AI-agent-based system for fact-checking support using large language models

Leonid Kupershtein[1], Oleksandr Zalepa[1], Volodymyr Sorokolit[1] and Serhii Prokopenko[2]

[1]*Vinnytsia National Technical University, 95 Khmelnytske Hwy., Vinnytsa, 21021, Ukraine*
[2]*Simon Kuznets Kharkiv National University of Econimics, 9A Nauky Ave., Kharkiv, 61166, Ukraine*

Abstract

In today's world, the problem of disinformation is becoming increasingly relevant due to the speed of information dissemination and the influence of social media. This article examines the impact of fake news on society and its political, economic and social consequences. Special attention is paid to the use of large language models (LLMs) to automate the fact-checking process. Describes the capabilities of LLMs in verifying information, including text analysis, comparison with reliable sources, and contextualization. At the same time, the risks of using LLMs to create fake news are highlighted. Proposes an architecture of an AI-based disinformation detection tool, which includes query processing modules, a database, work with web resources, and results analytics. This approach is aimed at improving the efficiency and accuracy of information verification.

Keywords

fake news, disinformation, fact-checking, large language model, AI-agent, RAG

1. Introduction

In today's world, the flow of information is constantly growing, and with it the number of unverified claims and myths. Social media and online platforms greatly facilitate this process, as each user can easily create and disseminate information. In the context of such a rapid spread of disinformation, the problem of identifying and refuting it becomes extremely important. Fake news, or deliberately fabricated false information, is used to mislead the audience or manipulate public opinion, which can have serious consequences for society. To combat disinformation, there is a need to develop effective tools for fact-checking.

Big language models represent one of the most advanced advances in artificial intelligence that can be used to automate fact-checking. These models are capable of analyzing and generating text with high accuracy, which makes them effective tools for verifying information. Using them for fact-checking will significantly improve the quality and speed of information verification, as well as integrate them into existing systems.

2. The impact of fakes on society

Studies show that fake news spreads faster and more widely than the truth due to the structure of social media, which rewards users for frequent sharing of information, regardless of its accuracy [1]. Fake news has a significant impact on society, causing a number of social, political, and economic consequences. Their harm can manifest itself in various forms and areas of human activity.

For example, the spread of false information about vaccination has led to growing doubts and fear among the population, which has complicated governments' efforts to fight epidemics. Fakes can also

CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine

kupershtein.lm@gmail.com (L. Kupershtein); sashasun2002@gmail.com (O. Zalepa); sorokolitvovan@gmail.com (V. Sorokolit); prokopenko.serhii@gmail.com (S. Prokopenko)
0000-0003-4712-3916 (L. Kupershtein); 0009-0008-2847-2006 (O. Zalepa); 0009-0001-9177-3054 (V. Sorokolit); 0000-0003-4712-3916 (S. Prokopenko)

contribute to the spread of social conflicts, enmity and hatred by deliberately manipulating people's emotions [2].

In the political sphere, fakes are used to discredit opponents or manipulate public opinion before elections. They can influence election results by distorting the electoral process. Falsification of information creates the illusion of distrust in decision-making processes and the government in general, which can lead to political instability [3].

Fakes also have the potential to affect the economy. For example, false reports on the financial condition of companies can seriously affect their stock prices and investment attractiveness. Such information attacks can lead to significant financial losses not only for investors, but also for ordinary consumers who may lose their jobs due to market instability. False information on the Internet about companies can cause significant fluctuations in the market, which negatively affects economic stability [4].

Disinformation is becoming an increasingly common problem in today's information environment, especially in the context of international politics. Ukraine is often one of the main targets of such attacks. According to the second report of the European External Action Service 'Manipulation of Foreign Information and the Threat of Interference', a significant number of disinformation cases were analyzed between December 2022 and November 2023 (table 1) [5].

Table 1
Analysis of recorded disinformation incidents.

Indicator	Number of cases	Share (%)
Incidents against Ukraine	160	21.3%
Incidents against the United States	58	7.7%
Incidents against Poland	33	4.4%
Incidents against Germany	25	4.1%
Incidents against France	25	3.3%
Attacks against the EU	142	19%
Attacks on NATO	113	15%
Attacks on the Armed Forces of Ukraine	105	14%
Attacks on the UN	22	3%

One of the most high-profile examples of disinformation that has had a significant impact on international politics and relations is the spread of fakes about events in Ukraine by Russia. Two of the most significant fakes were published in an article by the European Commission [6].

The first fake is that Ukraine and its government are the aggressor in relations with Russia. This false claim was used to justify Russia's military aggression against Ukraine, including the annexation of Crimea in 2014. Russian disinformation was aimed at convincing the international community and its citizens that Ukraine was allegedly provoking the conflict and posing a threat to Russia. However, the reality was that it was Russia that invaded Ukraine, violating its sovereignty and territorial integrity, as confirmed by numerous international agreements and legal documents.

The second fake is the genocide of the Russian-speaking population in Ukraine. This propaganda thesis has been used to justify Russia's criminal military actions in Ukraine, to strengthen support from the Russian population, and to create a negative image of Ukraine in the international arena. However, numerous international organizations, including the Council of Europe, have found no evidence of systematic violations of the rights of the Russian-speaking population in Ukraine. This myth has been dispelled as part of Russia's disinformation campaign aimed at destabilizing the situation in Ukraine and justifying its aggression.

These examples of disinformation have serious consequences not only for the domestic political situation in Ukraine, but also for international relations, as they contribute to the escalation of the conflict and influence international support and actions of other states towards the conflict. Distortion of facts and manipulation of information during military conflicts are among the most harmful aspects of modern information warfare.

Fact-checking is a critical process, as the speed of information dissemination can lead to the massive spread of disinformation. This process involves the use of various techniques and approaches that have evolved over time from traditional verification methods to modern technological solutions [7].

Initially, fact-checking depended mainly on manual verification of information, which required significant time and effort. Traditional methods included verification of primary sources of information, such as official documents, scientific studies or reports of independent organizations, cross-checking information from different sources, where data is confirmed through several independent and reliable resources, expert opinions from industry experts to analyze complex, or specialized topics that require deep knowledge in a particular area [8].

3. Possibilities of large language models for fact-checking

Large language models (LLMs) are one of the most advanced advances in artificial intelligence aimed at understanding and generating human speech. They are based on deep neural networks and use large amounts of text data for training, which allows them to perform complex language tasks with high accuracy. The basic concept of LLMs is to train the model to predict the next word in a sentence based on the context of the previous words.

LLMs can perform a wide range of tasks, namely [9]:

- generate text, which can be useful for writing articles, creating content, or even writing code;
- translate text from one language into another, taking into account the context and peculiarities of each language
- determine the emotional tone of the text, which can be useful for analyzing reviews, comments and other textual data;
- check information for accuracy, thanks to their ability to process large amounts of text and compare information from different sources.

LLMs have significant potential for use in fact-checking due to their ability to process and analyze large amounts of textual data. They can perform several key tasks that make them effective tools for verifying the accuracy of information [10]:

- quickly analyze textual materials, identifying potentially questionable or false statements;
- compare statements with available reliable sources of information;
- take into account the context in which the statement was made, which helps to avoid out-of-context quotes and manipulations;
- automate a part of the verification process, which significantly reduces the time and effort required to identify disinformation.

However, LLMs can be used not only for fact-checking, but also for creating fakes. They are capable of generating seemingly plausible text, which makes it quite easy to create articles, news and social media posts that can mislead readers. Such models can automatically formulate answers to questions by changing only key elements to create fake news that looks authentic. This process may include replacing facts or details in the original texts to form a new article containing disinformation [11].

For example, one way to create fake news is to first ask a question about a real news story, get an answer, and then modify that answer to create a new article. This article is then checked to see if the new answer matches the original question. If it differs significantly, the generated article is saved as fake news. This method allows you to create text that looks plausible but contains inaccurate information that can mislead readers [12].

Thus, while LLMs can be powerful tools for combating disinformation, they can also be used to create it, which underscores the importance of developing mechanisms for detecting fake news to ensure the information security of society.

Despite the potential harm, LLMs can also be used to detect text generated by different language models, based on stylistic and structural features of automated generation, which helps to identify fakes. They can be configured to identify texts that have such features. For example, language models often demonstrate certain patterns in word usage, parts of speech, and emotional tone, which can be detected by analyzing the syntactic and semantic features of a text. Studies show that machine-generated text has less vocabulary variation and is less emotionally charged than text written by humans [13].

Such methods include the use of pre-trained models that can distinguish between human-generated and LLM texts by analyzing word frequency, sentence length, syntactic structure, and other characteristics. Using these features, models can be trained on data containing both human and generated texts [14].

Several leading LLMs can be used for effective fact-checking, and they are highly effective due to their powerful algorithms and ability to process large amounts of textual data. A comparison of the technical characteristics and capabilities of the most advanced LLMs is shown in table 2 [15].

Table 2
Technical characteristics and capabilities of modern LLMs.

LLM	Context window size	Search for information on the Internet	Connect files	Support for RAG	Listing of found sources	Presenting an answer in a given form
GPT-3.5	16k tokens	No	Yes	Limited	No	Yes
GPT 4o	128k tokens	Yes	Yes	Yes	Yes	Yes
Llama 3.1	128k tokens	No	Yes	No	No	No
Claude 3	200k	No	Yes	Yes	Yes	Yes
Mistral	4k tokens	No	Yes	Yes	Yes	Yes
Gemini 1.5	1 million tokens	Yes	Yes	Yes	Yes	Yes

4. Architecture of AI-agent-based system for fact-checking

The architecture of an AI-based fact-checking system consists of several main modules that interact with each other to ensure efficient operation. Each module performs a specific function, and their coordinated interaction ensures high accuracy and speed of information verification. The main components of the tool's architecture are:

- user interface (UI);
- user request processing module;
- database;
- module for working with web resources;
- LLM;
- results analysis module.

Scheme of work an AI agent-based system to support fact-checking is shown in figure 1.

The interaction scheme of these components is shown in figure 1. The user query processing module is responsible for receiving queries from users through the interface, pre-processing them and passing them to the large language model for analysis. The module also provides feedback by returning the results of the check to users. Thanks to this module, the tool can quickly respond to queries, ensuring high accuracy and efficiency of fact-checking.

The database is the main repository of reliable sources of information or already verified facts used for verification. It contains a large amount of verified data that is constantly updated to ensure that the information is up-to-date. The vector database with a large language model implements the so-called RAG system. It allows quickly find and compare facts with existing ones, which ensures high accuracy and reliability of the results of checking those news items that have already been checked.

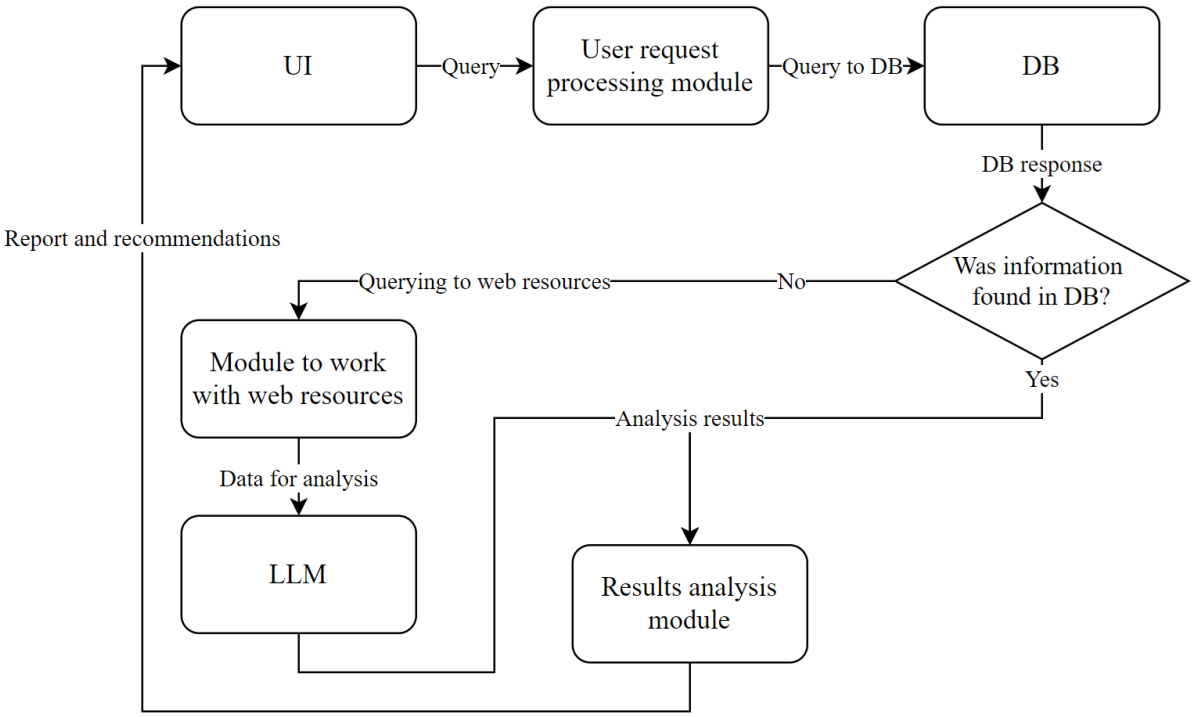

Figure 1: Architecture of AI-powered system for fact-checking support.

The web resources module is responsible for integrating the tool with external information sources. It makes queries to various web resources, extracting the necessary data for analysis and comparison if the data is not in the existing database. This module ensures that the application's checks are up-to-date, allowing it to use the latest data from the Internet. It also interacts with other modules of the system, transferring the received data for further processing and analysis.

LLMs are the key component of the application responsible for analyzing textual data and checking its validity. They use deep neural networks to understand and generate human speech, allowing them to effectively detect potentially false statements. LLMs will be used to parse the user's query and search for relevant information in the database. If the relevant data is not found in the database, the model converts the user's query into a search query for searching the Internet. After receiving the results, LLM analyzes extracted data and makes a decision on its reliability, determining whether the information is fake. This allows you to quickly and accurately verify a large amount of information, reducing the time and resources required for manual verification.

The results analysis module processes data obtained from a large language model and web resources, generates reports and provides recommendations on the reliability of information. It analyses the results of the check, determines the level of data reliability, and draws conclusions that are then provided to the user. The module can also detect patterns and trends in the data, which helps to identify new sources of disinformation and develop strategies to counter them. The module provides users with accurate and detailed verification results.

5. Development technologies and tools

To develop the fact-checking support system, we used modern tools that ensure efficiency, reliability and scalability. The main component is OpenAI's GPT-4o large language model, which is integrated through the OpenAI API. This choice was made due to GPT-4o's ability to quickly analyze large amounts of text, perform comparisons, take into account context, and provide accurate conclusions.

The system was implemented in the Python programming language, which provides flexibility and

a wide range of libraries for text processing, process automation, and neural networks. To process queries and retrieve relevant data from external sources, we used the google search and «serpapi» libraries. They provide integration with search engines, allowing us to collect relevant information for verification. The «htmldate» library was used to automatically determine the dates of webpage publications, which guarantees accurate extraction of chronological information for analysis.

The user interface was developed with the help of the Streamlit library, which allows creating interactive web applications without the need to use HTML, CSS or JavaScript. This greatly accelerated the implementation of a user-friendly interface for user interaction with the system.

Combining these tools with Python capabilities made it possible to implement the functionality of automatic text data verification, real-time fact analysis, and integration with other analytical systems. This ensured the creation of an efficient and reliable fact-checking system capable of meeting modern requirements.

6. Experiments

This section presents the results of experimental studies of a fake news detection software tool based on an AI agent system using large language model. The aim of the study is to evaluate the functionality, performance, and reliability of the developed system. The experiments were conducted in three stages: user interface research, testing of interaction with the database, and evaluation of information search on the Internet.

The first stage was to study the user interface to assess its intuitiveness, usability and correctness of interaction with the user. Figure 2 shows a view of the user interface.

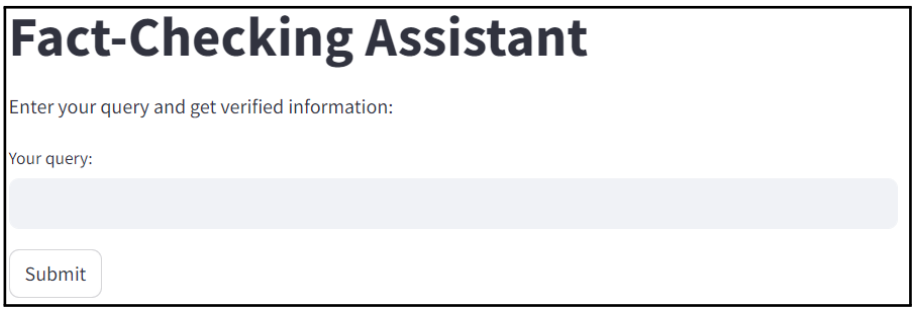

Figure 2: User interface design.

We checked the possibility of entering user queries and the correctness of the system's processing of the entered data. Figure 3 shows an example of a query used for this test. The results showed that the interface is intuitive and provides correct interaction with the user.

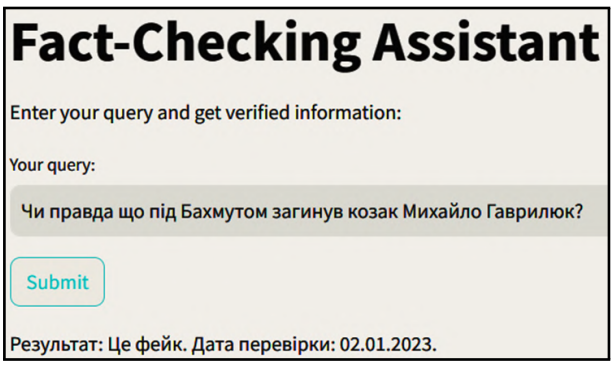

Figure 3: Test query view for checking the work user interface.

The second stage was to test the system's interaction with the verified facts database. The purpose

of this study was to make sure that the tool interacts correctly with the database and retrieves the necessary information. Figure 4 shows a fragment of the database.


```
{
    "fakeStatus": "false",
    "createdAt": "02.01.2023 14:45:32",
    "text": "NBSPПід Бахмутом загинув козак Михайло Гаврилюк",
    "id": "63b2ee0cceb6be2cd9eefd02"
},
```


Figure 4: View of the DB fragment.

A test was conducted with an initial query that matches the database record exactly. The results showed that the system successfully finds the relevant information. Next, the query was modified to be different from the database records (figure 5).

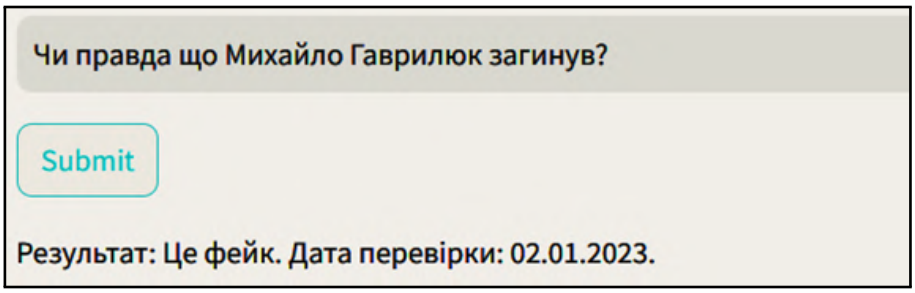

Figure 5: View of the modified query for testing work with DB.

The system still successfully identified the relevant information, demonstrating the ability to understand the context of the query and find the necessary data even with the changed wording.

The third stage involved testing the functionality of information search on the Internet. This allowed us to assess whether the program correctly forms queries to search engines, whether it can find relevant information and analyze it correctly.

First request: Verification of information about the introduction of electricity blackouts in Ukraine lasting up to 20 hours a day starting June 5, 2024. The results (figure 6) showed that the program correctly refuted this statement, providing a detailed conclusion with links to reliable sources.

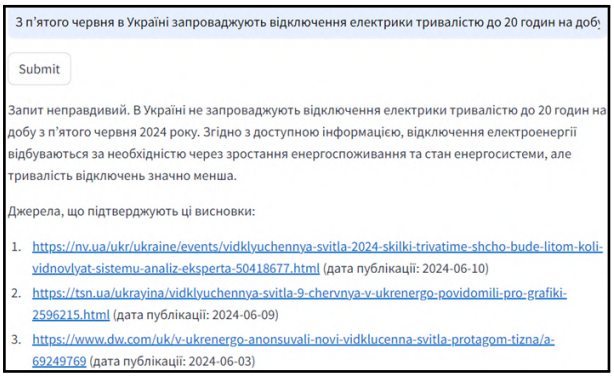

Figure 6: The first test query to check how the web search works.

Second query: Checking the information that English has become the second official language in Ukraine. The program refuted this claim, noting that English was granted the status of a language of international communication, but was not recognized as the state language (figure 7).

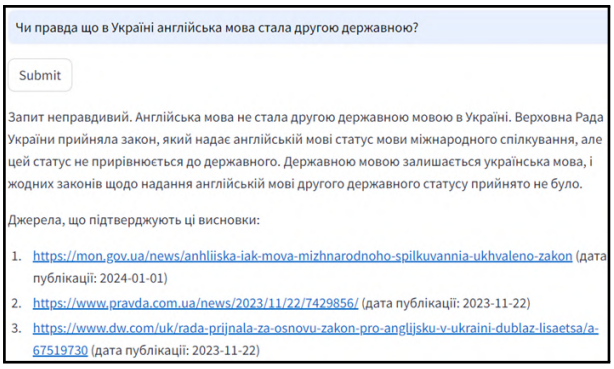

Figure 7: The second test query to check how the web search works.

Third request: Verification of information about the Kyiv Regional CCC's drawing of mobilization armor. Initially, the program gave a false answer, confirming this statement (figure 8). After a second request, the program gave the correct answer, refuting the fake and providing links to reliable sources (figure 8).

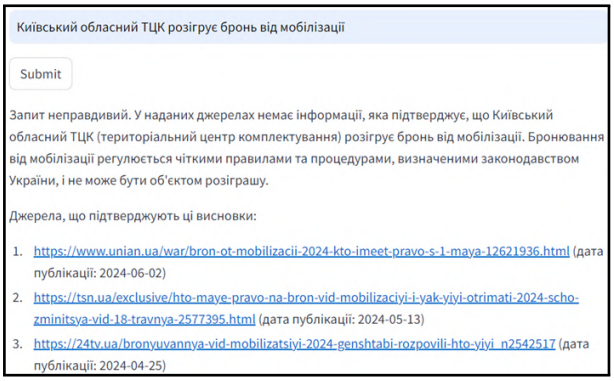

Figure 8: The third test query to check how the web search works.

The information checked in the test queries was also refuted by the fact-checking organization Gvara Media (figure 9). For example, they confirmed that the Kyiv Regional Military Commissariat does not sell reservations for mobilization, emphasizing that this information is fake.

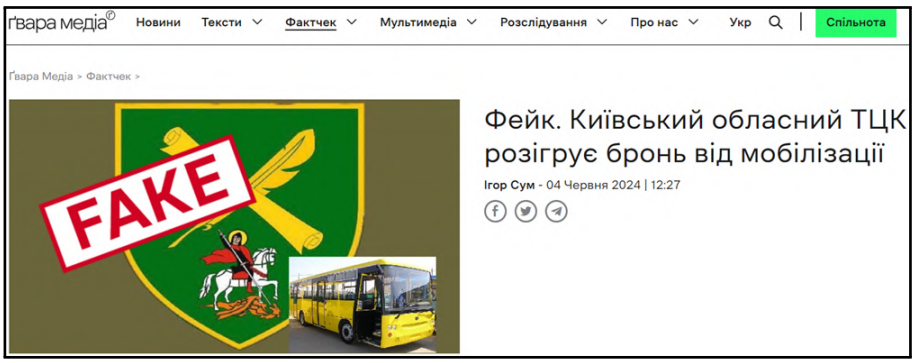

Figure 9: Refutation of the fake about the armor raffle from the mobilization by the fact-checking organization Gvara Media.

Figure 10 shows an erroneous result of the tool. The sources were found to be correct, but LLM gave an erroneous analysis result. This once again emphasizes the need for human control over the work of language models, as they can experience so-called "hallucinations".

After trying to make this test query a second time, the program gave the correct answer (figure 11).

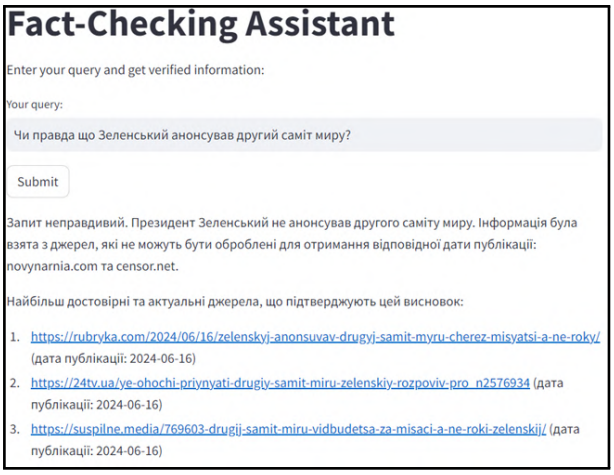

Figure 10: Incorrect response from the program.

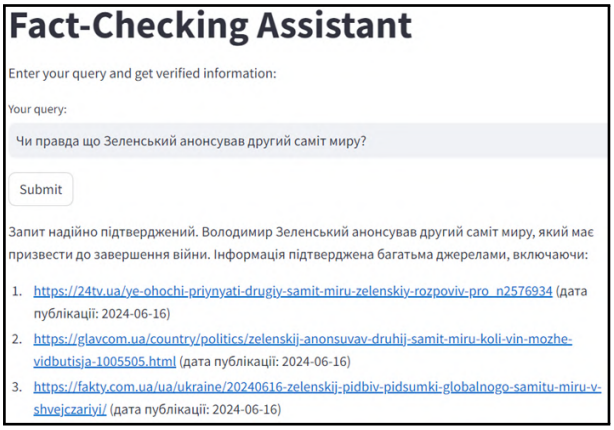

Figure 11: The correct answer of the program, after the second attempt of verification.

The results of the accuracy assessment show that the fake news detection tool has a high accuracy of 90%. The first and second types of errors amount to 10%, which indicates a relatively low level of false positives. The program's errors occurred due to the language model's "hallucinations" and because it sometimes received not the best sources for analysis. In general, the system demonstrates reliable performance, but there is still room for further improvement in detecting both true and false statements.

Testing has shown that the fake detection tool works correctly, meets the specified requirements, and is ready to be used in real-world conditions.

7. Conclusion

In this article, we develop and present a tool for automating fact-checking based on the use of large language model. The proposed system includes basic modules such as a user interface, a query processing module, a database of verified facts, a module for working with web resources, and a module for analyzing results.

The analysis of existing methods and tools for fact-checking has revealed the main problems and limitations of traditional approaches, which are often not effective enough in the modern information environment. The integration of LLM proved to be a promising solution and improved the quality and speed of information verification by automating the analysis of large amounts of text and searching for reliable sources.

Testing of the system on real data has confirmed its efficiency and flexibility in handling different

types of queries. The tool demonstrates a good ability to quickly adapt to different types of queries and provide accurate answers based on the analysis of a large number of sources. This demonstrates their versatility and ease of use for different categories of users.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

[1] D.-M. Ordway, Fake news and the spread of misinformation: a research roundup, 2017. URL: https://journalistsresource.org/politics-and-government/fake-news-conspiracy-theories-journalism-research/.

[2] University of Tokyo, New analysis shows anti-vaccination conspiracy theories gain political weight due to social media, 2024. URL: https://phys.org/news/2024-02-analysis-anti-vaccination-conspiracy-theories.html.

[3] The real impact of fake news: The rise of political misinformation–and how we can combat its influence, 2024. URL: https://sps.columbia.edu/news/real-impact-fake-news-rise-political-misinformation-and-how-we-can-combat-its-influence.

[4] M. Niessner, Does Fake News Sway Financial Markets?, 2018. URL: https://insights.som.yale.edu/insights/does-fake-news-sway-financial-markets.

[5] N. Lototska, Ukraina naichastishe ye zhertvoiu dezinformatsii ta feikiv, – zvit yevropeiskoi sluzhby zovnishnikh sprav, 2024. URL: https://lb.ua/society/2024/01/25/595364_ukraina_naychastishe_ie_zhertvoyu.html.

[6] Directorate-General for Neighbourhood and Enlargement Negotiations, Disinformation about the Current Russia-Ukraine Conflict – Seven Myths Debunked, 2022. URL: https://neighbourhood-enlargement.ec.europa.eu/news/disinformation-about-current-russia-ukraine-conflict-seven-myths-debunked-2022-01-24_en.

[7] Y. Baryshev, L. Kupershtein, V. Maidanovych, O. Voitovych, S. Prokopenko, Information System for the Fact-checker Support, in: A. Anisimov, V. Snytyuk, A. Chris, A. Pester, F. Mallet, H. Tanaka, I. Krak, K. Henke, O. Chertov, O. Marchenko, S. Bozóki, V. Tsyganok, V. Vovk (Eds.), Selected Papers of the X International Scientific Conference "Information Technology and Implementation" (IT&I-2023). Workshop Proceedings, Kyiv, Ukraine, November 20-21, 2023, volume 3646 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 127–138. URL: https://ceur-ws.org/Vol-3646/Paper_13.pdf.

[8] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. doi:10.1162/tacl_a_00454.

[9] L. Ardaiev, LLM: shcho tse take i yaki vidkryvaie mozhlyvosti, 2023. URL: https://aw.club/global/uk/blog/what-are-llms-and-what-opportunities-do-they-offer.

[10] N. Van Otten, Fact-Checking With Large Language Models (LLMs): Is It A Powerful NLP Verification Tool?, 2024. URL: https://spotintelligence.com/2024/02/26/fact-checking-verification-nlp/.

[11] Y. Sun, J. He, L. Cui, S. Lei, C.-T. Lu, Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges, 2024. URL: https://arxiv.org/abs/2403.18249. arXiv:2403.18249.

[12] J. Su, C. Cardie, P. Nakov, Adapting Fake News Detection to the Era of Large Language Models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1473–1490. doi:10.18653/v1/2024.findings-naacl.95.

[13] S. Abdali, R. Anarfi, C. Barberan, J. He, Decoding the AI Pen: Techniques and Challenges in Detecting AI-Generated Text, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 6428–6436. doi:10.1145/3637528.3671463.

[14] M. Bethany, B. Wherry, E. Bethany, N. Vishwamitra, A. Rios, P. Najafirad, Deciphering Textual Authenticity: A Generalized Strategy through the Lens of Large Language Semantics for Detecting

Human vs. Machine-Generated Text, in: 33rd USENIX Security Symposium (USENIX Security 24), USENIX Association, Philadelphia, PA, 2024, pp. 5805–5822. URL: https://www.usenix.org/conference/usenixsecurity24/presentation/bethany.

[15] S. Kazi, A. Elmahdy, Top Large Language Models (LLMs): GPT-4, LLaMA 2, Mistral 7B, ChatGPT, and More, 2023. URL: https://vectara.com/blog/top-large-language-models-llms-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other/.