Proscovia Svard[1]
Sanja Seljan[2]

THE USE OF LANGUAGE MODELS (LLMS) IN THE PUBLIC SECTOR AND THE IMPACT ON PUBLIC RECORDS: A CASE OF SWEDEN AND CROATIA

Abstract

Purpose: *This study is part of a bigger study being conducted under the auspices of the InterPARES project (https://interparestrustai.org/). Its purpose is to investigate the use of large language models (LLMs) in the public sector and the impact they have on the management of public records. The investigation was conducted in a Swedish and Croatian research settings. LLMs rely on substantial amounts of data from the Internet that might include a risk of generating inaccurate responses and biases. Since LLMs are used in decision-making processes, it is crucial that the records that derive are identified and managed for accountability and transparency purposes. The questions therefore records managers and archivists should be asking during the implementation of LLMs in government administrations is what type of critical public records can be identified in processes where they have been deployed and additionally, in case of wrong responses, how can transparency and accountability be maintained?*

Method/approach: *The authors have applied a case study method, interviews, and literature review to the investigation.*

Results: *The results reveal that despite the current deployment of LLMs in the public sector, little attention is paid to the identification and capture of public records. The focus is mostly put on the efficiencies that the LLMs are to bring, through the claimed lessened administrative burdens. Further the testing that was done of a chatbot demonstrated the challenges associated with ensuring the reliability and accuracy of information generated by LLMs.*

Conclusions/findings*: It is through the public records that the transparency of processes where the LLMs are being deployed can be maintained and where things go*

[1] Proscovia Svard, Associate Professor, History Department, Sorbonne University Abu Dhabi, email: Proscovia.Svard@Sorbonne.ae.

[2] Sanja Seljan, Professor, Faculty of Humanities and Social Sciences · Information and Communication Sciences University of Zagreb, Croatia, email: sanja.seljan@ffzg.hr.

wrong, those accountable can be held responsible. Despite the many opportunities that the LLMs are providing, it is equally of paramount importance to understand record making amidst the deployment of LLMs in government administrations.

Keywords: *Large Language Models (LLMs), Implementation, Public Records, Government Administrations, Transparency and Accountability*

L'USO DEI MODELLI LINGUISTICI (LLM) NEL SETTORE PUBBLICO E L'IMPATTO SUI REGISTRI PUBBLICI: UN CASO DI SVEZIA E CROAZIA

Abstract

Scopo: *questo studio fa parte di uno studio più ampio condotto sotto gli auspici del progetto InterPARES (https://interparetrustai.org/). Il suo scopo è quello di indagare l'uso dei grandi modelli linguistici (LLM) nel settore pubblico e l'impatto che ha sulla gestione dei registri pubblici. L'indagine è condotta in un contesto di ricerca svedese e croato. Gli LLM si basano su grandi quantità di dati da Internet che potrebbero includere un rischio di generare risposte imprecise e distorsioni. Le domande che gli archivisti dovrebbero porsi durante l'implementazione degli LLM nelle amministrazioni governative sono: che tipo di documenti pubblici critici possono essere identificati nelle aree in cui vengono implementati gli LLM e, inoltre, in caso di risposte errate, come possono essere mantenute la trasparenza e la responsabilità?*

Metodo/approccio: *gli autori hanno applicato un metodo di studio di caso, interviste e revisione della letteratura all'indagine.*

Risultati: *I risultati rivelano che, nonostante l'attuale impiego dei modelli linguistici di grandi dimensioni (LLM) nel settore pubblico, viene prestata poca attenzione all'identificazione e alla registrazione dei documenti pubblici. L'attenzione si concentra principalmente sulle efficienze che gli LLM dovrebbero apportare, in particolare attraverso la presunta riduzione degli oneri amministrativi. Inoltre, i test effettuati su un chatbot hanno dimostrato le sfide associate a garantire l'affidabilità e l'accuratezza delle informazioni generate dagli LLM.*

Conclusioni/risultati: *è attraverso i documenti pubblici che è possibile mantenere la trasparenza dei processi in cui vengono implementati gli LLM e, quando le cose vanno male, i responsabili possono essere ritenuti responsabili. Nonos-*

tante le numerose opportunità che gli LLM stanno offrendo, è ugualmente di fondamentale importanza comprendere la creazione di documenti durante l'implementazione degli LLM nelle amministrazioni governative.

Parole chiave: *Modelli linguistici di grandi dimensioni (LLM), implementazione, registri pubblici, amministrazioni governative, trasparenza e responsabilità.*

UPORABA JEZIKOVNIH MODELOV (LLM) V JAVNEM SEKTORJU IN VPLIV NA JAVNE EVIDENCE: PRIMER ŠVEDSKE IN HRVAŠKE

Izvleček

Namen: *Ta študija je del večje študije, ki poteka pod okriljem projekta InterPARES. Njegov namen je raziskati uporabo velikih jezikovnih modelov (LLM) v javnem sektorju in vpliv, ki ga imajo na upravljanje javnih evidenc. Preiskava poteka v švedskem in hrvaškem raziskovalnem okolju. LLM se zanašajo na velike količine podatkov iz interneta, ki lahko vključujejo tveganje ustvarjanja netočnih odgovorov in pristranskosti. Vprašanja, ki bi si jih arhivisti morali zastaviti med uvedbo LLM v javni upravi, so, katere vrste kritičnih javnih evidenc je mogoče prepoznati na območjih, kjer se uporabljajo LLM, in poleg tega, v primeru napačnih odgovorov, kako je mogoče ohraniti preglednost in odgovornost?*

Metoda/pristop: *Avtorici sta v raziskavi uporabila metodo študije primera, intervjuje in pregled literature.*

Rezultati: *Rezultati razkrivajo, da se kljub trenutni uvedbi LLM v javnem sektorju le malo pozornosti namenja identifikaciji in zajemanju javnih evidenc. Poudarek je predvsem na učinkovitosti, ki jo prinašajo z zmanjšanimi administrativnimi bremeni. Poleg tega je testiranje klepetalnega robota pokazalo izzive, povezane z zagotavljanjem zanesljivosti in točnosti informacij, ki jih generirajo LLM.*

Sklepi/ugotovitve: *Z javnimi evidencami je mogoče vzdrževati preglednost procesov, v katerih se uvajajo LLM-ji, in če gredo stvari narobe, se lahko od odgovornih zahteva odgovornost. Kljub številnim priložnostim, ki jih ponujajo LLM-ji, je prav tako izjemno pomembno razumeti izdelavo zapisov med uvajanjem LLM-jev v javni upravi.*

Ključne besede: *veliki jezikovni modeli (LLM), implementacija, javne evidence, javna uprava, preglednost in odgovornost.*

INTRODUCTION

Large Language Models (LLMs) are claimed to present unparalleled capabilities for numerous sectors such as business, communication, education, and government operations. These models offer a wide array of applications such as question-answering through chatbot interactions, summarization, translation, text generation, content creation, customer support, research, educational assistance, to mention but a few. The models use natural language processing algorithms to interpret and respond to text based human input. Governments are harnessing LLMs to deliver efficient public services and to ensure equal treatment, security, and privacy of citizens (Fang & Xu, 2023). Jingfeng et. al. (2024) investigated the practical aspects of LLMs in the real world and argued that their use requires an understanding of their limitations and capabilities, the quality of the data on which they are trained, the tasks they are meant to solve and knowledge regarding how to select the most suitable LLM. They confirmed that currently, there is no agreed-on definition of LLMs and posited that the end-users need to factor in their size, the computational requirements and the availability of domain-specific pre-trained models.

According to the Analysis and Research Team of the Council of the European Union, LLMs are being implemented in the public sector and will only continue to grow in sophistication (Council of the European Union, 2023). AI driven technologies are being incorporated in public administrations to substitute human intellectual work. This is because, they have the potential to analyze large amounts of data at a fast speed and accurately and hence deliver informed decisions. It is argued that this reduces bias and promotes fair and consistent outcomes. However, the challenge they pose is lack of transparency and yet, citizens who receive AI-based decisions need to understand how they were arrived at. Those who receive negative decisions have a right to know whether the decision was based on accurate information. AI technologies are also said to perpetuate and amplify biases and therefore to avoid such situations, public administrations need to develop and implement a framework and safeguards for the use of AI technologies (Carullo, 2023).

Though LLMs are praised for their innovative capabilities, scholars such as Weidinger et. al. (2021) argued that they equally pose risks that must be identified. They identified fairness and toxicity that emanate from unfair discrimination through stereotypes and social biases, risks related to private data because of the LLMs' capability to make inferences based on the private data in the training corpus. Furthermore, the LLMs can give misleading information and create less informed users which could erode trust in shared information. The authors argue that this could result in misinformation capable of harming all aspects of life. They further highlighted the possibility for misuse of LLMs by unscrupulous users with the intent to harm others. This could manifest in form of personalized scams, computer codes for viruses or weapon systems. The training and operation of LLMs also has a high environmental cost and yet, the applications that are based on them are only available to some groups of people, because the models are still inaccessible to many. They therefore concluded, that to measure and to mitigate the ethical and social risks that emanate from LLMs, will require collaboration, a broad range of expertise and inclusion of the affected communities.

Nashwan & AbuJaber (2023) on the other hand highlighted the potential benefits LLMs can bring to the management of e-health records. They argued that LLMs can revolutionalize healthcare practices by streamlining data-input processes, expedite information extraction from the patients' unstructured narratives and personalize communication. However, they also raised concerns regarding a patient's privacy, data security and biases which can hinder the delivery of equitable care.

It has not been easy to identify literature that has focused on public records management amidst the deployment of LLMs in the public sector, which has motivated the authors to undertake this study. We would like to understand what type of records grow in areas where LLMs have been deployed in the public sector and how they are captured to enhance transparency and accountability of government institutions. Though LLMs are delivering unprecedented advantages to society, the risks they pose are real. Therefore, we see this development as a rich ground for records managers and archivists to engage in, to defend the transparency and accountability of public administrations since it is underpinned by effective public records management. Records managers and archivists need to understand this development to be able to identify public records that should serve as evidence, especially in situations where citizens have been unfairly treated or hurt. The article offers an introduction, a literature review, a method, research findings that problematizes the issue of implementing LLMs in the public sector, findings, a discussion and conclusion and a summary.

THE RESEARCH PROBLEM

Large Language Models (LLMs) are being deployed in the public sector because of the claims that they enhance efficiency, reduce the administrative burden, deliver less biased decisions and hence equal treatment of citizens and their capability to deal with mass cases to mention but a few of the advantages (Fang & Xu, 2023). There are however also claims that counter the above stipulated advantages such as data privacy violations, biases, lack of representation and promotion of stereotypes, inaccurate information delivery and lack of transparency in the way they execute tasks (Weidinger et. al. 2021). The General Secretariat of the Council of the European Union argued that while it might be easy to understand how LLMs work, companies that develop them and mostly US based, are not willing to release detailed information and the parameters upon which they base the responses they provide to their users. This complicates their assessment once integrated in a workplace (Council of European Union, 2023). Public institutions are not for-profit organizations and are stringently held by transparency and accountability requirements, in most democratic societies. They must have safeguards against unfair treatment of the citizens (Carullo, 2023). Therefore, the implementation of new technologies requires that the open structure of public institutions is maintained. This is what makes it of paramount importance for the public sector records managers and archivists to understand what type of public records grow out of processes where LLMs have been deployed and to effectively capture them. This is to enhance democratic values of accountability and transparency, impartiality and reliability especially where decision making processes are concerned. Citizens have a right to know how decisions are arrived at, and this is through access to public records. Though big claims are being made about the many advantages that the LLMs bring to the public sector, there is less discussion on the management of public records which underpin the government administrations' open structure (Council of European Union, 2023).

LITERATURE REVIEW

We searched for articles related to the subject under investigation using words such as; 'large language models and the public sector,' 'LLM implementation in the public sector and public records' and 'challenges and opportunities of LLMs.'

We consulted databases such as Emerald, Science Direct and Google scholar. The literature search revealed that even though there is emerging literature on the deployment of Large Language Models (LLMs) in the public sector, we have not been able to find studies that are focusing on their impact on the management of public records.

LARGE LANGUAGE MODELS AND PUBLIC INSTITUTIONS

Recent research explores the potential of Large Language Models (LLMs) deployment in government administrations and highlights both opportunities and challenges. LLMs can enhance public services by automating responses to citizen inquiries, improving efficiency and accuracy in digital governance (Fang & Xu, 2023). However, the use of LLMs in government contexts raises significant security and privacy concerns, that include vulnerabilities such as data poisoning, and personally identifiable information leakage (Das et al., 2024). While LLMs can enhance code security and data privacy protection, they may also be exploited for malicious purposes due to their advanced reasoning capabilities (Yao et al., 2023). As governments increasingly adopt LLMs, addressing these security and privacy challenges become crucial to ensure the responsible and effective use of this technology in public service delivery. Das set al. (2024) reviewed the application-based risks of LLMs in various domains, such as transportation, education, and healthcare. They identified research gaps in the domain of LLM security and privacy and highlighted the need for more research in this direction.

Nagoudi et. al (2024), presented work that is being conducted under the auspices of the InterPARES TrustAI project: https://interparestrustai.org/ in the UNESCO SCEaR newsletter where they argued that, LLMs such as ChatGPT are offering unprecedented capabilities to generate text and understandability. They however warned that the models still pose a challenge of delivering inaccurate or fabricated information. They informed that a solution to this challenge and that could also be relevant to the archives domain is called Retrieval Augmented Generation (RAG). This solution now exists and enables the models to retrieve information from various external databases. They proposed a RAG model that is open source and that would ensure a robust and diverse knowledge base. RAG would enhance the reliability of LLMs by ensuring that output is based on verified data and would foster a trustworthy digital information environment. Papageorgiou et al. (2024) also proposed a modular framework using Retrieval-Augmented Generation (RAG) for integrating LLMs into e-government systems. Their approach aims to improve scalability and transparency, though it highlights the need for further empirical validation to optimize e-government service delivery. Like Nagoudi et. al. (2024), Ghodratnama (2024) who explored methodologies to optimize information retrieval was equally worried about the issue of inaccurate and misinterpreted data.

Raeini (2023) investigated privacy-preserving techniques that could be used to transform LLMs to preserve the privacy and security of data and users. LLMs have caused huge concerns related to transparency and privacy issues, which are discussed in large corporations and research centres. The Center for Democracy and Technology (2023) explored how LLMs can enhance government chatbots by simplifying legal and technical documents and discussed challenges like maintaining accuracy and the privacy risks associated with these systems. Oxford Insights (2023) in their report *LLMs in Government: Brainstorming Applications,* discussed potential applications of LLMs, such as improving procurement processes and citizen services. However, they cautioned against risks such as lack of transparency, environmental impacts, and the potential for hallucinations in high-stake decision-making.

The research presented by Fang and Xu (2023) presented the needs of government institutions that have to manage large volumes and diverse citizens' inquiries, which are primarily executed by human agents, with limited AI assistance. They argued that LLMs could be applied to address citizens' requests and to generate human-like answers. They however stated that, LLMs are general-purpose GPT models with limited understanding of professional expressions in the government domain and are unable to effectively respond like public officials. This research aimed to build a question and answer guidance system, specialized in government affairs based on LLMs and historical citizen question vector databases. Antonidas et al. (2021) used AI models to provide citizens with accurate, personalized and accessible information on Public Services (PS). For this purpose, they developed a chatbot named PassBot that gives personalized information on getting a Greek passport. Pena et al. (2024) used LLMs classify public documents, and analyse them which is crucial to promoting transparency and accountability.

Guererro et. al. (2024) investigated whether the exisitng AI regulations in Sweden, Finland and South African covered the management of public records. They concluded that there was a disconnect between AI regulations and the management of public records. Their study confirmed, that the focus was on the management of data and not records. They however highlighted the advantages AI technologies could bring to the records management field through enhanced efficiency and organization of information (classification, indexing and tagging of records), improved search and retrieval possibilities, increased transparency and accountability because even sensitive information could be redacted. This development would promote data-driven insights and innovation. They however recommended that governments need to invest in education, to create a workforce that will be equipped with the skills to embrace the opportunities that AI technologies offer but also deal with the challenges that they pose.

E-GOVERNMENT AND CHATBOTS

Seljan et al. (2020) examined transformations that encompass not only differences in ICT usage but also social, economic, educational, and business changes, which contribute to the risks associated with the "digital divide." One such development is e-Government, which has led to an ever evolving digital landscape. Governments now face new challenges as they strive to provide enhanced public services and empower citizens. The goal of e-Government is to streamline the delivery of information and services, improve efficiency and productivity, reduce costs, and enable greater citizen engagement in public policy decision-making. The interaction between government and users can include accessing information, completing forms, making payments, submitting online comments, and more.

Palvia & Sharma (2007) categorized e-government development through levels 1 to 5. Level 1 represented the most basic form, where it is limited to simple websites providing links to relevant institutions. At Level 2, the government offered an enhanced presence, providing a greater number of resources to the public, such as laws, regulations, policies, newsletters, and downloadable databases. Level 3 featured further improvement, with online services like downloadable forms for tasks such as tax payments and license renewals. Level 4 introduced transactional presence, enabling two-way communication between the government and citizens, allowing services such as online tax payments and applications for birth certificates or ID cards. Finally, Level 5 represents networked presence, where services are fully integrated, encouraging societal participation in open, constructive dialogue through online forms, comments, and other interactive features. Chatbots could be harnessed at this level, due to their potential to deliver responses to questions.

The use of chatbots to provide answers in a wide range of everyday situations has become a research topic for numerous researchers. Wangsa et al. (2024) conducted research on the use of five chatbots (ChatGPT, Bard, Llama, Ernie, and Grok) in education and healthcare, and underlined the importance of the context. The researchers used the criteria of accuracy, consistency, domain expertise, computational efficiency and scope of knowledge for evaluation, showing the strong performance of ChatGPT, followed by Bard and the Llama. Bahak et al. (2023) focused on question-answering evaluation of ChatGPT, using precision, recall, and F1 scores to measure performance and human evaluation for robustness and fluency. They found that while ChatGPT excels in conversational fluency, challenges remain in ensuring accuracy, confirming challenges in task-specific issues. Guo and Dong (2024) explored the adoption of chatbots for government use, outlining the main challenges, such as managing data and ensuring consistent, accurate responses, but also the importance of user expectations. In the context of e-government, there are also efforts to develop chatbot solutions that provide citizens with streamlined access to government services and information. These chatbots, like those in India's Digital India initiative, aim to centralize and simplify information from various agencies, making services more accessible (Kumar et al., 2024)

THE METHOD

This study builds on a bigger study that is being conducted under the auspices of the InterPARES project (https://interparestrustai.org/) and that focuses on ' Records Management Challenges Amidst AI Deployment in the Public Sector: The Case of Sweden, Finland and South Africa. The Croatian case is an addition to the study. The authors have employed the case study method because Creswell (2007) argued that it is an appropriate method, when an inquirer seeks an in-depth understanding of a phenomenon. Case studies offer a variety of evidence through interviews, documents, artefacts and observations (Yin, 2009). It is through data collection that a researcher can give a detailed description of the case being studied. Patton (2002) argued that researchers interview people because we cannot observe everything. Interviews therefore allow us to enter into other people's worlds. Interviews were carried out via Teams with an archivist and information security officer in the Swedish municipality hereto referred to as Municipality C, because they were the ones behind the implementation of a project that was relevant to the phenomenon under investigation that is, the Large Language Models (LLMs) implementation in the public sector. The interviews were carried out in December 2023 for the bigger InterPARES study mentioned above. The data presented in this case is therefore based only on the two interviews that were conducted with the archivist and information security officer who were in charge of the creation of a project that aimed to implement a ChatGPT-like model, to address the administrative burden. By the time of the interviews, the project was to be presented to the management to secure funding. The interviews were transcribed, analyzed and the relevant extractions from the gathered data are presented under the findings.

The second case looked at the implementation of a Chatbot in a Croatian setting. This research investigated the accuracy, reliability, and overall performance of responses generated by a general-domain chatbot, focusing on questions relevant to citizens of the Republic of Croatia. The evaluation was based on principles from the Archival Science domain, particularly in the areas of accessibility, verification, and trustworthiness of information sources. Emphasis was on the critical role of authoritative sources, such as:

- Official bank websites,
- The Croatian Institute for Health Insurance (HZZO),
- Faculty websites,
- Public transportation schedules (e.g., night tram schedules),
- The Ministry of Internal Affairs,

The chatbot responses were evaluated using a multi-faceted framework:

1. Accuracy: This measures whether the information provided is fully correct, partially correct, or inaccurate. Each response was cross-verified against reference sources. Accurate information was essential to determine the chatbot's utility, as prior research showed that chatbots often struggled with fact-based queries.

2. Reliability: This evaluated the credibility of the references cited or inferred by the chatbot. The criterion drew upon archival theory, where the provenance and authenticity of sources are paramount. For instance, information provided by public institutions (such as banks and HZZO) was regarded as reliable, whereas answers were not accurate
3. Efficiency: This pertained to the speed with which the chatbot retrieves and delivered information. Studies suggest that users place high value on response time in practical contexts, such as when accessing government services online.
4. Relevance: This assessed how closely the chatbot's answer aligned with the posed question, ensuring the response was both on-topic and specific. Relevance is often correlated with user satisfaction, especially when addressing legal, health, or transportation queries
5. Fluency: The linguistic quality of the chatbot's output, focusing on the clarity, coherence, and naturalness of the language used. A fluent response is essential for maintaining user understanding.

The study used a three-level rating system:

Positive: The response was accurate and complete.

Partially correct: The response contained correct elements but lacked full accuracy or completeness.

Incorrect: The information provided was wrong, non-existent, or irrelevant

The overall objective was to highlight the necessity for records managers and archivists to engage in these spaces where LLMs are being deloyed to effectively identify and manage public records because they are critical to decision-making processes, accountability and transparency of government institutions.

RESEARCH FINDINGS

The section below presents research findings from two cases: a Swedish and a Croatian research setting.

CASE 1 - THE USE OF LLMS IN A MUNICIPALITY – PROJECT UNDER CONCEPTION

The Archivist of Municipality C informed about a project was being conceived but had not yet been presented to the management. It was about an AI Generator likened to a ChatGPT that was to be suggested, and discussions were on-going with the supplier. The intent was to gather documents that were critical to the daily operations of the organization such as; steering documents to facilitate the administrative burden. The archivist did not have full knowledge of the AI generator but was to a certain degree participating in the conception of the project. He explained that the AI generator was to be fed with documents to enable it to generate answers. When the archivist was asked to share what he saw as a challenge with the project, he informed that, it would be the generation of wrong responses. This required an administrator using the ChatGPT to be critical.

The Information Security Officer in Municipality C was responsible for data protection. He pinpointed the importance of maintaining trustworthy and authentic information. He informed that he was together with the Archivist planning to present the conceived project to the municipality's management. An IT company was helping them to create an AI based administrative support function that is likened to ChatGPT 3.5-4.0. The company was using language models – machine learning models and selling licences to those interested in the support function. The language model was to be used as a base infrastructure upon which they could add their own data. Since the municipality was striving to facilitate the daily administrative routines for the employees, they were to start with the steering documents, guidelines for archives management, archival descriptions and later if the project was accepted, they would advance to the laws governing different areas and hence create their own ChatGPT. The system must be fed with the data that the municipality decided on. ChatGPT was supposed to help the administrators to get responses to questions that relate to their processes. He informed that through reading the responses from ChatGPT he as an information security officer could easily identify areas for improvement. He argued that this could also serve as support for the authors of the steering documents – to improve language and understanding. The support function would easily be accessible to the employees in their daily work.

The Security Officer however informed that one needs to include a disclaimer and reiterated what the archivist had also mentioned that the administrators need to embrace critical thinking when reading through ChatGPT responses. The administrators needed to deepen their understanding by reading the documents that ChatGPT based its responses on and had to consider the support function as guidance and not the absolute solution. Reading from the information that was being shared, one could not stop thinking about the efficiencies the solution was to bring but also the increased stress levels depending on the type of process an administrator was involved in. We also understood that the employees needed to be informed and educated about how the support function would work. The trustworthiness of the responses from the AI Generator function would require that it is only authorized that would upload documents upon which the responses would be based, to avoid erroneous responses. The language model had to be secured to avoid any data leakage – it had to be within the Swedish borders. This was why municipalities were paying licenses for this function to be able to access a language model to which they could add their own data. They also needed to ensure the right to own their information. The disadvantage of this development was the generation of wrong responses where wrong information was fed to the language model. The language model also needed to be trained on the Swedish language. Issues of Provenance and authenticity of the documents uploaded had to be clear. The Security Officer argued that it would be good to use a Swedish language model for all public sector organisations as a base infrastructure.

CASE 2 – VERIFICATION AND RELIABILITY OF INFORMATION SOURCES FROM A CHATBOT

This case revealed inconsistencies in the chatbot's ability to provide correct information. Out of five responses, two answeres were partly correct, but matched authoritative sources, such as banks and Croatian Health Insurance Fund. Two answers related to faculty enrollement criteria and public transport schedule were correct but did not provide links to exact web-sites containg the provided answer, while one answer related to Ministry of interior affairs was incorrect and the chatbot's link led to a non-authoritative source, highlighting a gap in source reliability. This finding aligns with studies that show chatbots often struggle to maintain updated knowledge, particularly when web links are involved. Despite these accuracy and relevance issues, the chatbot's performance was strong in areas of efficiency, relevance, and fluency. All responses were delivered promptly and in clear, grammatically sound language, demonstrating a high level of usability in terms of user experience. However, the reliability of some information was questionable, especially where certain links did not lead to trustworthy or expected sources. The analysis underscores the importance of integrating more robust verification mechanisms into chatbot systems, which aligns with results of Bahak et al. (2023). Ensuring that responses align with trusted or authoritative sources is critical for maintaining the reliability. As suggested by prior research, combining real-time data retrieval with a strong verification process could mitigate some of the issues observed, particularly around the use of outdated or incorrect links

DISCUSSION AND CONCLUSION

The potential of Large Language Models (LLMs) has been highlighted by some of the reviewed researchers. Governments face new challenges in delivering enhanced public services, aiming to reduce administrative burdens, and empowering citizens. Large Language Models (LLMs) present significant potential for improving the efficiency of public services, promoting equal treatment, and safeguarding the security and privacy of citizens. Specifically, LLMs can be deployed as chatbots to provide rapid, informative responses, thereby reducing administrative workloads. However, existing research underscores the need for specialized domain training to ensure the accuracy and reliability of the information provided, but also raise question of verification and reliability of information sources.

Since LLMs are currently being deployed in the public sector, it is crucial that government administrations equally have in place, safe guards where citizens are unfairly treated. This could be in form of unfair decisions. Citizens have a right to understand how such decisions are arrived and decisions have to be based on accurate, complete and trustworthy information. The Croatian case study demonstrated that given the vast amount of outdated or erroneous information on the web, there is a risk that LLMs may generate incorrect or uncomplete responses. However, in such cases, the question on responsibility for wrong answers remains open. The deployment of LLMs in the public sector has to consider that AI-generated information has to be reliable and trustworthy. This should be the basis upon which AI-decisions should be made. Traditionally, what has underpinned the open governance structure of government administrations has been public records. Records managers and archivists need to understand what type of records are growing in these new spaces and how can they be managed. This is what will promote accountability and transparency amidst the deployment of LLMs.

LLMs indeed can enable the delivery of efficient services but since the way they operate is not fully understood and companies behind their developments are not that transparent, then they need to be carefully implemented in the public sector. We understand that the ultimate goal is to enhance efficiency and productivity, reduce operational costs, and enable greater citizen engagement in public policy and decision-making processes but, the negative side of the LLMs need to be well studied to understand the impact they can have on the management of public reords and the citizens. Therefore, the use of LLMs in the government administrations still raises critical concerns related to privacy, security, data protection, fair decion-making processes, malicious misuse, the hindrance of accountability and transparency. Ensuring the trustworthiness and reliability of LLMs is crucial to maintaining public confidence in AI-driven public sector services.

REFERENCES

Antoniadis, P. and Tambouris, E. (2021). PassBot: A chatbot for providing information on Getting a Greek Passport. In M. A. Macadar, M. Meyerhoff Nielsen and M. Peixoto (eds.), *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance (ICEGOV 21),* (pgs. 292–297). Retrieved at https://dl.acm.org/doi/10.1145/3494193.3494233 (accessed on 20. 10. 2024).

Bahak, H., Taheri, F., Zojaji, Z., and Kazemi, A. (11. 12. 2023). Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models. *ArXiv*. Retrieved at https://arxiv.org/abs/2312.07592 (accessed on 20. 10. 2024).

Bateyko, D. (13. 12. 2023). Let LLMs Do the Talking? Generative AI Issues in Government Chatbots. *Center for Democrarcy and Technology*. Retrieved at https://cdt.org/insights/let-llms-do-the-talking-generative-ai-issues-in-government-chatbots/ (accessed on 20. 10. 2024).

Berman, A., de Fine Licht, K. and Carlsson, V. (2024). Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system. *Technology in society 76*. Retrieved at https://www.sciencedirect.com/science/article/pii/S0160791X24000198?via%3Dihub (accessed on 20. 10. 2024).

Carullo, G. (2023). Large Language Models for Transparent and Intelligible AI-Assisted Public Decision-Making Gherardo Carullo. *CERIDAP, 3*, 1–24. Retrieved at https://air.unimi.it/bitstream/2434/1007135/2/Carullo%20-%20Large%20Language%20Models%20for%20Transparent%20and%20Intellig.pdf (accessed on 20. 10. 2024).

Council of the European Union, General Secretariat. (24. 4. 2023). *ChatGPT in the Public Sectpr – overhyped or overlooked?* Retrieved at https://www.consilium.europa.eu/media/63818/art-paper-chatgpt-in-the-public-sector-overhyped-or-overlooked-24-april-2023_ext.pdf (accessed on 12. 9. 2024).

Das, B. C., Amini, M. H., and Wu, Y. (30. 1. 2024). Security and Privacy Challenges of Large Language Models: A Survey. *ArXiv*. Retrieved at https://arxiv.org/abs/2402.00888 (accessed on 12. 9. 2024).

Fang, K. and Xu, K. (2023). Automating Government Response to Citizens' Questions: A Large Language Model-Based Question-Answering Guidance Generation System. In *2023 3rd International Conference on Digital Society and Intelligent Systems (DSInS),* (pgs. 386–389). Chengdu, China: IEEE. Retrieved at 10.1109/DSInS60115.2023.10455136 (accessed on 12. 9. 2024).

Ghodratnama, S. and Zakershahrak, M. (2024). Adapting LLMs for Efficient, Personalized Information Retrieval: Methods and Implications. In F. Monti et al (eds.)*, Service-Oriented Computing – ICSOC 2023 Workshops. ICSOC 2023. Lecture Notes in Computer Science, vol 14518*. Singapore: Springer. Retrieved at https://doi.org/10.1007/978-981-97-0989-2_2 (accessed on 12. 9. 2024).

Guerrero, E., Svärd, P., Balugon, T., Saurombe, N., Hentonnen, P. and Jacobs, L. (2024). The Africa-Europe Gaps of AI Regulations for Managing Public Records. *SCEaR Newsletter Special Issue 2024*, 84–90. Retrieved at https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (accessed on 14. 9. 2024).

Guo, Y. and Dong, P. (2024). Factors Influencing User Favorability of Government Chatbots on Digital Government Interaction Platforms across Different Scenarios. *J. Theor. Appl. Electron. Commer. Res, 19*(2), 818–845. Retrieved at https://doi.org/10.3390/jtaer19020043 (accessed on 12. 9. 2024).

Jingfeng Y., Hongye J., Ruixiang T., Xiaotian H., Qizhang F., Haoming J., Shaochen Z., Bing Y. and Xia H. (26. 4. 2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data. 18*(6), Article 160, 1–32. Retrieved at https://doi.org/10.1145/3649506 (accessed on 12. 9. 2024).

Kumar, A., Rajpurohit, V., Sanjeev, Y. B. and Kumar, G. (2024). eGovernance Chatbot: Empowering Citizens through Chatbot Support. *Journal of Eerging Technologies and Innovative Research (JETIR)*, *11*(1). Retrieved at https://www.jetir.org/papers/JETIR2401169.pdf (accessed on 20. 10. 2024).

Nagoudi, B. E., Inciarte, A. A. and Muhammad A–M. (2024). Improving archives-Focused LLMs with Retrieval Augmented Generation. *SCEaR Newsletter Special Issue 2024*, 21–27. Retrieved at https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (accessed 14. 9. 2024).

Nashwan A. and AbuJaber A. A. (29. 7. 2023). Harnessing the Power of Large Language Models (LLMs) for Electronic Health Records (EHRs) Optimization. *Cureus, 15*(7). Retrieved at doi: 10.7759/cureus.42634 (accessed 14. 9. 2024).

Oxford Insights (19. 5. 2023). *LLMs in Government: Brainstorming Applications*. Retrieved at https://oxfordinsights.com/insights/llms-in-government-brainstorming-applications/ (accessed 14. 9. 2024).

Palvia, S. C. J. and Sharma, S. S. (2007). E-government and e-governance: definitions/domain framework and status around the world. *In 1st International Conference on Theory and Practice of Electronic Governance (ICEGOV2007) took place in Macau, China.* Retrieved at https://www.researchgate.net/figure/Palvia-and-Sharma-Framework-for-e-Government-versus-e-Governance_tbl1_268411808 (accessed 14. 9. 2024).

Papageorgiou, G., Sarlis, V., Maragoudakis, M. and Tjortjis, C. (2024). Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models. *Applied Sciences,14*(18), 8259. Retrieved at https://doi.org/10.3390/app14188259 (accessed on 12. 9. 2024).

Peña, A., Morales, A., Fierrez, J., Serna, I., Ortega-Garcia, J., Puente, I., Córdova, J. and Córdova, G. (15. 8. 2023). Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. In M. Coustaty and A. Fornés (eds), *Document Analysis and Recognition – ICDAR 2023 Workshops. ICDAR 2023. Lecture Notes in Computer Science, vol 14193*. Springer, Cham. Retrieved at https://link.springer.com/chapter/10.1007/978-3-031-41498-5_2 (accessed on 12. 9. 2024).

Patton, Q. M. (2002). *Qualitative Research & Evaluation Methods*. London: Sage Publications.

Raeini, M. (24. 7. 2023). Privacy-Preserving Large Language Models (PPLLMs). *SSRN*. Retrieved at http://dx.doi.org/10.2139/ssrn.4512071 (accessed on 14. 9. 2024).

Seljan, S., Miloloža, I. and Pejić Bach, M. (2020). E-government in European countries: gender and ageing digital divide. *Interdisciplinary management research, XVI* , 1581–1602.

Wangsa, K., Karim, S., Gide, E. and Elkhodr, M. A. (2024). Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok. *Future Internet 2024, 16*(7), 219. Retrieved at https://doi.org/10.3390/fi16070219 (accessed on 14. 9. 2024).

Weidinger, L., Mellor, J., Maribeth Rauh, M., Griÿn, C., Uesato, J., Huang, P-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton. Z, Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A, Isaac, W., Legassick, S., Irving, G., and Gabriel, I. (8. 12. 2021). Ethical and social risks of harm from Language Models. *arXiv.* Retrieved at https://arxiv.org/abs/2112.04359 (accessed on 14. 9. 2024).

Williamson, K. (2002). *Research methods for students, academics and professionals. Information management and systems*. Chandos Publishing.

Yin, K. R. (2009). *Case Study Research, Design and Methods, 4th ed*. Thousand Oaks, CA: Sage Publications.