**Bozhko Oleksandr**

postgraduate student

«Kharkiv National University of Radio Electronics», Ukraine

**Petrov Konstantin**

Dr. Sc., Prof., Head of Information Control Systems Department

«Kharkiv National University of Radio Electronics», Ukraine

**Utilizing Large Language Models for Information Recognition in Numismatic Descriptions**

In the context of economic informatics, the forecasting of price fluctuations in the numismatic market, characterized by a high degree of uncertainty, is an urgent task. An important aspect in this process is the analysis of textual descriptions of objects using natural language processing (NLP) technologies. The peculiarity of collecting information on the pricing of numismatic objects is its extraction from user descriptions made in an informalized form and in different languages. This adds complexity to the process of data analysis and extraction.

The use of large language models (LLMs) for identifying and structuring information from textual descriptions is a topic that has recently been extensively discussed in the scientific literature. However, most publications focus on building, fine-tuning [1] or retraining [2] these models.

This study can be seen as part of a wider project on price forecasting of numismatic material, which involves further development and comparison of different LLM approaches and models.

It aims at solving the NLP problem by means of a methodological approach to the use of pre-trained language models (PLM) and, in the context of numismatics, to outline the initial results of its application. Particular attention is paid to assessing the accuracy of the model, analyzing its applicability to this task, and considering its cost-effectiveness compared to other methods and models.

In general, the Information Extraction task is a subset of the natural language processing industry and consists of analyzing and processing textual data for the purpose of identifying and structuring specific [3], relevant data from unstructured textual sources. Specifically, this study focuses on extracting and classifying key attributes of numismatic material, such as issue period, geographical origin, material used, and condition of coins, from textual descriptions.

The focus is on the use of large language models for processing and analyzing textual data in numismatics. The task includes aspects of named entity recognition (NER) recognition and categorization in text, which is critical for transforming large bodies of unstructured text into an organized and systematic form suitable for further analysis within the numismatic discipline. The process includes elements of semantic normalization such as unification of synonyms, resolution of polysemousness, and standardization of formats. The research focuses on the development and validation of a textual data processing method based on the application of advanced language models. A distinctive feature is the use of pre-trained models as a tool for extracting information from textual descriptions of coins. The principles of deep learning and context adaptability make these models particularly suitable for analyzing complex and unstructured data specific to numismatic descriptions. The challenge involves a comprehensive approach to information extraction and classification to accurately identify each element of numismatic material, including the development and validation of algorithms capable of considering both explicit and implicit linguistic characteristics inherent in numismatic descriptions.

To address this challenge, we investigated the feasibility of using pre-trained language models that have the potential for deep understanding of natural language and can handle a wide range of linguistic nuances. An important aspect of the research is to analyze the ability of LLMs to extract key attribute values from text. This research seeks not only to create effective tools for classifying numismatic material, but also to broaden the understanding of the capabilities of modern NLP technologies as applied to a variety of applications.

In the initial stages of this research, GPT-3.5 Turbo and GPT-4 Turbo (gpt-4-1106-preview)[4] developed by OpenAI were chosen as the primary text processing tool. This choice is due to several key factors. First, the advanced natural language generation and understanding ability of these algorithms allows them to efficiently process unstructured textual descriptions of coins. Second, due to their extensive training base and powerful learning algorithms, GPT models can identify complex patterns and extract key information from text, which is critical for identifying attributes of numismatic material. Finally, the flexibility of the GPT API allows for easy integration into an existing data processing system.

Integrating the GPT model into a data processing system goes through several steps. First, lot descriptions are collected and systematized. The data is then passed to a Python script that serves as an interface between the database and the GPT. The script generates and sends queries to the GPT, receiving in return attribute values in JSON format. This data is then compared to a reference in the database to identify or create new records. The GPT model can supplement missing attributes based on the context of the description, which increases the accuracy and efficiency of the identification process.

By applying this approach, the potential of Large Language Model (LLM) learning can be effectively harnessed to extract and classify information from textual data in numismatics, providing a high degree of accuracy and efficiency in query processing.

During the study, the effectiveness of GPT 3.5 Turbo and GPT 4 Turbo models in the tasks of recognizing and classifying numismatic material based on textual descriptions was evaluated. On a sample of about 1000 records, both models showed high accuracy in determining attribute values (table 1) [5].

*Table 1*

**Evaluation results of information retrieval algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1-** **Score** |
| ChatGPT 4,5 | 0,997 | 0,996 | 0,997 |
| ChatGPT 3,5 Turbo | 0,988 | 0,987 | 0,987 |

The study found that pre-trained language models (PLMs) are effective in identifying and classifying numismatic material based on pre-defined attributes. However, their ability to autonomously extract and identify complex sets of attributes from text was found to be limited.

The use of GPT models is characterized by accessibility and ease of use, not requiring significant computational resources or deep programming knowledge to deploy and train your own models. The main disadvantage is the cost of using the GPT 4 Turbo API. Nevertheless, the simplicity of implementation and the lack of need for complex hardware make this approach economically attractive, especially for tasks that do not require frequent processing of large amounts of data.

The results of this study show that the use of GPT Large Language Models (LLMs) for classification and identification of numismatic materials is efficient in terms of accuracy and can be cost-effective, especially in resource-constrained environments. However, the cost of using the API may limit its application in projects with high query volume.

This research intends to analyze and compare various available pre-trained language models with APIs to evaluate their performance in tasks of similar nature. Exploring the diverse models will help to deepen the understanding of the potential of PLM in the context of information extraction from texts. It is noted that despite the superiority of the GPT 4 Turbo model in terms of key features at the time of writing, there are various models that may be more suitable for certain tasks.

To optimize the process and reduce the number of queries to LLM, it is planned to modify the algorithm to include the creation of dictionaries in the process of recognizing text descriptions to fill in attribute values, and to use LLM only for objects not recognized using these dictionaries. This will improve the efficiency of the system and reduce the cost of query processing. To expand the functionality of the system, the introduction of an image recognition module is envisaged, which will allow solving a wider range of tasks, including visual identification of numismatic materials.

In general, at this stage, the study has shown that the use of OpenAI's large language models is effective for the task of classifying numismatic materials based on textual descriptions. This approach provides accurate extraction of key attributes from different descriptions and offers significant promise for their use in similar tasks. Pre-trained language models show significant potential in automating tasks related to analyzing and interpreting textual information in different disciplinary contexts.

**List of References:**

1. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Dive into Deep Learning. Cambridge, UK: Cambridge University Press.
2. Dilmegan, C. (2024, January 11). Model Retraining: Why & How to Retrain ML Models? Retrieved from https://research.aimultiple.com/model-retraining.
3. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLAMA: Open and efficient foundation language models. Retrieved from arXiv:2306.07377.
4. OpenAI. (n.d.). Official Models Documentation. Retrieved from https://platform.openai.com/docs/models/.
5. Bozhko O. (2024) Using large language models to recognize information in numismatic descriptions. Science and technology today. Issue No 1(29) Retrieved from https://doi.org/10.52058/2786-6025-2024-1(29)