and the use of an English (sometimes Croatian) word or phrase in a Slovenian sentence. Considering that LLMs generate discourse based on the available texts in this language, and that the size of the available pool of Slovenian texts (in particular grammatically impeccable texts) is much smaller compared to that of English texts, the chatbot's generated Slovenian is rather good. There are rare cases of missing punctuation (e.g., Ni jasno omenjeno (missing comma) na kateri strani se nahaja informacija o lokaciji sestanka), while the majority of answers, including those with complex syntactic structure, contain all the necessary commas (e.g., Po poročanju magnetograma, ki ga vsebuje ta vir, se o tem, kako so si sodelujoči na sestanku interpretirali predlagano odložitev osamosvojitvenih dejanj Slovenije in Hrvaške, ni izrecno govorilo, vendar pa se je v poročilu o tem sestanku omenilo, da so ...). In the chatbot's answers, we occasionally found grammatical objects in incorrect cases (... neavtoriziran magnetogram razgovora med Milan**a** Kučan**a** in predsednik**a** Republike Hrvaške dr. Franj**a** Tuđman**a**). The –**a** suffix is a genitive case ending, while the Slovenian sentence structure (the preposition "med") would require the locative (5th) case ending -**om**. The mistake, however, is likely due to the Croatian sentence structure, where the preposition "između" (Slovenian: "med", English "between") goes with the genitive case; thus, the chatbot must have applied the Croatian syntax rule to a Slovenian sentence, possibly owing to the mention of the Croatian president. Not unexpectedly, the chatbot could not always master the dual case, which is a Slovenian grammatical particularity (e.g., V dokumentu so prisotni le 2 jeziki sporazumevanja ...). The phrase "so prisotni le 2 jeziki" is a plural sentence structure, requiring a dual structure "sta prisotna le 2 jezika". An example of a non-Slovenian word in a Slovenian sentence is "ni razvidno, da **document** vsebuje sprotne opombe"; however, these, like most other types of mistakes listed, are not very frequent, so all-in-all, the chatbot's Slovenian is relatively satisfactory.

5.2 Interpretative Questions

The chatbot struggled with all interpretative questions. It was unexpectedly strongest in IQ 5, the most complex of interpretative questions, indicating some ability to make value judgments and construct arguments for the judgment, which are crucial skills for a historian. It managed 6 conditionally acceptable answers (60% accuracy) and 0 entirely acceptable ones. In the "conditionally acceptable" cases, it accurately found that a definitive answer was not possible based solely on the document. For example, it suggested consulting additional sources to determine the meeting's success. On the other hand, it could not justify these judgments but simply reiterated irrelevant details from the document, for example: "At the meeting, the participants discussed how the situation in Slovenia and Croatia should continue and presented their demands and positions. During the meeting, conflicting views were highlighted, and no common ground was found on the proposed postponement of independence". In two cases, it searched the document for explicit mentions of success or failure and responded that "It is not clearly stated in this PDF source whether the meeting was successful or not".

The chatbot showed limitations in handling IQ 1, which required a more substantial answer located within a single paragraph of the document (0% accuracy). It either claimed the answer was not in the document or responded that it could not answer the question. It even recommended a rephrasing of the question: "Please rephrase the question for it to be more specific". The chatbot also performed poorly on IQ 2 (0% accuracy), which means that it cannot extract the main point of a more extended exposition from a paragraph. In all experiment runs, it stated, usually directly but sometimes also indirectly, that the document did not contain information about the required answer or that the answer could not be determined: "It cannot be derived from document Doc. no. 27 that the European "Troika" opposed the Slovenian and Croatian declarations of independence". It could be argued that the chatbot questioned our knowledge of the topic.

In IQ 3, we tested the chatbot's ability to provide a substantial interpretative answer for which the information should be derived from multiple paragraphs. This was not successful (0% accuracy). In six experiment runs, it focused on the wrong paragraphs and/or paragraph passages, and consequently provided wrong answers or summaries, for example: "Considering the text, it seems that the discussion topics were mainly related to the war in Slovenia and Croatia at the time. The details of the key positions of the Slovenian representatives are, therefore, not discernible from this document". The chatbot could not provide a more extended and substantial answer. Moreover, it "hallucinated" in one of the experiment runs, stating that it searched for the answer in "two magnetograms," which is inaccurate. Furthermore, in another run it ignored all but the first and second page of the document. Finally, in a total of 4 experiment runs, it refused to cooperate entirely.

The chatbot also failed at the more complex IQ 4 (0% accuracy), as it could not derive a concise interpretative answer from several paragraphs and/or paragraph passages. In half of the experiment runs, it wrote that the document contained no details about the interpretations of the proposed postponement of the Slovenian and Croatian independence process, which was the main point of our question, but it still tried to help, although unsuccessfully. In three experiment runs, the chatbot confirmed that the document contained the interpretations of the meeting participants we were looking for but gave an incorrect answer. In two of these runs, it also stated the supposed pages of these interpretations, which contain part of the answer and can thus be helpful to the researcher. The chatbot only replied that Poos had proposed a three-month postponement of Slovenia's and Croatia's independence acts without explaining how Poos understood this postponement, which was the gist of the question. Also, in the case

of IQ 4, the chatbot refused to cooperate with the participant in two experiment runs. In one run, it again stated that it had analysed two magnetograms.

According to our findings, answering complex interpretative historical questions, for which critical analysis and reasoning is essential, remains a significant challenge for the chatbot. This is reflected by how it answers factual questions, struggles with demanding fact-retrieval tasks, and its tendency to generate unsubstantiated claims or "hallucinations". Apart from these limitations, our results reveal the chatbot's inability to critically evaluate information and identify cause-and-effect relationships, which are crucial skills for historical reasoning. These limitations suggest that the chatbot, in its current form, cannot accurately perform tasks that require historical reasoning based on in-depth interpretation of historical documents.

6 CONCLUSIONS

Our study investigated the capability of the PDFGear Copilot's chatbot powered by GPT-3.5 in supporting historical document analysis. We explored three key areas:

Factual Information Identification and Extraction

The chatbot exhibited mixed success in answering factual questions, its success in the four factual categories (FQs 1-19 & 21-24) was 39.6% (91 correct answers from 230). It was able to navigate the document structure, which is crucial for locating relevant information. Its strengths were primarily in named entity recognition, where it successfully identified the names and roles of key individuals (regardless of spelling variations) and recognised the absence of entities like document author. Additionally, it successfully located page numbers for acronyms and recognised the absence of specific terms like "ethical conflicts". In extracting basic document information like page count and creation date, it achieved moderate-high to high accuracy. It also exhibited some potential in distinguishing subgroups from larger groups, as well as in identifying active meeting participants. However, its performance was weaker in other tasks. Document type identification and meeting details like location and duration proved challenging. Group member identification also showed limitations. The chatbot struggled with extracting full names from acronyms, distinguishing document titles from volume titles, counting named entity occurrences, document characters, or footnotes, and handling information across languages and lines (such as line breaks). It also struggled with complex sentence structures, which can potentially lead to misinterpretations of factual details. Moreover, the chatbot exhibited a concerning tendency to generate "hallucinations".

Interpretative Questions

The chatbot's limitations were most evident in its inability to answer complex, interpretive historical inquiries (IQs 1-5), with only 12% success (6 acceptable answers out of 50). It lacks skills for critical analysis, it cannot link cause-and-effect relationships, and it cannot manage tasks that require in-depth interpretation of the document.

Handling Language Complexities

The chatbot exhibited some success in handling typos and misspellings, and it demonstrated a degree of robustness against minor language errors. This could be beneficial for historical documents where typos and inconsistencies might be present. However, the chatbot faced challenges with tasks requiring a deeper understanding of language (FQs 20 & 25), managing only 20% success (2 acceptable answers out of 10). Complex sentence structures with intricate grammar and phrasing were often perceived as obstacles. Finally, the chatbot struggled with domain-specific terminology in Slovenian, potentially limiting its ability to grasp the full context of the document.

Overall, our findings suggest that while the chatbot holds some promise for assisting with certain factual retrieval tasks from digitised non-English historical documents, it is not currently suitable for tasks requiring historical reasoning grounded in critical analysis, interpretation, and understanding of cause-and-effect relationships. Further development is needed to address the identified limitations, particularly the "hallucinations" tendency and the complete lack of critical reasoning skills. Nonetheless, our study represents a pioneering step in integrating AI into historical research, demonstrating both its potential as well as its challenges. By incorporating the chatbot as a novel tool in the historian's methodological toolkit, we have laid the groundwork for future research to explore the potential of the developing AI in historical inquiry.

Finally, some limitations and considerations for future research are worth noting. The analysis is based on a limited number of questions and a single historical document type and format. Moreover, the study was conducted in March 2024, preceding advancements like the release of GPT-4o and its document analysis capabilities. Future research could explore a wider range of document types, compare performances of various chatbots, and incorporate metrics like response time and user experience. A deeper analysis of errors, question wording, and potential biases in the chatbot's decision-making process would also be valuable. Additionally, instead of using zero-shot prompting, we could repeat the study using few-shot prompting, or per-haps chain-of-thought prompting, which has recently shown potential (Cheng et al., 2024). These findings may also offer insights applicable to history-related educational settings that use English as a language of instruction since, according to Picciuolo (2023, 95), studies investigating the role of educational technology in the EMI (English as a Medium of Instruction) classroom remain scarce.

7 APPENDIX

Appendix contains a reference table with the factual and interpretative questions used in the experiment. To each question, the experts have added a reference answer (an expected, correct/acceptable answer), mostly relying on Repe (see section 3.2 above) as the main reference source for the analysed document. Finally, the table contains a brief justification/ explanation of the reference answer. The experiment is laid out in subsection 3.5., and the results of the experiment referring directly to Section 4.

8 REFERENCES

Abadie, A., Chowdhury, S. & Mangla S. K. (2024). A shared journey: experiential perspective and empirical evidence of virtual social robot ChatGPT's priori acceptance. *Technological Forecasting and Social Change*, 201, 123202. DOI: https://doi.org/10.1016/j.techfore.2023.123202.

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K. & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change Volume*, *199*, *123076*. DOI: https://doi.org/10.1016/j.techfore.2023.123076.

Altmäe, S., Sola-Leyva, A. & Salumets, A. (2023). Artificial Intelligence in Scientific Writing: A Friend or a Foe? *Reproductive BioMedicine Online, 47*, 3–9. DOI: 10.1016/j.rbmo.2023.04.009.

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S. & Kaplan, F. (2024). Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study. In Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, & S. Szpakowicz (Eds.), *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 133–159). Stroudsburg, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.latechclfl-1.pdf (accessed on April 4 2024).

Buholayka, M., Zouabi, R. & Tadinada, A. (2023). Is ChatGPT Ready to Write Scientific Case Reports Independently? A Comparative Evaluation Between Human and Artificial Intelligence. *Cureus*, *15*, *e39386*. DOI: 10.7759/cureus.39386.

Carretero, M. & Gartner, E. (2024). Artificial Intelligence and historical thinking: a dialogic exploration of ChatGPT / Inteligencia Artificial y pensamiento histórico: una exploración dialógica del ChatGPT. *Studies in Psychology*, *45*, 80–102. DOI: https://doi.org/10.1177/02109395241241379.

Čepič, Z. (2005). Plebiscit o samostojni Sloveniji. In J. Fischer et al. (Eds.), *Slovenska novejša zgodovina: od programa Zedinjena Slovenija do mednarodnega priznanja Republike Slovenije: 1848-1992* (pp.1294–1297). Ljubljana: Inštitut za novejšo zgodovino & Mladinska knjiga.

Cheng, X., Li, J., Zhao, W. X. & Wen, J. (2024). ChainLM: Empowering Large Language Models with Improved Chain-of-Thought Prompting. *arXiv*. Available online: arXiv:2403.14312 (accessed on April 2 2024).

Chiarello, F., Giordano, V., Spada, I., Barandoni, S. & Fanton, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, *133*, *103002*. DOI: https://doi.org/10.1016/j.technovation.2024.103002.

Currie, G. M. (2023). Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy? *Seminars in Nuclear Medicine*, *53*, 719–730. DOI: https://doi.org/10.1053/j.semnuclmed.2023.04.008.

Deng, J. & Lin, Y. (2023). The benefits and challenges of ChatGPT: an overview. *Frontiers in Computing and Intelligent Systems*, *2*, 81–83. DOI: https://doi.org/10.54097/fcis.v2i2.4465.

Diederich, S., Brendel, A. B., Morana, S. & Kolbe, L. (2022). On the design of and interaction with conversational agents: an organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, *23*, 96–138. DOI: 10.17705/1jais.00724.

Drumm, K., & Tran, V. (2023). Examining the Effectiveness of Chatbots in Gathering Family History Information in Comparison to the Standard In-Person Interview-Based Approach. *Catalyzex*. Available online: https://www.catalyzex.com/paper/examining-the-effectiveness-of-chatbots-in (accessed on April 4 2024).

Dwivedi, Y .K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M. et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, *102642*. DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, *613*, *423*. DOI: 10.1038/d41586-023-00056-7.

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative model. *Philosophy & Technology*, *36*, *15*. DOI: 10.1007/s13347-023-00621-y.

Gonzalez Garcia, G & Weilbach, W. (2023). If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History. *arXiv*. Available online: arXiv:2310.10808 (accessed on April 2 2024).

Henrickson, L. (2023). Chatting with the dead: The hermeneutics of Thanabots. *Media, Culture & Society*, *45*, 949–966. DOI: https://doi.org/10.1177/016344372211476.

Hribernik, M. (2018). Die Schlacht um Vukovar im Jahr 1991. *Studia Historica Slovenica*, *18 (1)*, 251–276. DOI: https://dx.doi.org/10.32874/SHS.2018-10.

Jambrek, P. (2021). Plebiscit. In M. Avbelj et al. (Eds.), *Osamosvojitev: prispevki za enciklopedijo slovenske osamosvojitve, državnosti in ustavnosti* (pp. 220–222). Nova Gorica: Nova univerza.

Kádár, D. Z., House, J., Todorović, T., Onič, T., Hazemali, D., Plemenitaš, K. & Brown, D. (2024). The language of diplomatic mediation: a case study of an emergency meeting in the wake of the Yugoslav wars. *Language & Communication: an interdisciplinary journal*, *96*, 54–66. DOI: 10.1016/j.langcom.2024.02.004.

Kansteiner, W. (2022). Digital Doping for Historians: Can History, Memory, and Historical Theory be Rendered Artificially Intelligent? *History and Theory*, *61*, 119–133. DOI: https://doi.org/10.1111/hith.12282.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, *102274*. DOI: https://doi.org/10.1016/j.lindif.2023.102274.

Lorenz, M. & Konečný, M. (2023). Digital Archives as Research Infrastructure of the Future. *Acta Informatica Pragensia*, *12*, 327–341. DOI: 10.18267/j.aip.219.

Lozić, E. & Štular, B. (2023). Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet*, *15*, *336*. DOI: https://doi.org/10.3390/fi15100336.

Makhortykh, M., Zucker, E. M., Simon, D. J., Bultmann, D. & Ulloa, R. (2023). Shall androids dream of genocides? How generative AI can change the future of memorialization of mass atrocities. *Discover Artificial Intelligence* 2023, *3*, *28*. DOI: https://doi.org/10.1007/s44163-023-00072-6.

Mezeg, A. & Žigon, T. (2023). "A Carniolan also learns Latin and French at grammar school" : France in the light of the articles of the Ljubljana German weekly newspaper for benefit and amusement. *Annales : anali za istrske in mediteranske študije*, *33 (2)*, 299–314. https://doi.org/10.19233/ASHS.2023.15.

Osojnik, J. (2023). Demosova plebiscitna pobuda: analiza spominske literature in dogajanje konec oktobra in v začetku novembra 1990. *Annales: Series historia et sociologia*, *33 (3)*, 527–536. DOI: 10.19233/ASHS.2023.27.

Osojnik, J. (2022). Predlog Socialistične stranke Slovenije oktobra 1990 za izvedbo plebiscita o samostojnosti Republike Slovenije in odzivi nanj v Sloveniji. *Studia Historica Slovenica*, *22 (2)*, 463–502. DOI: 10.32874/SHS.2022-13.

Pentzold, C., Lohmeier, C. & Birkner, T. (2024). Communicative remembering: Revisiting a basic mnemonic concept. *Memory, Mind & Media*, *2*, 1–15. DOI: 10.1017/mem.2023.7.

Pesek, R. (2007). *Osamosvojitev Slovenije: ali naj Republika Slovenija postane samostojna in neodvisna država?* Ljubljana, Slovenia: Nova revija.

Pessanha, F. & Salah, A. A. A. (2021). A Computational Look at Oral History Archives. *Journal on Computing and Cultural Heritage*, *15*, 1–16. DOI: 10.1145/3477605.

Picciuolo, M. (2023). An ELF-Oriented Corpus-Based Analysis into the EMI Lecturers' Use of Spatial Deixis across Two Different Teaching Media. *ELOPE: English Language Overseas Perspectives and Enquiries*, *20 (1)*, 89–112. https://doi.org/10.4312/elope.20.1.89-112.

Pirjevec, J. (1995). Jugoslavija 1918–1992: nastanek, razvoj ter razpad Karadjordjevićeve in Titove Jugoslavije; Koper, Slovenia: Lipa.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. DOI: 10.1016/j.iotcps.2023.04.003.

Repe, B. (ed.). (2002a). *Viri o demokratizaciji in osamosvojitvi Slovenije. I. del: Opozicija in oblast*. Ljubljana, Slovenia: Arhivsko društvo Slovenije.

Repe, B. (ed.). (2003). *Viri o demokratizaciji in osamosvojitvi Slovenije. II. del: Slovenci in federacija*. Ljubljana, Slovenia: Arhivsko društvo Slovenije.

Repe, B. (ed.). (2004). *Viri o demokratizaciji in osamosvojitvi Slovenije. III. del: osamosvojitev in mednarodno priznanje*. Ljubljana, Slovenia: Arhivsko društvo Slovenije.

Repe, B. (2001). *Slovenci v osemdesetih letih*. Ljubljana, Slovenia: Zveza zgodovinskih društev Slovenije.

Repe, B. (2002b). *Jutri je nov dan: Slovenci in razpad Jugoslavije*. Ljubljana, Slovenia: Modrijan.

Repe, B. (2015). *Milan Kučan, prvi predsednik*. Ljubljana, Slovenia: Modrijan.

Repe, B. & Kerec, D. (2017). *Slovenija, moja dežela: družbena revolucija v osemdesetih letih*. Ljubljana, Slovenia: Cankarjeva založba.

Repe, B. (2022). Slovensko-srbski konflikt v osemdesetih letih. *Studia Historica Slovenica* 2022, *22 (2)*, 305–341. DOI: 10.32874/SHS.2022-08.

Salvagno, M., Taccone, F. S. & Gerli, A.G. (2023). Can Artificial Intelligence Help for Scientific Writing? *Critical Care* 2023, *27*, *75*. DOI: 10.1186/s13054-023-04380-2.

Sisto, D. (2020). *Online Afterlives: Immortality, Memory, and Grief in Digital Culture*. Cambridge, United Kingdom: MIT Press.

Siu, S. C. (2023). ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. Preprint. DOI: 10.2139/ssrn.4448091.

Spina, S. (2023). Artificial Intelligence in archival and historical scholarship workflow: HTS and ChatGPT. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2308.02044.

Stevenson C., Smal, I., Baas, M., Grasman, R. & Van Der Maas, H. (2022). Putting GPT-3's creativity to the (alternative uses) test. *arXiv*. DOI: 10.48550/arXiv.2206.08932.

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, *613*, 620–621. DOI: 10.1038/d41586-023-00107-z.

Thomas, A., Gaizauskas, R. & Lu, H. (2021). Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages* (pp. 116–121). Torino, Italy: ELRA Language Resources Association & International Committee on Computational Linguistics.

Tirado-Olivares, S., Navío-Inglés, M., O'Connor-Jiménez, P. & Cózar-Gutiérrez, R. (2023). From Human to Machine: Investigating the Effectiveness of the Conversational AI ChatGPT in Historical Thinking. *Education Sciences*, *13*, *803*. DOI: https://doi.org/10.3390/educsci13080803.

Trichopoulos, G., Konstantakis, M., Caridakis, G., Katifori, A. & Koukouli, M. (2023). Crafting a Museum Guide Using ChatGPT4. *Big Data and Cognitive Computing* 2023, *7*, *148*. DOI: https://doi.org/10.3390/bdcc7030148.

Varitimiadis, S., Kotis, K, Pittou, D. & Konstantakis, G. (2021). Graph-Based Conversational AI: Towards a Distributed and Collaborative Multi-Chatbot Approach for Museums. *Applied Sciences*, *11*, *9160*. DOI: https://doi.org/10.3390/app11199160.

Vuković Vojnović, D. (2023). "Experience Norfolk! Experience Fun!" vs. "Doživi više od očekivanog" – A Corpus-Based Contrastive Study of Reader Engagement Markers on the Web. *ELOPE: English Language Overseas Perspectives and Enquiries*, *20 (1)*, 133–150. https://doi.org/10.4312/elope.20.1.133-150.

Zajc, M. (2020). Poletni aferi kritičnih misli. Tomaž Mastnak in Dimitrij Rupel, slovenska kritična intelektualca med jugoslovansko in slovensko javnostjo v letu 1986. *Studia Historica Slovenica*, *20 (3)*, 921–955. DOI: 10.32874/SHS.2020-26.

FACTUAL QUESTIONS (FQ)

Question	Reference answer	Justification of the question
FQ 1 What is the title of the document?	The title of the document is Unauthorised tape recording of the conversation between the President of the Presidency of the Republic of Slovenia, Milan Kučan, and the President of the Republic of Croatia, Dr Franjo Tuđman, with the Delegation of the European Community, Zagreb, 28/29 June 1991 from 23:30 to 01:15.	To assess the chatbot's ability to distinguish between the title of the document itself and the title of the volume in which the document was published.
FQ 2 What is the document type?	It is a PDF document.	To check the chatbot's capacity to identify the file type or format of the document (e.g., PDF, Word).
FQ 3 When was the record made?	The exact date of creation is unknown.	To check the chatbot's ability to extract the date when the document was created or modified.
FQ 4 Who is the author of the document?	The author of the document is unknown.	To check the chatbot's capacity to recognise and identify the author(s) of the document.
FQ 5 How many pages does the document have?	The document is 10 pages long.	To check the chatbot's ability to determine the number of pages in the document.
FQ 6 How many non-spaced characters does the document have?	The document has 36,744 characters excluding spaces.	To check the chatbot's capacity to count the number of characters (without spaces) in the document.
FQ 7 How many footnotes does the document have?	The document has 4 footnotes.	To check the chatbot's ability to recognise and count footnotes within the document.
FQ 8 How long did the meeting last?	The meeting lasted one hour and 45 minutes (105 minutes).	To check if the chatbot can calculate the length of the meeting based on the listed beginning and end times of the meeting.
FQ 9 Where did the meeting take place?	The meeting took place in Zagreb.	To check the chatbot's ability to identify the location of the meeting.
FQ 10 How many different interlocutors actively participated in the meeting?	From the document, it is not clear how many speakers were present. At one point during the meeting, "all" present joined the debate at the same time (simultaneously); their number is not known.	To check if the chatbot was able to discern that at some point "all" participants had spoken at the same time, while their number is not explicitly stated in the document.
FQ 11 Who was the President of the Presidency of the Republic of Slovenia at the time of the meeting?	At the time of the meeting, the President of the Presidency of the Republic of Slovenia was Milan Kučan.	To check if the chatbot was able to identify the President of the Presidency of the Republic of Slovenia.
FQ 12 Who were the representatives of the European "Troika", present at the meeting?	Jacques Poos, Foreign Minister of Luxembourg, Gianni de Michelis, Foreign Minister of Italy, and Hans Van den Broek, Foreign Minister of the Netherlands.	To check if the chatbot was able to identify the European "troika" among the attendees. There is no explanation in the document about the formation of troika.

FQ 13 Who was the Foreign Minister of the Netherlands at the time of this meeting?	(Hans) van den Broek.	To check if the chatbot can read the footnotes. Van den Broek is mentioned as Foreign Minister of the Netherlands in footnote 150.
FQ 14 On which page of the document does the acronym KEVS first appear?	On page 109 of the document or on page 2 of the clipping.	To check if the chatbot can identify the location of a simple factoid in a document when target searching for it.
FQ 15 How many times does the surname Poos appear in the document?	8 times.	To check if the chatbot can count the exact number of occurrences of a particular surname.
FQ 16 How many times does the surname Van den Broek appear in the document?	5 times.	The purpose was twofold. First, we wanted to check if the chatbot was capable of cumulative counting both in the text of the document and in the footnotes. Second, we wanted to see whether the chatbot would recognise the name of Van den Broek, misspelt as Van den Brook.
FQ 17 Which language did Van den Brook speak at the meeting?	Van den Broek spoke in English/English language.	The purpose was twofold. First, to check if the chatbot would notice the (incorrect and deceiving) phrase before Van den Broek's speech, saying, "here follows the English translation"; since Van den Broek spoke in English, the phrase should say "here follows the translation from English". Secondly, to check if the intentional misspelling of Van den Broek's surname would affect the chatbot's response.
FQ 18 In what capacity did Franjo Tuđman appear at the meeting?	Franjo Tuđman appeared at the meeting as the President of the Republic of Croatia.	To check if the chatbot can recognise Franjo Tuđman as the President of the Republic of Croatia, in spite of occasional incorrect spellings of his name (Tudjman, Tudjaman).
FQ 19 What percentage of the plebiscite participants voted in favour of "an independent and sovereign Slovenia"?	93% of the plebiscite participants voted in favour of "an independent and sovereign Slovenia".	To check if, in a complex sentence, the chatbot can distinguish between two different categories of people: all eligible voters versus those who attended and voted "yes".
FQ 20 What percentage of the referendum participants voted in favour of "an independent and sovereign Slovenia"?	93% of the plebiscite participants voted in favour of "an independent and sovereign Slovenia".	To check if the chatbot recognises synonyms: plebiscite and referendum. In the document, the word plebiscite is used in the report about the Slovenian referendum result, and "referendum" in the report about the Croatian referendum result.
FQ 21 Does the document mention the date 17. 8.?	Yes.	To check if, when asked a Question in Slovenian, the chatbot can retrieve the answer from the document written in Croatian. The date 17. 8. appears twice in the document, on page 111: "17. kolovoza".
FQ 22	Yes.	The JNA is referred to once with its full name, the Yugoslav National Army, on page 110. This question was designed to

Does the document refer to the JNA by its full name?		test whether the chatbot can infer that the abbreviation JNA stands for Yugoslav National Army from a given context. The abbreviation JNA does not appear in the document.
FQ 23 Are "ethical conflicts" mentioned in the text?	No, there is no mention of ethical conflicts in the document.	Firstly, to determine whether the chatbot can recognise (intentional) misspellings in a question. Given the content of the document, we were interested in references to "ethnic" conflicts, not "ethical" conflicts. Secondly, to check if the chatbot would answer the question in a way that made sense in the given context, considering the misspelt term (mention of "ethnic conflicts") or take it literally (mention of "ethical conflicts").
FQ 24 Does the text mention Ambassador Zimmerman?	Yes, he is mentioned on page 113.	To check if the chatbot can recognise factual data spanning over two lines. In the question, the surname was spelt correctly, even though in the document it is spelt with only one "m". We also deliberately used the synonym "veleposlanik" (Slovene for "ambassador") for the Slovenian term "ambasador" used in the document.
FQ 25 How many different languages of communication are present in the document?	Three different languages are present: Slovenian, Croatian and Serbo-Croatian.	To determine whether the chatbot can recognise different related languages.

INTERPRETATIVE QUESTIONS (IQ)

Question	**Reference answer**	**Justification of the question**
IQ 1 What were the European Troika's initial questions at the Zagreb talks about the possibility of ending the conflict?	The European "Troika" had the following starting questions: Can an immediate ceasefire and withdrawal of troops to barracks be agreed upon? Is it possible to delay the implementation of the Slovenian and Croatian declarations of independence for three months? Do the Slovenian and Croatian participants in the meeting support the election of Stipe Mesić as President of the Presidency of the SFRY?	To determine whether the chatbot can provide a longer, more substantial answer, which is easy to find in the document and is contained in a single paragraph. CRITERION: All or nothing; this question eliminates the conditionally acceptable option.
IQ 2 What arguments did the Troika use to oppose the Slovenian and Croatian declarations of independence?	The declaration of Slovenian and Croatian independence was contrary to the Helsinki and Berlin Declarations and international law.	To check if the chatbot recognises that the answer was given by one of the members of the "troika", who is referred to by his surname. Furthermore, to assess whether the chatbot can extract the main point from a slightly longer exposition.
IQ 3 What were the key starting points of the Slovenian representatives at the Zagreb negotiations?	Slovenia was attacked; the destruction in Slovenia is a consequence of the war/conflict; Yugoslavia showed no willingness to hold talks after the plebiscite; Slovenia was prepared to talk about a gradual takeover of sovereign functions; Slovenia did not intend to stay in Yugoslavia because of the broken trust within Yugoslavia; Slovenia wanted to internationalise the Yugoslav problem; the	To check if the chatbot can give a longer substantial answer, which spans more than one paragraph, which was the case for the first two questions.

	Slovenian side wanted to stop the war but not at the cost of independence; the Yugoslav side initially refused to establish contact with Slovenia after the outbreak of the conflict; Slovenia did not obstruct transit from Yugoslavia – the outbreak of the war changed this, and it was the JNA's fault; Slovenia was economically exhausted, since it was the workhorse of the Yugoslav economy; the EEC's insistence that Yugoslavia remains united led to the suspension of democratic processes in Yugoslavia.	
IQ 4 How did the participants at the meeting interpret the proposed postponement of Slovenia's and Croatia's independence activities?	The representatives of the Slovenian and Croatian republican authorities understood that a postponement would mean a cessation of the planned independence process, while the "troika" understood that a postponement would mean a "temporary" withdrawal from the adopted independence acts and a return to the old status quo, i.e. the one before the declaration of independence.	To check if the chatbot can give an answer that can be found in several paragraphs, i.e. in the words of several speakers at the meeting. The aim was to determine whether the chatbot can give an answer in a concise form.
IQ 5 Was the meeting a success?	There is no clear answer to this question, but statements about successful and unsuccessful aspects can be found. The EEC has successfully acquired first-hand perspectives from the Slovenian and Croatian sides on the developments in Slovenia and Yugoslavia in general. Tuđman's positive response to all initial questions of "Troika" was also a success. The fact that there was no ceasefire in Slovenia must be considered a failure.	To check if the chatbot can make a value judgment and, consequently, whether it is able to provide an interpretation, which is crucial for academics working in the humanities.

POVZETEK

VREDNOTENJE POGOVORNEGA SISTEMA KOT POMOČNIKA PRI ANALIZI ZGODOVINSKIH DOKUMENTOV

asist. dr. David HAZEMALI
Univerza v Mariboru, Filozofska fakulteta, Slovenija
david.hazemali@um.si

asist. Janez OSOJNIK
Univerza v Mariboru, Filozofska fakulteta, Slovenija
janez.osojnik1@um.si

izr. prof. dr. Tomaž ONIČ
Univerza v Mariboru, Filozofska fakulteta, Slovenija
tomaz.onic@um.si

asist. dr. Tadej TODOROVIĆ
Univerza v Mariboru, Filozofska fakulteta, Slovenija
tadej.todorovic@um.si

asist. dr. Mladen BOROVIČ
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Slovenija
mladen.borovic@um.si

Članek preučuje potencial pogovornega sistema PDFGear Copilot, ki ga poganja veliki jezikovni model GPT-3.5, za pomoč pri analizi zgodovinskih dokumentov. Želeli smo ugotoviti, ali je mogoče pogovorni sistem kot nov pristop umestiti v okvir metodologije, temelječe na tradicionalnih raziskovalnih praksah, pri raziskovalnem delu zgodovinarja. Namen članka je analizirati, kako učinkovito lahko obravnavani pogovorni sistem pomaga pri analizi in interpretaciji primarnih zgodovinskih virov.

V članku smo (o)vrednotili uspešnost pogovornega sistema kot pomočnika pri analizi dokumenta, povezanega s slovensko osamosvojitveno vojno. Izbrani in analizirani dokument je v slovenskem jeziku, prav tako je tudi eksperiment potekal v slovenskem jeziku, s čimer smo hkrati ocenjevali večjezične zmožnosti in jezikovne spretnosti pogovornega sistema.

Sposobnost pogovornega sistema kot pomočnika pri omenjeni analizi smo preverili na podlagi 25 faktografskih in 5 interpretativnih vprašanj, ki so spraševala po različnih vidikih dokumenta, vključno z njegovimi lastnostmi, vsebinskimi podrobnostmi in specifičnimi izrazi. Obenem smo pozornost posvetili še jezikovni rabi in robustnosti pogovornega sistema ter njegovi sposobnosti poglobljene interpretacije in kritične analize. Cilj ni bil le preizkusiti točnost njegovih odgovorov na manj zahtevna faktografska vprašanja, temveč tudi ovrednotiti njegovo zmožnost reševanja nalog, ki zahtevajo razumevanje in sklepanje na višji taksonomski ravni, torej spretnosti, ki so nujne pri raziskovalnem delu zgodovinarja.

Raziskava je pokazala, da je bil pogovorni sistem deloma uspešen pri odgovarjanju na faktografska vprašanja. Njegova uspešnost se je sicer razlikovala glede na različne vidike dokumenta in njegove vsebine. Pokazal je določeno zmožnost prepoznavanja strukture dokumenta, razpoznavanja imenskih entitet in pridobivanja osnovnih informacij o dokumentu. Po drugi strani je bil bistveno manj uspešen pri prepoznavanju vrste dokumenta, podrobnostih o vsebini sestanka, ki ga dokument obravnava, prepoznavanju prisotnih članov na sestanku in nalogah, ki zahtevajo poglobljeno analizo besedila. Pogovorni sistem se je pri odgovarjanju na interpretativna vprašanja odrezal zelo slabo.

Ni uspel prepoznati vzročno-posledičnih povezav ter prikazati zahtevane poglobitve in natančnosti, ki so potrebne pri zgodovinopisnih raziskavah. Ugotavljamo, da je izbrani pogovorni sistem precej omejen, da bi bil zmožen v zadovoljivi meri pomagati zgodovinarju pri njegovem raziskovalnem delu. Ti sistemi sicer kažejo obetaven potencial, a bo potreben še znaten napredek, da bodo lahko nudili podporo na ustrezno visokem nivoju.

Opozoriti velja na nekatere omejitve te študije in možnosti za prihodnje raziskave. Analiza temelji na omejenem številu vprašanj ter eni vrsti in obliki zgodovinskega dokumenta. Poleg tega je bila študija izvedena marca 2024, torej pred izidom jezikovnega modela GPT-4o in njegovih zmogljivosti za analizo dokumentov. Prihodnje raziskave bi lahko preučile širši nabor vrst dokumentov, primerjale delovanje različnih pogovornih sistemov in vključile metrike, kot sta odzivni čas in uporabniška izkušnja. Dragocena bi bila tudi poglobljena analiza napak in morebitnih pristranskosti v procesu odločanja pogovornega sistema.

About the authors:

David Hazemali is a teaching assistant at the Faculty of Arts, University of Maribor, Slovenia. His academic expertise includes migration and ethnic studies, 20th-century military history, collaboration of university and industry, Yugoslav and contemporary Slovenian history, and the language of diplomacy. Hazemali is actively engaged in interdisciplinary research projects exploring cultural, social, and environmental transitions. These include investigating religious and spiritual care in the Slovenian Armed Forces, developing strategies for effectively communicating the climate crisis to support the transition to a green society, and examining historical interactions with foreigners in Upper Adriatic cities during transformative periods of societal change. Hazemali has published in leading peer-reviewed journals and presented his research at several international and local conferences and symposiums.

Janez Osojnik is a teaching assistant at the Department of History, Faculty of Arts, University of Maribor. He is completing his PhD thesis under the supervision of Prof. Dr. Gorazd Bajc. His PhD thesis is a comprehensive study of the plebiscite held on 23 December 1990 in the Republic of Slovenia. His general research interests are mainly in the study of plebiscites and their circumstances, both in the Slovenian ethnic territory and international relations – particularly in the context of British diplomacy – and in the historical issues of Slovenian national identity. He received the Miklošič Award at the Faculty of Arts, University of Maribor, which is awarded for the best Master's Thesis in a particular department in one academic year.

Tomaž Onič is an Associate Professor of English and American Literature at the Department of English and American Studies, Faculty of Arts, University of Maribor. He teaches courses in Anglophone literatures and literary translation. His research focuses on contrastive literary stylistics, graphic novels and, more recently, on pragmatics and discourse analysis of literary and non-literary texts. He often connects literary studies with other fields of the humanities and social sciences, such as theatre, music, history, and intercultural studies. He has published numerous scholarly articles and book chapters on these topics. He is also the editor of several scientific monographs and special issues of academic journals. He has led and participated in various research projects. He regularly attends international scientific conferences and other events abroad as well as in the local community.

Tadej Todorović is a teaching assistant of linguistics and philosophy with a PhD in philosophy from the University of Maribor. His area of interest includes pragmatics and philosophy of language, with an interest in speech act analysis in pragmatics, specifically regarding politeness and interaction. His special interest is the analysis of the influence and relevance of the analytic philosophy of language on pragmatics and speech act analysis of literary works, specifically drama. Tadej Todorović actively participates in several research projects related to climate change, digitalisation, and artificial intelligence in education. He has published several articles in peer-reviewed journals and has participated in several international conferences and symposiums. He is also the editor of the Slovenian philosophy journal *Analiza* and the technical editor of the international philosophy journal *Acta Analytica*.

Mladen Borovič is a teaching assistant and researcher at the Faculty of Electrical Engineering and Computer Science at the University of Maribor. His research areas include applied artificial intelligence, search engines and recommender systems, natural language processing, similar content detection, high-performance computing, and generative artificial intelligence. He actively works on projects related to natural language processing and the application of artificial intelligence in the form of conversational systems. He has participated in notable projects such as the National Open Access Infrastructure, Slovene in the Palm of Your Hand, HPC RIVR, and the Development of Slovene in a Digital Environment. As part of his teaching, he lectures on topics related to computer networks, high-performance computing, and artificial intelligence. He is also a reviewer for several peer-reviewed scientific journals with an impact factor in the field of computer science.

O avtorjih:

David Hazemali je asistent na Filozofski fakulteti Univerze v Mariboru. Njegovo akademsko področje vključuje migracijske in etnične študije, vojaško zgodovino 20. stoletja, sodelovanje med univerzo in industrijo, jugoslovansko in sodobno slovensko zgodovino ter jezik diplomacije. Hazemali aktivno sodeluje v interdisciplinarnih raziskovalnih projektih, ki raziskujejo kulturne, družbene in okoljske spremembe. Ti vključujejo raziskovanje verske in duhovne oskrbe v Slovenski vojski, razvoj strategij za učinkovito komuniciranje o podnebni krizi v podporo prehodu v zeleno družbo ter preučevanje zgodovinskih interakcij s tujci v zgornjejadranskih mestih v transformacijskih obdobjih družbenih sprememb. Hazemali je objavil članke v vodilnih recenziranih revijah ter sodeloval na številnih mednarodnih in domačih konferencah in simpozijih.

Janez Osojnik je asistent na Oddelku za zgodovino Filozofske fakultete Univerze v Mariboru. Zaključuje doktorski študij pod mentorstvom red. prof. dr. Gorazda Bajca. V doktorski disertaciji se posveča celostni obravnavi plebiscita, izvedenega 23. decembra 1990 v Republiki Sloveniji. Nasploh se raziskovalno ukvarja predvsem s preučevanjem plebiscitov in njihovih okoliščin tako na slovenskem etničnem ozemlju kot v mednarodnih odnosih – zlasti v okviru britanske diplomacije – ter s historičnimi vprašanji slovenske narodne identitete. Za magistrsko nalogo je prejel Miklošičevo priznanje na Filozofski fakulteti Univerze v Mariboru, ki se podeljuje za najboljšo magistrsko nalogo na posameznem oddelku v tekočem študijskem letu.

Tomaž Onič je izredni profesor za angleško in ameriško književnost na Oddelku za anglistiko in amerikanistiko Filozofske fakultete Univerze v Mariboru. Poučuje predmete s področja anglofonih književnosti in literarnega prevajanja, raziskovalno pa se ukvarja s kontrastivno literarno stilistiko, z grafičnim romanom, v zadnjem času pa tudi s pragmatiko in diskurzno analizo literarnih in neliterarnih besedil. Literarnovedne študije pogosto povezuje z drugimi področji humanistike in družboslovja, kot so

gledališče, glasba, zgodovina in medkulturne študije. Objavil je več znanstvenih člankov in poglavij v monografijah z omenjenih področij. Je tudi urednik več monografij in tematskih številk znanstvenih revij, je vodil in sodeloval v več raziskovalnih projektih, redno pa se udeležuje tudi mednarodnih znanstvenih konferenc in drugih dogodkov v tujini in lokalnem okolju.

Tadej Todorović je asistent za jezikoslovje in filozofijo z doktoratom iz filozofije na Univerzi v Mariboru. Njegovo področje delovanja vključuje pragmatiko in filozofijo jezika, pri čemer ga zanima analiza govornih dejanj v pragmatiki, zlasti z vidika vljudnosti in interakcije. Še posebej se posveča analizi vpliva in pomena analitične filozofije jezika na pragmatiko in analizo govornih dejanj v literarnih delih, zlasti v drami. Tadej Todorović aktivno sodeluje v več raziskovalnih projektih, povezanih s podnebnimi spremembami, digitalizacijo in umetno inteligenco v izobraževanju. Objavil je več člankov v recenziranih revijah ter sodeloval na več mednarodnih konferencah in simpozijih. Je tudi urednik slovenske filozofske revije *Analiza* in tehnični urednik mednarodne filozofske revije *Acta Analytica*.

Mladen Borovič je asistent in raziskovalec na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Njegova raziskovalna področja so aplikativna umetna inteligenca, iskalniki in priporočilni sistemi, obdelava naravnega jezika, detekcija podobnih vsebin, visokozmogljivo računalništvo in generativna umetna inteligenca. Aktivno deluje na projektih povezanih z obdelavo naravnega jezika in uporabo umetne inteligence v obliki pogovornih sistemov. Sodeloval je pri odmevnejših projektih kot so Nacionalna infrastruktura odprtega dostopa, Slovenščina na dlani, HPC RIVR in Razvoj slovenščine v digitalnem okolju. V sklopu svojega pedagoškega dela poučuje predmete povezane z računalniškimi omrežji, visokozmogljivim računalništvom in umetno inteligenco. Je tudi recenzent za več znanstvenih revij s faktorjem vpliva na področju računalništva.