ISSN 2224-087X. Електроніка та інформаційні технології. 2024. Випуск 25. С. 102–116
Electronics and information technologies. 2024. Issue 25. P. 102–116

УДК 004.6 DOI: https://doi.org/10.30970/eli.25.9

EXPLORATORY DATA ANALYSIS POSSIBILITY IN THE MEANING SPACE USING LARGE LANGUAGE MODELS

L. Lyashkevych, V. Lyashkevych, R. Shuvar

System Design Department,
Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine

liliia.liashkevych@lnu.edu.ua

Today, data analysis plays a key role in studying and using large amounts of information to make informed decisions. Thus data is one of the most valuable resources and analytical tools are becoming increasingly important. Data analysis allows us to identify patterns and relationships between different factors. As for the methods of analysis, they are very diverse and depend on the specific task and the nature of the data, each method has its advantages and limitations, and the choice of a particular one depends on the context.

Recent investigations have proved that LLM showed good results in textual information analysis and might be one of the textual analytical tools. Therefore, the main focus of our research is centred around the ability of LLM to perform data interpretation operations, arithmetic and statistical operations in the meaning space. In the evaluation of the proposed concepts, simple cases were considered. It was enough to better understand the effectiveness of LLM as one of the tools for exploratory data analysis.

The practical results of the research indicate that the concept has some advantages over the closest analogues, as well as identify several scientific problems that can be solved in subsequent studies. Additionally, research tools were developed as a chatbot system in the Telegram environment.

Keywords: exploratory data analysis, prompt engineering, large language models, chatbot systems, “meaning” space.

Introduction. Big data and advanced technologies bring new opportunities and challenges for engineers and data scientists. [1] The world's leading firms have accumulated huge amounts of data and this process continues, because business and science see great benefit in this, especially in data analysis. According to estimates, the information accumulated on the Internet is huge. In 2024, approximately 149 zettabytes of data should exist, and this number is growing exponentially, as is the world's population capable of generating it. [2] Typically, data is collected from multiple sources at different points in time using various technologies. This creates problems of heterogeneity, experimental variation and statistical variation, and requires us to develop more adaptive and robust methods. [1]

Exploratory data analysis (EDA) [3] has been known for a long time, since 1969 and provides a preliminary data analysis to identify the most general features of information. For example, [4] guides how to begin any statistical analysis and two goals to achieve: data description and model formulation. [4] Volumes, formats, means, and dimensions of data analysis space have changed over the past decades. According to [5], the two main goals of multivariate data analysis are to develop effective methods that can accurately predict future observations, and at the same time gain insight into the relationship between characteristics and response for scientific purposes.

The objectives of EDA can be set up differently, and it depends on many factors. In the work [6], the authors described the research goals as profiling (data quality assessment) and discovery (obtaining new ideas). As was shown in [6], some domain experts can provide a comprehensive EDA. However, there are problems with EDA and statistical data analysis measuring. In the work [7], the authors concluded the effectiveness of combining statistical analysis with EDA and visual tools for data analysis.

Observation of the latest approaches for solving EDA problems. Despite the enormous amount of information we produce, only 0.5% is being analysed and used to discover and structure the data. This may seem like a small number, but it is still huge according to the volume of that information. [1-2] Considering this scale of data, data analysis plays a crucially important role in solving business problems. According to [4, 6-7], it helps:

- make effective decisions;
- accumulate the customer experience;
- gain an edge over competitors;
- support effective decision-making through innovative technologies, etc.

Even though EDA requires in-depth analytical skills, experience, and domain knowledge, many proposals and related systems have been developed to automate it over the past decade.

According to [3-4, 8-11], data analysis includes collecting, modelling and studying data using various mathematical equations, statistical models, modelling tools, machine learning models, etc. The data analysis is a fairly standardised and formalised cross-industry process of intelligent data analysis, known as CRISP-DM (Cross-Industry Standard Process for Data Mining) - which is one of the most common among accepted approaches. [11-12] The CRISP-DM cross-industry standard counts the iterative nature of the data analysis process and ensures a successful iterative search for the best solution based on EDA. Data mining examines and analyses large blocks of information to gather meaningful patterns and trends. It is used in credit risk management, fraud detection, spam filtering, etc. It is also a market research tool that helps you discover the sentiments or opinions of a particular group of people. [11-12] Usually, in the data preparation stage of modelling, data mining counts on intelligent data analysis for various subject domains [11].

Sometimes we group elements into many sets based on their common features. These sets are called clusters, hence the name. [13-14] There is no target variable during clustering, the method is often used to find hidden patterns in data and to provide additional context to trends. The elements are similar to one another in the same cluster. Multidimensional Scaling (MDS) [15] is a method which uses similarities or differences between objects for observation. Objects are represented using an "MDS map" that positions similar objects together and separates ones far apart. Common features of objects were represented in the form of one or more dimensions that can be observed using a numerical scale.

In general, many methodologies in the industry have already been proposed for EDA automation. In [10], the proposed concept of developing integrated methodologies for data management in scientific projects has three fundamental methods: project, team, data and information management. The proposed concept assumes that it must constantly develop, improve and adapt to new challenges in the data area.

EDA describes data features [3], so we can assume: that if algorithms could "speak", they would tell us about data features "verbally" and not use various graphs, charts, etc. According to [16:13], the nature of EDA has changed due to the emerging methods and convergence between EDA and other methodologies, such as data analysis and resampling, so the usual conceptual frameworks of EDA may no longer cope with this trend.

Machine learning (ML) application for solving EDA. ML is a subset of artificial intelligence (AI) that focuses on computer algorithm development that improves automatically through experience and the use of data. In other words, ML allows computers to learn from data and make decisions or predictions without explicit programming, namely, it is the creation and implementation of algorithms capable of facilitating decisions and forecasts. [8-9, 17] The progress in the ML area has created powerful opportunities for the automation of the EDA processes. [18] Indeed, in the ML industry, EDA has been widely used, for example: to detect fraud in the banking system [17], in the study of chronic kidney diseases in medicine [8] or for analog-mass spectrometry of the ocean world [9].

Recently, deep learning (DL) is increasingly used for data analysis. The best examples can be given for identifying anomalies, finding similar examples, building recommendation systems, etc. [19-22] DL, unlike traditional machine learning and data mining algorithms, can create extremely high-level data representations from vast amounts of raw data. [21] DL algorithms extract complex high-level abstractions as data representations using a hierarchical learning process, extract complex patterns from huge volumes of data, perform semantic indexing, data tagging, fast information retrieval, and simplification of discrimination tasks. A key advantage of DL is the analysis and learning of huge amounts of unsupervised data making it a valuable tool for big data analytics, where the raw data is mostly unlabelled. [22] Despite the many advantages of neural networks, they also have a significant drawback: they are generally difficult to test. Some neural network processes can behave like a "black box" where input data is fed, complex processes are performed, and only the output is reported.

Large language models and meaning space. A large language model (LLM) is a DL model that understands and generates text in human language. These models are trained on huge amounts of textual data (books, articles, websites and other sources) and describe billions of parameters. Parameters are variables that are in the model, they change during the learning process. Conventionally, a large language model contains more than one billion parameters. LLMs can understand, translate, predict or generate text or other content. Usually, one or more sentences are given as input to such models, based on which the model tries to understand what they want from it, and generates an answer. [23-26]

The way of word representation is a key factor in the work of LLMs. [23] Earlier forms of machine learning used a numerical table representing each word [27]. However, this representation form could not detect relationships between words with similar meanings. This limitation was overcome by using multidimensional vectors, the so-called "embeddings" of words so that words with similar contextual meanings or other relationships are close to each other in the vector space. Since we can perform operations on words in a vector space, we can claim that this space can be considered a space of "meaning". In [27] we have a good explanation of meaning space and its role in text-logical methods.

Using word embedding, transformers can preprocess text into numerical representations through an encoder, and understand the context of words and phrases with similar meanings, and other relationships between words, such as parts of speech. The LLM then uses this knowledge of the language using a decoder to produce output data. [23, 26]

The emergence of LLM has caused special scientific interest in studying their possibilities for application in data area analysis. According to the reviewed publications [23-26], it becomes clear that LLM is a powerful tool in research practice and can be applied to qualitative data analysis avoiding discrepancies in the analysis. Obviously, in the "space of meaning" [27], where LLM works, concepts depicted on graphs or diagrams can be verbally described and vice versa.

Taking into account all the above-described tools for conducting EDA, trends in the industry regarding the construction of analytical systems [2] and the peculiarities of data preparation for EDA [11, 28-29], it can be stated that LLM is capable of performing some data transformations based on relevant prompts. The variety of existing LLMs, which could be applied for data analysis, is quite big today:

- DataLang performs data analysis based on hints and requires no technical knowledge. It presents the result in human-readable form using GPT-3 technologies. The tool can handle a wide range of queries and provide information on metrics such as the number of users over some time or total values in a specific field;
- Owlbot provides multi-source integration of information: documents, spreadsheets, and databases;
- ChatCSV works with CSV files and with some chart visualisation tools for responses;
- Hal9 enable communication with the company's cloud data, etc.

Tabular data investigation by EDA is one of the important areas in the industry because it is common for the industry to use structured or tabular data that resides in databases or data warehouses. EDA requires primarily execution of data transformation operations related to data processing pipelines and being executed by data engineers. [28, 30] Thus, as usual, the data scientists should develop the specialised tools for scientific investigation by themselves. These applications might connect to databases and provide the required data for EDA.

Taking into account the relevance of the scientific study of the possibilities of LLM for performing EDA on tabular data, as well as their practical application in the construction of data analysis systems, the tasks for scientific investigation can be formulated as follows:

- to investigate the possibilities of supporting complex computing operations on data in EDA tasks;
- to explore the possibilities of data transformation according to the LLM legend;
- to explore the possibilities of data analysis using LLM;
- to develop the concept of managing data analysis using prompts;
- to create a system architecture for performing EDA on tabular data;
- to implement a chatbot system for investigating the proposed solutions.

Overall, we investigate EDA automation opportunities using LLM capabilities.

The ChatGPT capabilities exploration to resolve EDA tasks in meaning space. As it is known, we cannot directly communicate with LLM. Prompting is one way of communication with LLM. Therefore, we generated tabular data with the next prompt body: “Prompt 1: Generate data for a table with the following columns: Company name (this attribute must not be repeated); Field of activity; Country of origin; Company size (number of employees); Annual income (in US dollars); Net profit for the last fiscal year; Number of branches/departments; The average salary level of employees; Customer satisfaction rating; Advertising and PR expenses”.

The results of the generated data are shown in Figure 1. The generated data contains random values which correspond to reality in LLM understanding. To have more accurate data, you can specify additional conditions in tooltips. In this case, the network will count instructions in prompt and generate data in the given range. The only drawback of this approach is that the network cannot generate large amounts of tabular data, we should iterate this process.

Company Name	Field of Activity	Country of Origin	Company Size	Annual Income (USD)	Net Profit (USD)	Number of Branches
TechSolutions Inc.	Information Technology	USA	500	$25,000,000	$5,000,000	10
Global Pharma	Pharmaceuticals	Germany	1200	$120,000,000	$20,000,000	25
BrightTech Solutions	Software Development	UK	300	$30,000,000	$8,000,000	6
GreenEnergy Solutions	Renewable Energy	Canada	150	$15,000,000	$3,000,000	4
MedLife	Healthcare	USA	800	$80,000,000	$15,000,000	15

Fig. 1. The results of tabular data generation by ChatGPT 3.5 with Prompt 1

Let's clarify a range of values for generated data. The next prompt body is: "Prompt 2: Generate data for a table with the following columns: Company name (this attribute must not be repeated); Field of activity; Country of origin; Company size (number of employees); Annual income (in US dollars); Net profit for the last fiscal year; Number of branches/departments; The average salary level of employees; Customer satisfaction rating; Advertising and PR expenses. Use these ranges for the data attributes: Company size (number of employees) - from 10 to 100; Annual income (in US dollars) - from $50000 to $1000000". The results are shown in Figure 2.

Company Name	Field of Activity	Country of Origin	Company Size	Annual Income (USD)	Net Profit (USD)	Number of Branches	Average Salary (USD)
InnovateTech	Information Technology	USA	55	$650,000	$120,000	3	$70,000
HealthCare Solutions	Healthcare	UK	35	$200,000	$40,000	2	$55,000
EcoSolutions	Environmental	Canada	75	$750,000	$150,000	5	$80,000
GlobalTrade	Import/Export	Germany	45	$400,000	$80,000	4	$65,000
EduLearn	Education	Australia	20	$120,000	$20,000	1	$50,000
FoodDelight	Food & Beverage	France	65	$600,000	$100,000	3	$75,000
TechInnovate	Software Development	India	30	$250,000	$50,000	2	$60,000

Fig. 2. The results of tabular data generation by Prompt 2

As shown in Figure 2, the LLM understand us clearly, so, we assume that LLM can execute the number comparison operations in meaning space correctly. Continuing with data generation by prompting we can get the data as demanded.

For working with the statistical analysis we generated different prompts, for example: "Prompt 3: Create statistics for annual revenues with a histogram: from the company that earns the most to the one that earns the least". The results of the Prompt 3 execution are shown in

Figure 3. Of course, ChartGPT didn't show the histogram but prepared a compiled code in Python for execution. It is an acceptable output for us because we will use it later building the analytical chatbot system for EDA execution above tabular data.

ChatGPT

To create a histogram for annual revenues from the company that earns the most to the one that earns the least, we need to gather the annual revenue data and then plot it in a histogram. Here's how we can do it:

1. **Gather Annual Revenue Data:** We'll use the annual revenue data from the generated dataset.
2. **Sort Companies by Annual Revenue:** Sort the companies in descending order based on their annual revenue.
3. **Plot the Histogram:** Plot the histogram with the annual revenues.

Let's generate the histogram:

python Copy code

```python
import matplotlib.pyplot as plt

# Annual revenue data
annual_revenues = [
    900000, 750000, 650000, 600000, 550000,
    400000, 300000, 250000, 200000, 120000
]

# Company names
company_names = [
    "RealEstate Group", "EcoSolutions", "InnovateTech", "FoodDelight", "FashionTrend'
    "GlobalTrade", "MediaHub", "TechInnovate", "HealthCare Solutions", "EduLearn"
]
```

Fig. 3. The results of Prompt 3 execution

As seen in Figure 3, the code has inputs from LLM as a response to our prompt. Thus, we can conclude that the network prefers to execute computations with code, rather than directly performing them in the space of "meaning".

Let's ask ChatGPT to explain the output without code. The next prompt body is: "Prompt 4: Create statistics for annual revenues with a histogram: from the company that earns the most to the one that earns the least and describe the output values without coding in Python".

ChatGPT

To create statistics for annual revenues with a histogram from the company that earns the most to the one that earns the least, we can visualize the distribution of annual revenues among the companies in our dataset. Here's a description of the output values without coding in Python:

1. **Company Names:** The companies are listed along the x-axis of the histogram from left to right. Each bar represents a company, with the company that earns the most revenue on the left and the one that earns the least on the right.
2. **Annual Revenue (USD):** The y-axis represents the annual revenue in US dollars. The height of each bar corresponds to the annual revenue of the respective company.
3. **Distribution:** The histogram shows how the annual revenues are distributed across the companies. A taller bar indicates a higher annual revenue, while a shorter bar indicates a lower annual revenue.
4. **Range:** The range of annual revenues is displayed on the y-axis, showing the minimum and maximum revenue values. This range helps understand the spread of revenue values among the companies.
5. **Frequency:** The frequency of annual revenues within certain ranges is depicted by the height of the bars. Higher bars indicate a higher concentration of companies with revenues falling within that range.
6. **Trends:** By observing the histogram, we can identify any trends or patterns in the distribution of annual revenues. For example, we might notice clusters of companies with similar revenue levels or outliers with exceptionally high or low revenues.
7. **Comparison:** The histogram allows for easy comparison of annual revenues between companies. We can quickly see which companies earn the most and least revenue, as well as how they compare to each other in terms of revenue magnitude.

Fig. 4. The results of Prompt 3 execution

As was mentioned earlier, the data interpretation techniques, as usual, are being executed in the stages before EDA, when we store data in databases, data warehouses or else. Namely, during the data interpretation stage, the ETL operations are executed and raw data are represented in the desired data structure. These operations are one of the main problems in the industry, especially, for working with Web Analytics because of possible losses in data.

Data governance and quality assessment are still big problems for working with Big Data or streamed data in Enterprise because of the wide range of possible solutions here which should be upgraded in the future due to different reasons.

Continuing with EDA, let's define ChatGPT capabilities for working with data interpretation operations. The data interpretation The appropriate prompt is: "Prompt 5: Provide the normalised values in a range from 0 to 1 for the data attribute: "Annual Income (USD)". The results of execution are shown in Figure 5.

ChatGPT

To normalize the values of the "Annual Income (USD)" attribute to a range from 0 to 1, we can use min-max scaling. The formula for min-max scaling is:

$$\text{Scaled value} = \frac{\text{Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}}$$

Here's how we can normalize the "Annual Income (USD)" values to a range from 0 to 1:

1. Find the minimum and maximum values of the "Annual Income (USD)" attribute.
2. Use the formula to calculate the normalized values for each data point.

Let's say we have the following annual income values:

$[\$50,000, \$120,000, \$200,000, \$300,000, \$400,000, \$550,000, \$600,000, \$650,000, \$750,($

Then, the normalized values would be:

$$\text{Min Value} = \$50,000$$
$$\text{Max Value} = \$900,000$$

$$\text{Normalized Value} = \frac{\text{Value} - 50,000}{900,000 - 50,000}$$

$$\text{Normalized Values} = [0.0, 0.033, 0.067, 0.1, 0.133, 0.2, 0.233, 0.267, 0.333, 1.0]$$

These normalized values represent the "Annual Income (USD)" attribute scaled to a range from 0 to 1. The highest value (900,000) is scaled to 1, and the lowest value (50,000) is scaled to 0. All other values fall between 0 and 1 based on their relative positions between the minimum and maximum values.

Fig. 5. The results of Prompt 5 execution

As shown in Figure 5, LLM supports the operations about the digits and real numbers, it proved that LLM can operate the digits in the meaning space. As can be seen from Figure 5, it is not required to represent operations in the form of formulas or mathematical descriptions because LLMs perfectly understand all concepts in the meaning space, at least simple transformations. In this case, we can talk about the technology of prompts.

EDA is a sequence of operations on data relying on the experience of conducting EDA in [4, 6-10, 17]. The specified operations depend on the goals and objectives of the EDA. Even when we have different goals, some parts or operations of EDA can be used for them. The EDA can be represented mathematically in the form of Boolean algebra. In this case, let's create the prompt template for EDA processes as a chain of consequent operations, which is shown in Figure 6.

""" Original data: "..."

{Instruction 1} U {Original data} U {Instruction 2} U {Operation 1}
U {Operation 2} U {Operation 3} U {Operation 4} U {Operation 5}
"""

where:

- Instruction 1: “Perform a statistical analysis on the original data:“,
- Instruction 2: “In the order of performing the following operations: ”,
- Operation 1: “Calculate the average value of prices for goods;”,
- Operation 2: “Find the median number of goods in stock;",
- Operation 3: "Let's count the total number of goods for each year of production;",
- Operation 4: "We will determine the most expensive and cheapest goods in the table;",
- Operation 5: "Let's build a histogram of the distribution of prices for goods."

Fig. 6. The example of prompt template form for EDA execution

Changed the chain of instruction and operation we manage the goals and targets for EDA. Enhancing the instructions and operation with extra conditions and restrictions we can easily manage the end-to-end EDA execution without programming.

Analysing the EDA approaches represented in [4, 6-7, 9-11] we have created a list of prompts which cover the most widespread operations in EDA pipelines. Indeed, by configuring the prompts, LLMs provide useful analytical information about the given data. Received results of the investigations show that LLM can execute more complex operations but it will be a challenge for the next investigations.

Resolving EDA tasks with an analytical chatbot system. Received results of the conducted research, have convinced that LLMs perform some simple operations above the data in the meaning space. This means that the existing rule-based or ML-based analytical tools should be upgraded because they work on the algorithmic presentation of the results of data analysis. Thanks to the capabilities of LLM for performing analytical operations, in particular, performing EDA, we can automate the EDA processes by chatbot systems with a prompt formation support system, which supports the conversation flow with the right prompt templates. The question of the algorithms for prompt template detection is out of the scope of this investigation. Let’s consider some chatbot systems for analytics.

In [31], were shown the results of chatbot system investigations. Integrating big data as a knowledge base into chatbots can generate dynamic responses to user queries and improve the analytical capabilities of chatbots with data from a distributed environment. This cutting-edge technology directly opens up the world of big data to chatbots, allowing chatbots to be used as a business intelligence analysis tool. [31]

An example of a professional system for bibliometric analysis of publications was shown in [32]. The authors have developed a chatbot based on ChatGPT, which allows searching for the most relevant publications. This solution is one of the most widespread LLM-based chatbot solutions which includes the components for processing the user queries with a user-friendly interface, RAG system, and access to LLM.

An appropriate chatbot system has been developed for effective EDA on tabular data. With this tool, we can continue to investigate the complex capabilities of LLM in data analysis.

The architecture of the chatbot system is shown in Figure 7.

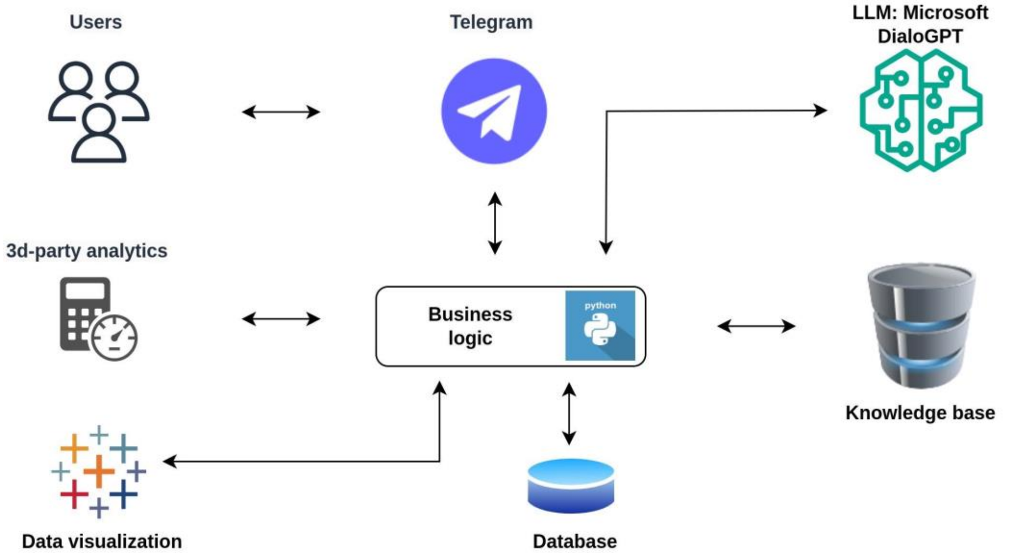

Fig. 7. The architecture of the proposed analytical chatbot system

At the centrepiece of the chatbot is the business logic that controls the operation of the entire chatbot system. Through the Telegram channel, users can communicate with our analytical chatbot system. Sometimes we don't need to ask LLM but only communicate with the chatbot system. In this case, we type an appropriate command directly in the channel. This is very convenient, especially for those who have been using it for a long time and do not need to print large and long queries for LLM.

The business logic connects with the database and the knowledge base. The database contains tabular data. In the knowledge base, we keep prompt templates and other useful knowledge. In the future, the functionality of the chatbot system will be extended due to new requirements for the EDA or other analytics.

Following the instructions for Telegram developers, we have built the chatbot, named "Data Visualizer" (fig. 8).

Fig. 8. Prepared chatbot in the Telegram environment

The chatbot is ready for integration with the back-end part of the chatbot system. To develop the business logic, we used Python which supports several programming paradigms and has a simple syntax and a wide range of libraries, such as python-telegram-bot for interacting with the Telegram API, libraries for ML, such as transformers, etc. This makes Python a good choice for efficient chatbot development.

The Microsoft DialoGPT autoregressive language model was chosen among LLMs because of the free license and easy access through API. This LLM can predict the next future word from a sequence of words given as input. This network was trained on text data using GPT-3. One of its features is the choice of the tone of the chatbot's responses: creative, balanced and accurate. To get started, users need to write a text topic and then define the characteristics of the text according to their needs.

As a database, we used “Cardiovascular_Disease_Dataset.csv”, taken from kaggle.com. This is a research-accessible heart disease dataset collected at a multispecialty hospital in India. It consists of 1000 subjects with 14 features. Among them are ID, age, gender, and levels of many medical indicators of patients. Before running the bot, we created a MySQL database and uploaded the database to it. When the chatbot starts, it loads a database in memory and is ready to use it. If you want to see the data attributes, ask the chatbot (fig. 9). Ask the chatbot about the bar plot for some attributes and the chatbot will show a bar plot for them (fig. 10).

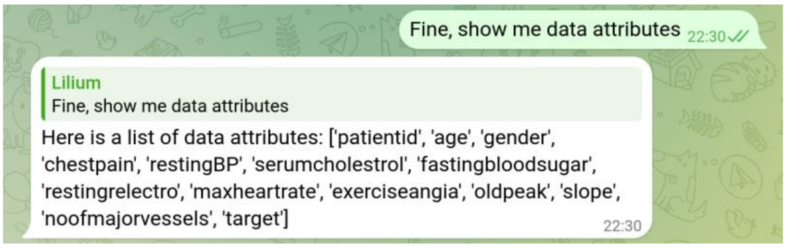

Fig. 9. The chatbot shows the data attributes

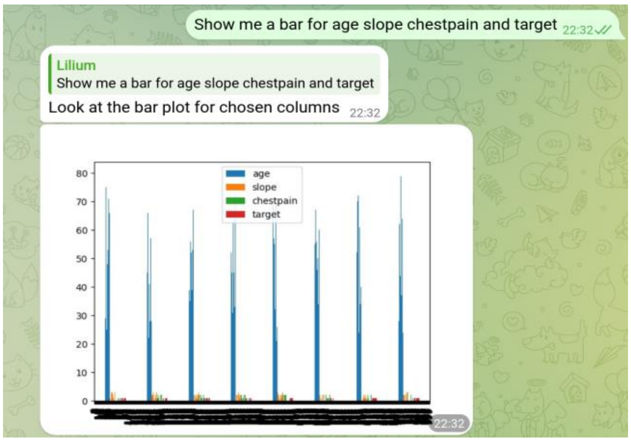

Fig. 10. The chatbot shows the bar plot for chosen data attributes

Continuing working with the chatbot system, we can discover the LLM capabilities for complex operations execution within EDA.

Summary. AI technologies play a crucially important role in EDA implementation. Thanks to AI, various processes are automated and optimised through intelligent chatbots and this is a starting point for further research. Today, it is possible to create even more complex analytical systems capable of performing tasks that overcome humans using new approaches and algorithms as LLM is.

The literature sources analysis indicates that the chosen direction of scientific research is relevant, and the topic of performing data analysis operations in the meaning space has a great scientific and practical interest. Among the received results we want to highlight:

- the concept of EDA execution in meaning space is firstly proposed in the part of EDA pipeline dynamical structuring without so far programmed strong pipelines. The EDA is being executed directly as a response to a user query. This approach reduces time for EDA and excludes redundant information due to user requests.
- the possibility of supporting complex computational operations on data in EDA tasks was investigated, and it was indicated that complex computational operations should be decomposed or supported with extra analytics or code, generated by LLM;
- the possibilities of data transformation according to the LLM tradition were investigated and it was determined that LLMs work well with data formats, easily identify entities, normalise the data and text analytics are also available to them. In the case of complex data conversion operations, they should be divided into simpler ones;
- the possibilities of LLM for data analysis were investigated, and it was determined that the network in the meaning space fully understands the EDA operations, and can also offer Python code for its running. Clarifying the prompts, the network won't provide a code but can explain it and the results of the calculations. Meanwhile, LLM works fine with real numbers with a good prompting approach as well;
- the proposed concepts of managing data analysis using a sequence of operations and instructions in prompts was successfully tested;
- an architecture was designed and a chatbot system was developed to perform EDA on tabular data for approbation of the proposed solution.

The main functionality of the chatbot is its ability to maintain a conversation thanks to the used LLM, reaction to keywords, EDA of the database, provided by the user, communication with Matplotlib libraries and data visualisation support. The chatbot is potentially useful for those who need to understand the data. Therefore, it automates EDA processes, avoiding the programming part. As usual, EDA is being implemented by many data engineers, data analysts and data scientists, because real analytical systems are programmed and cannot dynamically change the order of displaying certain graphs and charts.

Considering all the above, we propose the following tasks for further research:

- decomposition of complex EDA operations;
- capability of LLM in deep data analysis;
- support of EDA with ontology-driven prompt engineering;
- support the chatbot with extra analytic tools and RAG system;
- development of typical EDA templates for its various purposes and tasks;
- improvement of real-time analytical chatbot systems.

REFERENCES

[1] *Jianqing Fan, Fang Han, Han Liu*. Challenges of Big Data analysis: National Science Review, June 2014. - V. 1. - I. 2. - pp. 293–314. - DOI: 10.1093/nsr/nwt032

[2] *Pranay Ahlawat, Justin Borgman, Samuel Eden, Steven Huels, Jess Iandiorio, Amit Kumar, and Philip Zakahi*. A New Architecture to Manage Data Costs and Complexity: BCG,

February 7, 2023. - URL: https://www.bcg.com/publications/2023/new-data-architectures-can-help-manage-data-costs-and-complexity?linkId=200819392

[3] Exploritary data analysis: Wikipedia. - URL: https://en.wikipedia.org/wiki/Exploratory_data_analysis

[4] *Chris Chatfield*. Exploratory data analysis / European Journal of Operational Research, 1986. - V. 23, I. 1, pp. 5-13. - DOI: 10.1016/0377-2217(86)90209-2

[5] *Peter Bickel*. Discussion on the paper “Sure independence screening for ultrahigh dimensional feature space” by Fan and Lv. J. Roy. Statist. Soc. Ser. B, 70(5):883–884, 2008

[6] *Kanit Wongsuphasawat, Yang Liu, Jeffrey Heer*. Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study, 1 Nov 2019. - 10 p. - arXiv: https://arxiv.org/pdf/1911.00568.pdf

[7] *Jessica Hullman, Andrew Gelman*. Challenges in Incorporating Exploratory Data Analysis Into Statistical Workflow: Harvard Data Science Review, 2021. - V.3. - I. 3. - 11 p. - DOI: 10.1162/99608f92.9d108ee6

[8] *Mehta V, Batra N, Poonam, Goyal S, Kaur A, Dudekula KV, Victor GJ*. Machine Learning based Exploratory Data Analysis (EDA) and Diagnosis of Chronic Kidney Disease (CKD): EAI Endorsed Transactions on Pervasive Health and Technology, 2024. - 8 p. - DOI: 10.4108/eetpht.10.5512

[9] *Da Poian V, Theiling B, Clough L, McKinney B, Major J, Chen J and Hörst S*. Exploratory data analysis (EDA) machine learning approaches for ocean world analog-mass spectrometry, 2023. - 17 p. - DOI: 10.3389/fspas.2023.1134141

[10] *Inigo Martinez, Elisabeth Vilesb, Igor G Olaizolaa*. Data Science Methodologies: Current Challenges and Future Approaches, Jan 2022. - 22 p. - arXiv: https://arxiv.org/pdf/2106.07287.pdf

[11] *Christoph Schocka, Jonas Dumlerb, Prof. Dr.-Ing. Frank Doeppera*. Data Acquisition and Preparation – Enabling Data Analytics Projects within Production // 54th CIRP Conference on Manufacturing Systems: Procedia CIRP, 2021. - V. 104, pp. 636-640. - DOI: 10.1016/j.procir.2021.11.107

[12] *F. Martínez-Plumed et al*., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," in IEEE Transactions on Knowledge and Data Engineering, 1 Aug 2021. - V. 33. - N 8. - pp.3048-3061, - DOI: 10.1109/TKDE.2019.2962680

[13] *N. Bratchell*. Cluster analysis: Chemometrics and Intelligent Laboratory Systems, 1989. - V. 6. - I. 2. - pp. 105-125. - DOI: 10.1016/0169-7439(87)80054-0

[14] *Brady Lund, Jinxuan Ma*. A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. Performance Measurement and Metrics, 2021. - N 22. - pp.161-173. - DOI: 10.1108/PMM-05-2021-0026

[15] *Mugavin, Marie*. (2008). Multidimensional Scaling: A Brief Overview. Nursing research. 57. 64-8. - DOI: 10.1097/01.NNR.0000280659.88760.7c

[16] *Chon Ho, Yu*. (2010). Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research, 3(1), 9-22. - URL: https://www.redalyc.org/pdf/2990/299023509014.pdf

[17] *Miguel Ângelo Lellis Moreira, Claudio de Souza Rocha Junior, Diogo Ferreira de Lima Silva and others*. Exploratory analysis and implementation of machine learning techniques for predictive assessment of fraud in banking systems: Procedia Computer Science, 2022. - V. 214. - pp. 117-124. - DOI: 10.1016/j.procs.2022.11.156

[18] *Tova Milo, Amit Somech*. SIGMOD '20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, June 2020. - pp.2617-2622. - DOI: 10.1145/3318464.3383126

[19] *Tabassum, Lubna*. (2020). Fundamentals of artificial intelligence and deep learning techniques. 2020

[20] *Hend A. Selmy, Hoda K. Mohamed, Walaa Medhat*. Big data analytics deep learning techniques and applications: A survey: Information Systems, 2024. - V.120. - DOI: 10.1016/j.is.2023.102318

[21] *Sarker, I.H*. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021). - DOI: 10.1007/s42979-021-00815-1

[22] *Najafabadi, Maryam & Villanustre, Flavio & Khoshgoftaar, Taghi & Seliya, Naeem & Wald, Randall & Muharemagic, Edin*. (2016). Deep Learning Techniques in Big Data Analytics. - DOI: 10.1007/978-3-319-44550-2_5

[23] *Parishad Behnam Ghader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, Siva Reddy*. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, 9 Apr 2024. - arXiv: https://arxiv.org/pdf/2404.05961v1.pdf

[24] *Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, Ingo Weber*. LLMs for Science: Usage for Code Generation and Data Analysis, 7 Dec 2023. - arXiv: https://arxiv.org/pdf/2311.16733.pdf

[25] *Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G.* (2024). An examination of the Use of Large Language Models to Aid Analysis of Textual Data. International Journal of Qualitative Methods, 23. - DOI: 10.1177/16094069241231168

[26] *Jacqueline A Jansen, Artür Manukyan, Nour Al Khoury, Altuna Akalin*. Leveraging large language models for data analysis automation, 2024. - 18 p. - DOI: 10.1101/2023.12.11.571140

[27] *Lokazyuk V.* Software development for texts with diagnostic information processing [Electronic resource] / V. Lokazyuk, V. Lyashkevych, O. Olar // Radio electronic and computer systems. - 2007. - № 6. - P. 123–129. - URL: http://nbuv.gov.ua/UJRN/recs_2007_6_25

[28] *Fernandes, A.A.A., Koehler, M., Konstantinou, N. et al.* Data Preparation: A Technological Perspective and Review. SN COMPUT. SCI. 4, 425 (2023). - DOI: 10.1007/s42979-023-01828-8

[29] Preprint of Atzmueller, M., Schmidt, A., Hollender, M. (2016) Data Preparation for Big Data Analytics: Methods & Experiences. In: Enterprise Big Data Engineering, Analytics, and Management, IGI Global.

[30] *Costello, Tim & Blackshear, Lori*. Prepare Your Data for Tableau: A Practical Guide to the Tableau Data Prep Tool, 2020. - DOI: 10.1007/978-1-4842-5497-4

[31] *Sankar, Reshmi*. (2018). EMPOWERING CHATBOTS WITH BUSINESS INTELLIGENCE BY BIG DATA INTEGRATION. International Journal of Advanced Research in Computer Science. 9. 627-631. - DOI: 10.26483/ijarcs.v9i1.5398

[32] *Hamed Khosravi, Mohammad Reza Shafie, Morteza Hajiabadi, Ahmed Shoyeb Raihan, Imtiaz Ahmed.* Chatbots and ChatGPT: A Bibliometric Analysis and Systematic Review of Publications in Web of Science and Scopus Databases: arXiv, 2023. - 30 p. - arXiv: https://arxiv.org/pdf/2304.05436.pdf

МОЖЛИВОСТІ РОЗВІДКОВОГО АНАЛІЗУ ДАНИХ У ПРОСТОРІ СМИСЛУ ВИКОРИСТОВУЮЧИ ВЕЛИКІ МОВНІ МОДЕЛІ

Л. Ляшкевич, В. Ляшкевич, Р. Шувар

Кафедра системного проектування,
Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, Львів, 79005, Україна

liliia.liashkevych@lnu.edu.ua

Сьогодні аналіз даних відіграє ключову роль у вивченні та використанні великих обсягів інформації для прийняття обґрунтованих рішень. Таким чином, дані є одним із найцінніших ресурсів, а аналітичні інструменти стають дедалі важливішими. Аналіз даних дозволяє виявити закономірності та зв'язки між різними факторами. Що стосується методів аналізу, то вони дуже різноманітні і залежать від конкретного завдання та характеру даних, кожен метод має свої переваги та обмеження, а вибір конкретного залежить від контексту.

Недавні дослідження довели, що LLM показали хороші результати в аналізі текстової інформації та можуть бути використані як один із текстових аналітичних інструментів. Таким чином, основна увага нашого дослідження зосереджена навколо здатності LLM виконувати операції інтерпретації даних, арифметичні та статистичні операції в просторі "смислу". При дослідженні запропонованої концепції розглядалися прості випадки. Цього було достатньо, аби зрозуміти ефективність LLM як одного з інструментів для розвідкового аналізу даних.

Практичні результати та дослідження свідчать про певні переваги концепції перед найближчими аналогами, а також визначають ряд наукових проблем, які можуть бути вирішені в подальших дослідженнях. Крім того, інструменти дослідження були розроблені як чат-бот система у середовищі Telegram.

Ключові слова: дослідницький аналіз даних, оперативна інженерія, великі мовні моделі, системи чат-ботів, простір "смислу".

Стаття надійшла до редакції 05.03.2024.
Прийнята до друку 12.03.2024.