Integration of Technologies in Cybersecurity: Information Retrieval and Artificial Intelligence

Oleksandr Puchkov [1], Dmytro Lande [1,2] and Oleksandr Rybak [1]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[2] Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

Abstract

Modern challenges in cybersecurity require new approaches to information retrieval and data analysis. The growth of data volumes and the speed of their updates make traditional information processing methods insufficiently effective. This paper proposes the integration of large language models (LLMs) into information retrieval systems to enhance analytical capabilities and automate data processing tasks. The main goal of the research is to translate the analytical component of the information retrieval system to LLMs, significantly improving the accuracy, completeness, and relevance of information searches.

The system Cyber Aggregator, used for monitoring and analyzing social media content in the context of cybersecurity, demonstrates the effectiveness of the proposed approach. The integration of LLMs into Cyber Aggregator allows for the automation of semantic indexing processes, enhances the formulation and modification of user queries, and provides more precise summarization and analysis of search results. This includes creating analytical digests, identifying key events, constructing semantic maps, and conducting semantic analysis.

The proposed methodology is based on leveraging the powerful capabilities of LLMs, such as understanding complex relationships between concepts, analyzing context, and automatically forming conclusions. The application of this technology in cybersecurity contributes to more effective threat monitoring, improved situational awareness, and enhanced real-time threat response capabilities. The paper also presents a UML diagram illustrating the key components of the system, along with a mathematical formalization of the main processes related to the integration of LLMs into information retrieval systems.

The research findings indicate that the use of LLMs combined with information retrieval technologies opens new opportunities for automating data analysis and ensuring cybersecurity. This makes the proposed approach an important tool for cybersecurity professionals engaged in open-source intelligence (OSINT) and other analytical tasks in today's information environment.

Keywords

Cybersecurity, Information Retrieval, Large Language Models, Data Analysis Automation, Social Media Monitoring, Semantic Analysis, Cyber Aggregator

1. Introduction

The rapid evolution of threats in cyberspace requires the use of advanced tools and methodologies for timely detection and neutralization of threats. Open Source Intelligence (OSINT) is a key component in cybersecurity, utilizing publicly available information to identify and minimize risks [1]. The emergence of generative artificial intelligence models, particularly large language models (LLMs), opens up new opportunities for automating the collection, processing, and analysis of data across various fields [2, 3]. In particular, paper [4] provides a comprehensive overview of the application of LLMs in computational linguistics, while paper [5] outlines the fundamentals of semantic networking, methodologies for forming semantic networks, and domain models through engagement with LLMs.

The authors of this work have already developed the Cyber Aggregator system [6], which is used for monitoring and analyzing content on social media, particularly in the context of cybersecurity.

ITS-2023: Information Technologies and Security, November 30, 2023, Kyiv, Ukraine
iszzi@iszzi.kpi.ua (O. Puchkov); dwlande@gmail.com (D. Lande); rybak.oleksandr01@gmail.com (O. Rybak)
0000-0002-8585-1044 (O. Puchkov); 0000-0003-3945-1178 (D. Lande)

This system has proven effective in collecting and processing large volumes of data from open sources, enabling the prompt detection and analysis of relevant cyber threats.

Based on Cyber Aggregator, the authors propose the implementation of new capabilities of large language models (LLMs) for automating the analysis of textual data. The application of LLMs significantly enhances the accuracy and completeness of searches, improves the formulation and modification of user queries, and provides deeper and more precise analysis of results [7, 8]. This not only reduces the volume of routine work but also increases the overall efficiency of the system, allowing specialists to focus on more complex analytical tasks.

2. Problem Statement

Despite advancements in information retrieval technologies, traditional methods often face issues with the completeness and accuracy of the obtained information. This problem is particularly acute in the field of cybersecurity, where timely access to relevant data is critically important. The challenge lies in the need to develop more sophisticated systems that can intelligently process user queries, identify the most significant information, and present it in a concise and understandable format.

As the volume of information and the complexity of cyber threats increase, traditional methods of information retrieval and analysis become less effective. In a context where cybercriminals employ increasingly sophisticated attack methods, the need for rapid and accurate threat detection is critical. This requires automation of the processing of large data volumes and enhancement of analytical quality.

Recently, large language models (LLMs) have demonstrated significant potential in improving the quality of text information processing. The availability of open-source software and models like LLama-2 [9] opens new opportunities for their integration into closed corporate systems. This is particularly relevant for systems dealing with cybersecurity issues, where ensuring reliable and timely monitoring of the information space is a priority.

Integrating such technologies into existing systems, such as Cyber Aggregator, can significantly enhance the effectiveness of cyber threat detection, optimize information retrieval and analysis processes, and automate the generation of analytical summaries and the construction of semantic networks. This provides a new level of protection for information systems and enables more effective responses to modern cyber threats.

3. Goal

The main goal of this research is to develop and formalize a methodology that integrates LLM into a social media monitoring system focused on cybersecurity to enhance the accuracy and relevance of information retrieval. This methodology includes semantic indexing, query modification, and result summarization performed using LLM. The research also aims to provide a clear mathematical formalization of the processes involved in the proposed system.

To achieve this goal, it is necessary to address the following tasks:

1. Development of a methodology for semantic indexing of textual data using LLM. This task involves creating an algorithm that allows for the preprocessing of textual data by identifying key concepts, their relationships, and forming an index for effective database searching.
2. Integration of LLM into the process of modifying user queries to enhance their accuracy and completeness. The goal is to develop approaches for dynamic modification of user queries based on semantic analysis of the text, which will provide more relevant information retrieval results.
3. Development of algorithms for summarizing search results using LLM. This task includes creating a methodology for automatically generating digests, summaries, and other analytical products based on relevant documents obtained from the search.

4. Formalization of processes involved in the proposed methodology. The task is to develop mathematical models and formalisms that accurately describe the stages of semantic indexing, query modification, and result summarization.
5. Integration and testing of the proposed methodology in real conditions within the Cyber Aggregator system. This task involves implementing the developed approaches into the existing social media monitoring system, conducting test studies, and analyzing the results.

4. System Architecture

The architecture of a social media monitoring system for cybersecurity, integrated with large language models (LLMs), consists of several key components, each performing specific functions and interacting with other parts of the system to achieve a common goal. The main components of the architecture include:

1. Data Collection Module: Responsible for aggregating data from various sources, such as social media, forums, blogs, and other public platforms. This module ensures regular and efficient collection of textual data, including real-time data, with the ability to pre-filter and clean the data.
2. Database and Data Storage: Utilizes specialized data storage systems, such as Elasticsearch, to store the collected information and ensure quick access to it [10]. The database is structured to support efficient semantic indexing and searching, as well as scalability for handling large volumes of information.
3. Semantic Indexing Module: Performs functions of text analysis and building semantic indexes based on key concepts and their relationships [11]. Integration with LLM allows for the creation of more complex and accurate indexes that consider the context and meanings of words in different domains of knowledge.
4. Search Optimization Module: Uses LLM to modify user queries to improve search results. This module automatically analyzes input queries, supplementing or refining them to ensure maximum relevance and accuracy of the results.
5. Results Processing Module: Responsible for summarizing and analytically processing search results. The application of LLM allows for the automatic creation of digests, analytical summaries, detecting events, and constructing semantic maps to visualize relationships between data.
6. User Interface: Provides user interaction with the system. The interface includes dashboards for customizing queries, viewing search results, and obtaining analytical products in a user-friendly format. Sometimes, integration with LLM may also allow interaction through chatbots or other interactive interfaces.
7. Security and Access Management Module: Ensures data protection and access management to the system. This component is particularly important in the context of integration with corporate systems, where strict cybersecurity requirements must be followed.

5. Methodology

The proposed methodology consists of three main stages:

1. The proposed methodology consists of three main stages:

- Data collection: Gathering data from various open sources using the CyberAggregator system.
- Preprocessing: Cleaning and normalizing data to prepare it for indexing.

- Semantic indexing: Using LLMs for semantic indexing, identifying key concepts and relationships in the data. The indexed data is stored in an Elasticsearch database.

2. Query Processing:

- Query analysis: Analyzing and modifying user queries to improve completeness and accuracy. LLM offers synonyms, related terms, and alternative query structures.
- Information retrieval: Searching for relevant documents in the Elasticsearch database based on the modified query.

3. Summary and Analysis of Results**:**

- Summary: Automatic generation of digests, summaries, semantic maps, and other analytical materials using LLM.
- Event detection and semantic map construction: Identifying significant events and creating semantic maps that visualize the connections between key concepts.

6. Mathematical formalization

6.1. Semantic indexing

Let there be a set of documents $D = \{d_1, d_2, \ldots, d_n\}$ and "a set of terms" $T = \{t_1, t_2, \ldots, t_m\}$ used for indexing. The indexing process assigns a weight w_{ij} to each term t_j in the document d_i, which can be formalized as follows:

$$I(d_i) = \{(t_j, w_{ij}) | t_j \in T, w_{ij} \geq 0\}.$$

The weight w_{ij} is determined using LLMs that evaluate the relevance of each term in the context of the document.

6.2. Modification of queries

For the user query $q = \{t_1, t_2, \ldots, t_l\}$, the LLM modifies the query by expanding it with additional relevant terms $t_k{}'$, forming the extended query q':

$$q' = q \cup \{t_1', t_2', \ldots, t_p'\}.$$

The relevance of the document d_i to the query q' is assessed using a similarity function:

$$sim(q', d_i) = \sum_{t_j \in q'} w_{ij}.$$

A document is considered relevant if its similarity score exceeds a defined threshold ε:

$$sim(q', d_i) > \varepsilon.$$

6.3. Summary of Results

The summarization process aggregates information from relevant documents $R = \{r_1, r_2, \ldots, r_k\}$to create a set of summaries U:

$$U = \{u_1, u_2, \ldots, u_k\}.$$

Each summary is generated using LLM (denoted as a function and the corresponding prompt), which highlights the most significant points from the relevant document:

$$u_i = LLM(r_i, prompt).$$

6.4. Detection of connections

Construction of the term relationship matrix:

- We will create a term matrix A of size $m \times m$, where n is the number of terms in the document.
- The element $a_{ij}$of this matrix defines the relationship between terms $t_i$and t_j.

1. Calculation of values in the matrix:

- The significance a_{ij}can be defined as the frequency of co-occurrence of terms t_i and t_j in a document d. This can be implemented by counting how many times the terms appear in the same context, or through metric values such as mutual information.

6.5. Formation of the network

1. Creating a graph:

- Let $G = (V, E)$ be a graph, where V is the set of vertices (terms), and E is the set of edges (connections between terms).

2. Definition of nodes and edges:

- Vertices V correspond to terms t_i from the set $T(d)$.
- Edges E connect pairs of terms t_i and t_j if a_{ij} exceeds a certain threshold θ.

$$E = \{(t_i, t_j) | a_{ij} > \theta\}.$$

where θ is the significance threshold that determines which connections between terms are substantial.

7. Implementation

As a result of integrating the CyberAggregator system with the large language model Llama, significant improvements have been achieved in the system's analytical capabilities in the areas of social media monitoring and information retrieval. In this section, we will explore how Llama's new features enhance various operational modes of CyberAggregator, including information search, dynamic analysis, digest generation, and network construction.

7.1. Information Summaries (Digests)

The combination of search technology with the capabilities of Llama enables the automatic analysis of news reports and the creation of summaries. The Llama model allows the Cyber Aggregator system to generate detailed information digests that include:

1. The system automatically generates digests by processing large volumes of news, identifying key events and facts, and creating a concise overview of the main events based on them.

2. The linguistic capabilities of Llama help to better understand the context and meaning of events, improving the accuracy and usefulness of the digests for users.

7.2. Networks of Hacker groups networks

The Cyber Aggregator system, enhanced with Llama capabilities, effectively visualizes the connections between hacker groups:

1. Detection of connections between groups. Llama assists in identifying and analyzing the connections between different hacker groups, their activities, involvement in cyberattacks, and their relationships with law enforcement agencies of specific states.
2. Information visualization. Integration with Llama allows for the automatic creation of visualizations that depict the connections between groups, facilitating analysis and the detection of patterns in their activities.

7.3. Term networks

The functional capabilities of Llama also enable the analysis and construction of term networks:

1. Term Analysis The model helps automatically identify key terms and their relationships, enabling a better understanding of the context of the information.
2. Building Semantic Networks Using Llama, semantic networks can be created to visualize the connections between terms and concepts, facilitating the understanding of complex concepts and their interrelations.

7.4. Personal Networks

AI systems leverage their linguistic capabilities to create and analyze networks of individuals involved in cyber warfare. This process begins with analyzing individuals' activities, where an LLM model helps detect connections based on their social media interactions and mentions in various sources. By examining this data, the model identifies both direct and indirect relationships among individuals. This approach enables the formation of more accurate and comprehensive networks, providing deeper insights into the dynamics of cyber warfare actors.

To identify the actors involved in the world's first cyberwar, a methodology has been proposed for analyzing selected documents available in electronic sources on the Internet, using a generative artificial intelligence system.

At the first step of the methodology, a query is formed for the search aggregator, such as CyberAggregator, using keywords that must be included in the document for further analysis. This query should include the keywords that are essential for the document's presence for further examination. After finding a sufficient number of text messages, these documents are filtered using generated LLM code, for example, in Python, to search for pairs of concepts formatted as “First Last Name”.

In the next step, the filtering of the provided phrases is carried out. The information is converted into a PDF file, and a prompt is formulated for the LLM with the following wording:

→ Extract names and surnames from the given file, ignoring proper names and organization names.

In our case, approximately 700 names were extracted from over 30,000 phrases. To optimize the construction of the network, a software code was developed in Python, which counts the number of

occurrences and removes all appearances except for the first one, as well as eliminates words that are mentioned less than a specified number of times (in our case, 3), as they lack statistical significance and only clutter the network with unnecessary information. Connections between actors are created using ChatGPT:

→ Find connections between characters linked by their activities to build a cohesive network, and use all names in the connections in the format "character1; character2".

In the third step, after establishing connections between participants in the specified format, the obtained information is recorded in a CSV file.

In the fourth and final step, using a special software application developed based on the GraphViz library [12], a graphical representation of the cyberwar actors and their connections is created (Fig. 1).

Let's provide the mathematical formalization of the method for detecting cybersecurity subjects.

7.4.1. Initial assumptions

Document set $D = \{d_1, d_2, \ldots, d_N\}$ — a collection of documents obtained through OSINT systems based on thematic queries.

Hacker group set H— a set of names of hacker groups that need to be identified from the document texts.

Contextual connections C— a set of connections between hacker groups extracted from the document texts.

7.4.2. Step 1: Formation of the publication information array

For each set of thematic queries Q (for example, queries based on cyberattacks in Ukraine or Israel), we obtain a set of documents Dthat correspond to these queries.

$$D = \bigcup_{q \in Q} OSINT(q)$$

where $OSINT(q)$ is a function that returns a set of documents D for the thematic query q.

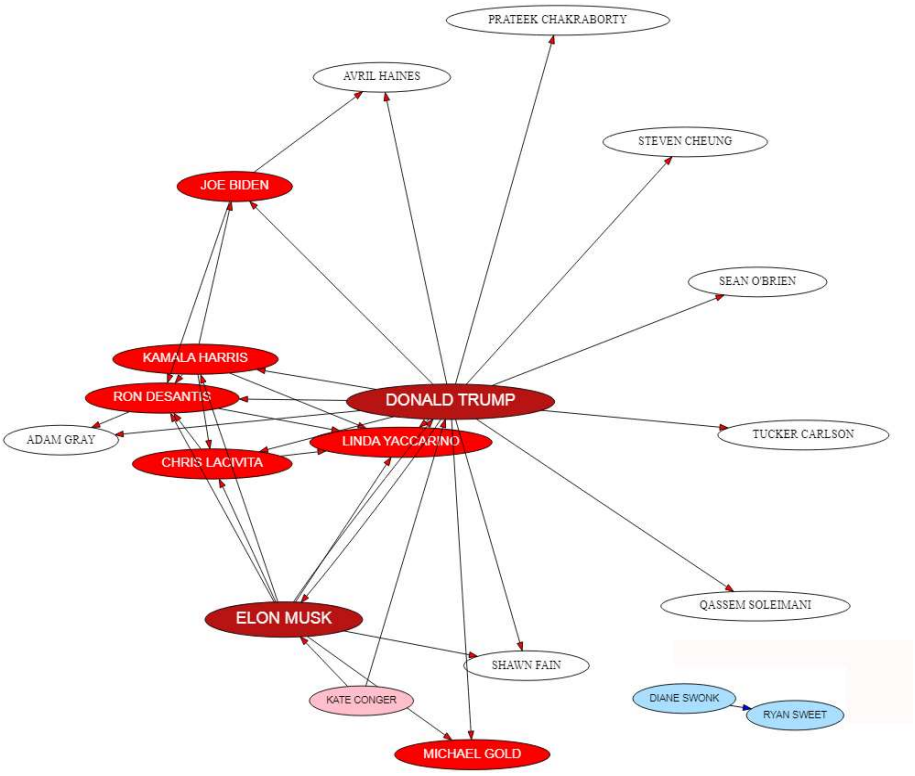

Figure 1 – Fragment of the cyberwar actors network

7.4.3. Step 2: Extraction of hacker group names

For each document $d \in D$, we create the corresponding prompt for the ChatGPT system to extract the names of hacker groups:

$$H(d) = ChatGPT(prompt, d)$$

where $H(d)$ is the set of hacker groups extracted from document d, and prompt is the substantive query to the ChatGPT system.

7.4.4. Step 3: Building a network of connections

Based on the extracted names of hacker groups, we form a set of contextual connections for each document:

$$C(d) = \{(h_i, h_j) | h_i, h_j \in H(d)\}$$

where $C(d)$ is the set of paired connections between hacker groups from document d. The overall set of connections for all documents is defined as:

$$C = \bigcup_{d \in D} C(d)$$

7.4.5. Step 4: Visualization and analysis of the network

The network of connections between hacker groups H, constructed based on a set of groups $H$and a set of connections C, can be represented as a graph $G = (H, C)$, where:

- H— the set of vertices (hacker groups);

- C — the set of edges (contextual connections between the groups).

7.4.6. Computational Complexity

1. Formation of an Information Array of Publications:

   The complexity depends on the number of queries Q and the number of documents N. The complexity of forming a set of documents can be estimated as $O(|Q| \times N)$.

2. Extraction of Hacker Group Names:

   For each document d, a request is made to the ChatGPT system. Let t_{GPT} be the average processing time for one request to the system. Then, the total complexity of this stage is: $O(N \times t_{GPT})$.

3. Construction of a Network of Connections:

   For each document d, the connections between the groups are extracted. If hacker groups $|H(d)|$ are found in the document d, the number of connections between them can be estimated as $O(|H(d)|^2)$. The overall complexity of constructing the network will be: $O(\sum_{d \in D} |H(d)|^2)$.

4. Visualization and Analysis of the Network:

   The complexity of visualization depends on the number of vertices $|H|$and edges $|C|$ in the graph $G = (H, C)$. In the worst case, the complexity of visualization and analysis can be estimated as $O(|H| + |C|)$.

Taking into account all stages, the overall complexity of the algorithm is:

$$O(|Q| \times N + N \times t_{GPT} + \sum_{d \in D} |H(d)|^2 + |H| + |C|)$$

This method allows for the effective extraction and analysis of relationships between hacker groups based on data from textual sources, using generative artificial intelligence tools.

8. Usage

Thus, the proposed approaches enabled:

1. The system successfully identified and summarized key events occurring in the field of cybersecurity.
2. It automatically created analytical digests from a large volume of documents.
3. Semantic maps were constructed to visualize the relationships between key concepts in cybersecurity.

The integration of Llama into the CyberAggregator system significantly improved the quality and accuracy of information retrieval and analytical processes. The system is now capable of automatically generating more detailed and useful informational digests, creating accurate networks of individuals and groups, and conducting deeper semantic analysis of terms. These enhancements contribute to increased efficiency in detecting important events and patterns within large volumes of information, which is critically important for ensuring cybersecurity.

9. Discussion

The proposed methodology has shown significant potential in enhancing information retrieval systems in the context of cybersecurity. The integration of LLMs provides a deeper understanding of user queries and enables the retrieval of more relevant and high-quality information. This opens up new opportunities for automating OSINT processes and improving the efficiency of cybersecurity analysts.

10. Conclusion

In this study, a methodology for integrating large language models (LLMs) into a social media monitoring system focused on cybersecurity has been developed and formalized. The main objective was to improve the accuracy and relevance of information retrieval by implementing new capabilities of LLMs into the CyberAggregator system. The results of the research confirm the success of achieving this goal.

The integration of information retrieval technologies and artificial intelligence has great potential in the field of cybersecurity. The proposed system demonstrates how LLMs can be used to enhance the accuracy and completeness of information retrieval, as well as to automatically summarize results. In the future, the development of this system may lead to the creation of more advanced tools for OSINT, enabling better responses to modern threats in cyberspace.

The methodology developed in the study involves several key stages. It begins with semantic indexing, where LLM is used for the automatic analysis and classification of textual data. The Llama model excels in accurately recognizing terms and their relationships, which helps create a high-quality information index that greatly improves search efficiency. Next, query modification takes advantage of Llama's linguistic capabilities to automatically refine user queries, enhancing the accuracy and completeness of the search results by adjusting them according to the context and specifics of the requested information. Finally, result summarization employs Llama to generate concise summaries and digests based on the search results, offering a clear and understandable presentation of the key facts and events.

The research delves into the mathematical formalization of processes within the proposed system. The Llama model enhances semantic indexing by employing algorithms for semantic analysis that create indexes based on vector representations of words and their contexts. This involves constructing a term-document matrix, where both terms and documents are depicted as vectors in a multidimensional space.

For modifying queries, the mathematical formalization incorporates algorithms designed to refine queries through contextual analysis. This is achieved by optimizing a query's utility function, which allows the model to automatically adjust queries for greater accuracy in results.

Additionally, the generalization of outcomes, such as digests and summaries, is accomplished through text generation algorithms that leverage clustering and data summarization techniques. This process aggregates information while taking into account its significance and context.

The tasks defined in the research objective were implemented as follows: The integration of Llama into the CyberAggregator system allowed for the automation of the analysis of large volumes of textual data. The model automatically processes news messages, creates summaries, and generates reports, which enhances the speed and accuracy of analytical processes. The application of Llama has led to a significant improvement in the accuracy of searches and the relevance of results. The model adapts queries according to the context and specifics of the requested information, enabling the retrieval of more precise and useful results. The developed mathematical models provide a clear understanding and implementation of the processes within the system, allowing for the enhancement of its functionality and integration with Llama.

The developed methodology and integration of Llama into CyberAggregator have a significant impact on the practical application of the system in the field of cybersecurity. It enables effective monitoring and analysis of social media, as well as automatic responses to emerging threats and trends in the information space. This enhances the system's ability to predict and detect potential cyberattacks and threats, which is critical for ensuring cybersecurity.

References

[1] D. Meredith. The OSINT Handbook: A practical guide to gathering and analyzing online information., Birmingham, UK: Packt Publishing, 2024. 198 p. ISBN: 1837638276

[2] S. Wolfram. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media, Inc., 2023. ISBN: 9781579550813, 978-157-9550-82-0.

[3] Chat GPT AI Revolution 2023: A Guide to GTP Chat Technology and Its Social Impact. Technology Summary, 2023. 64 p. ISBN 979-837-7089-14-8.

[4] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, M. Vassilakopoulos. Large Language Models versus Natural Language Understanding and Generation. PCI 2023: 27th Pan-Hellenic Conference on Progress in Computing and Informatics, Lamia, Greece, November 2023. DOI: https://doi.org/10.1145/3635059.3635104

[5] Dmytro Lande, Leonard Strashnoy. GPT Semantic Networking: A Dream of the Semantic Web - The Time is Now. Kyiv: Engineering, 2023. - 168 p. ISBN 978-966-2344-94-3

[6] Dmytro Lande, Olexander Puchkov, Ihor Subach. Method of Detecting Cybersecurity Objects Based on OSINT Technology. Selected Papers of the XXII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2022) — Vol-3503. — pp. 115-124.

[7] ChengXiang Zhai. Large Language Models and Future of Information Retrieval: Opportunities and Challenges. SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pp. 481 – 490. DOI: https://doi.org/10.1145/3626772.3657848

[8] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, D. Guha. A Literature Survey on Open Source Large Language Models. ICCMB '24: Proceedings of the 2024 7th International Conference on Computers in Management and Business. Pp. 133 – 143. DOI: https://doi.org/10.1145/3647782.3647803

[9] F. Castanedo. Run Llama-2 Models. O'Reilly Media, Inc., 2023. ISBN: 9781098163198

[10] Pranav Shukla, Sharath Kumar M N. Learning Elastic Stack 7.0. Distributed Search, Analytics, and Visualization Using Elasticsearch, Logstash, Beats, and Kibana, 2nd Edition. Packt Publishing, 2019. ISBN 9781789958539, 1789958539. — 474 p.

[11] J. Gavilanes, Y. Bozhilov, U. Dodeja, G. Valtas, A. Badrajan. Use of LLM for Methods of Information Retrieval. Report of University of Twente, 2024. Available: https://bachelorshowcase-eemcs.apps.utwente.nl/content/TytQHsvY/Design_report.pdf

[12] Tamilla Triantoro. Graph Viz: Exploring, Analyzing, and Visualizing Graphs and Networks with Gephi and ChatGPT (March 30, 2023). ODSC Community.