UDC 004.8:004.91
DOI: 10.15587/2706-5448.2025.341926

**Oleksandr Shyshatskyi,
Borys Moroz,
Maksym Ievlanov,
Ihor Levykin,
Dmytro Moroz**

EVALUATION OF THE EFFICIENCY OF LARGE LANGUAGE MODELS FOR EXTRACTING ENTITIES FROM UNSTRUCTURED DOCUMENTS

The object of research is arrays of unstructured documents located on public websites of rural and urban communities of Ukraine.

The study is devoted to solving the problem of choosing a large language model (LLM), which is the best for applied use in solving named entity recognition (NER) problems during document processing. Modern researchers recognize that such a choice is significantly influenced by the features of the subject area and the language of document creation. However, when studying the feasibility of using LLM to solve NER problems, the features of the operation of such models are practically not taken into account. The issues of evaluating such features remain largely unexplored.

A method for recognizing selected varieties of legal unstructured texts in the Ukrainian language is proposed. Unlike existing ones, this method solves the NER problem for those documents that are subject to recognition/classification. Metrics for the cost of processing input and output tokens are proposed and a methodology for evaluating the cost of using LLM is developed. Based on these results, a comparative evaluation of the application of common LLMs to solve the NER problem on Ukrainian texts that need to be recognized was conducted. According to the evaluation results, it was recognized that: (I) GPT-4o is the best in terms of accuracy and quality of processing (Precision = 0.919; Recall = 0.954; F1 = 0.936); (II) GPT-4o-mini with discounts is the best in terms of average document processing cost (0.00045 USD per document); (III) GPT-4.1-mini with discounts is the best in terms of quality/cost ratio (the indicator value is 0.938). The GPT-4.1-mini LLM is recommended as the best for applied application.

The evaluation results obtained allow to significantly simplify the choice of LLM, which is advisable to use for creating information systems and technologies for processing unstructured documents created in Ukrainian.

Keywords: *legal unstructured document, structured document annotation, token processing cost, GPT-4.1-mini.*

Received: 01.08.2025
Received in revised form: 08.09.2025
Accepted: 29.09.2025
Published: 29.12.2025

How to cite

Shyshatskyi, O., Moroz, B., Ievlanov, M., Levykin, I., Moroz, D. (2025). Evaluation of the efficiency of large language models for extracting entities from unstructured documents. Technology Audit and Production Reserves, 6 (2 (86)), 57–67. https://doi.org/10.15587/2706-5448.2025.341926

1. Introduction

The rapid growth of digital information, a significant part of which is presented in an unstructured form, creates a fundamental challenge for modern data processing systems. Unlike structured sources (relational databases, machine-readable document formats, etc.), unstructured text sources (reports, articles, emails, instant messaging messages, etc.) do not have a predefined structure [1]. This requires special approaches to automate the analysis and extraction of information contained in unstructured sources [1, 2]. The scale of this problem is enormous: according to the graph given in [3] (Fig. 1), the volume of data in the world from 2020 to 2025 will grow by 181.93%, and, based on the volume, 80% of the data is unstructured.

According to other estimates, unstructured data accounts for 90% of all data generated by enterprises [1]. Such estimates emphasize the need to develop effective solutions for managing this volume of data in order to obtain valuable information and knowledge from it.

Most often, the problem of extracting entities from unstructured documents in modern research is considered as the problem of named entity recognition (NER). The essence of NER is to process structured and unstructured data and classify named entities that were detected during the processing into predefined categories (names, locations, companies, time, monetary value, events, etc.) [4].

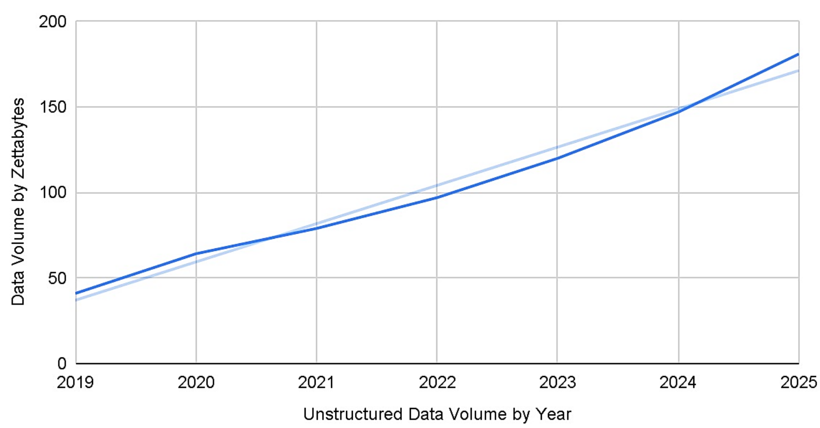

Fig. 1. Forecast and actual growth graphs of the volume of unstructured data in the world from 2019 to 2025 [3]

To solve the NER problem, modern experts use various approaches and technologies, among which in [4] they distinguish:

– dictionary-based systems (the simplest and most fundamental approach);
– rule-based systems (pattern-based rules, context-based rules, etc.);
– machine learning-based systems;
– deep learning technologies (RNN, Transformer, etc.);
– hybrid methods.

The selection of specific tools for solving the NER problem, taking into account the specifics of a specific subject area, conditions and limitations of the scientific and applied problem, is usually carried out on the basis of the following assessments [4]:

– *Accuracy*: measures how many entities identified by the NER model are actually correct, helping to assess the accuracy of the model in predicting named entities;
– *Recall*: assesses how many actual entities present in the text were successfully recognized by the NER model, indicating its ability to find all relevant entities;
– *F1 score*: provides a balanced indicator, combining precision and completeness, offering a single metric that reflects both precision and completeness.

In addition to this, metrics such as overall accuracy and average accuracy can provide additional insight into the performance of the NER model [4].

Practical experience in the use of NER problem-solving tools has allowed to identify their main advantages and limitations or challenges that negatively affect the applied application of such tools. Thus, in [4], the following are especially highlighted among the advantages of the applied application of NER problem-solving tools:

– the possibility of information extraction (identifies key data, facilitating information search);
– content organization (helps to classify content useful for databases and search engines);
– the expansion of user experience (clarifies search results and personalizes recommendations);
– the possibility of deep analysis (facilitates sentiment analysis and trend detection);
– the possibility of workflow automation (helps save time and resources).

As limitations or challenges of the applied application of NER problem-solving tools in [4], the following are highlighted:

– the problem of resolving ambiguity (the need to distinguish similar entities);
– problems of domain-specific adaptation (NER models and tools differ in resource intensity in different subject areas);
– language problems (the effectiveness of NER models and tools depends on slang and regional differences in the base language);
– the problem of labeled data scarcity (large labeled data sets are required to train NER models and tools);
– the need to apply advanced technologies for processing unstructured data;
– the problem of measuring performance (accurately assessing the performance of NER tools is a complex procedure);
– real-time processing problems (in particular, the problem of complex balancing of speed and accuracy of NER tools);
– the problem of context dependence (the accuracy of NER models and tools depends on understanding the nuances of the surrounding text);
– the problem of data sparsity (large labeled data sets are required, especially for specialized areas).

The considered advantages and problems of the applied use of NER models and tools have led to a large number of scientific and applied research studies devoted to solving individual issues of the NER problem. Thus, in the scientometric database Scopus alone, for the period from 2021 to 2025, 2463 publications (of which 694 were scientific articles) were recorded in the field of computer science under the keywords "Named Entities Recognition". The general taxonomy of the NER problem, which was built in [5] based on various training methods, modeling paradigms, and NER tasks, is shown in Fig. 2.

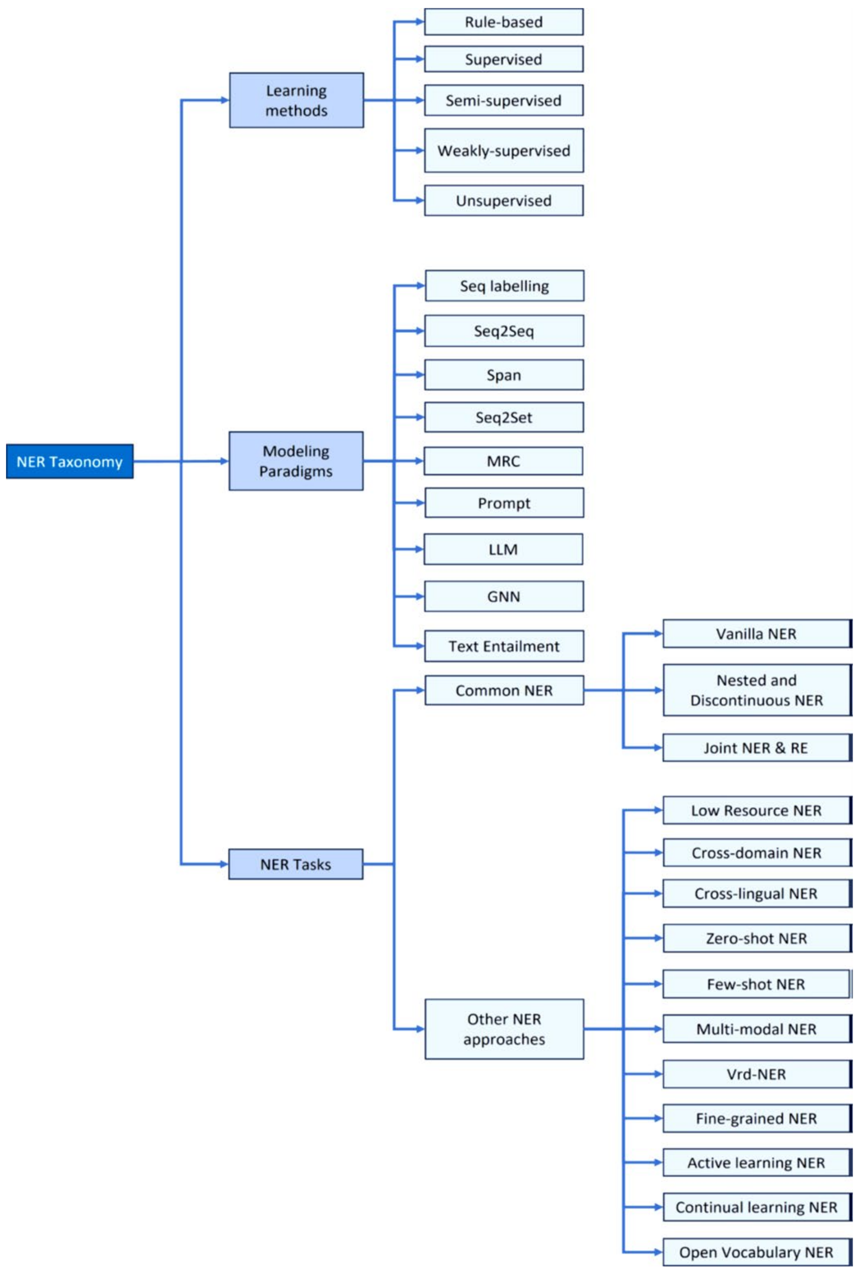

Fig. 2. NER taxonomy based on different training methods, modeling paradigms and NER tasks [5]

Based on this taxonomy, a study of the advantages and disadvantages of NER modeling paradigms was conducted in [5]. Of the entire set of studied paradigms, the "Large Language Models" (LLMs) paradigm should be especially highlighted. The advantages and disadvantages of this paradigm are given in Table 1 [5].

Table 1

Advantages and disadvantages of the NER modeling paradigm "LLMs"

Modelling paradigms	Advantages	Disadvantages
LLMs	– Scalable and versatile; – Able to fine tune and adapt to downstream tasks	– Large training dataset required; – High computational cost; – Bias and hallucination

Note: based on [5]

Such attention to the "LLMs" paradigm is due to its advantages, namely the scalability and universality of NER models built on the basis of this paradigm. These advantages are especially important for solving scientific and applied and applied NER problems in various subject areas. All other NER modeling paradigms focus on solving problems of individual aspects of the general NER problem. This complicates the application of models developed on the basis of these paradigms in cases of inconsistency of the conditions of a specific scientific and applied or applied problem with the features of the corresponding aspects of the NER problem. The universality of models based on the "LLMs" paradigm is confirmed by the results of a number of modern studies. Thus, in [6] it was proposed to use LLM to simplify the extraction of food objects from culinary recipes. The obtained research results allow automating decision-making support in the field of healthy and sustainable nutrition. The authors [6] claim that applying their proposed methodology, focused on rapid response engineering, to small LLMs with only 7b parameters allowed to increase their efficiency in terms of time, minimizing the required resources. However, these results, according to the researchers, are not conclusive and require further research for improvement.

The culinary subject area was used in [7] to comparatively study the capabilities of four LLMs (GPT-4o, GPT-4o-mini, Llama3.1:70b and Llama3.1:8b) to convert unstructured recipe text into a specialized structured Cooklang format. In this case, not only traditional metrics were used during the comparative evaluation, but also specialized metrics for identifying semantic elements. The following results were obtained during the study [7]:

– recognition of the ability of LLMs to reliably convert subject-oriented unstructured text into structured formats without significant training;
– LLM performance typically scales with size (experiments have shown that GPT-4o with few hints achieved breakthrough performance in ROUGE-L (0.97) and Word Error Rate (0.0730));
– in smaller models such as Llama3.1:8b, there is an amazing potential for optimization through targeted fine-tuning.

A limitation (and quite significant) of the study [7] is that it is focused on a specific subject area. And while the first of the study's results is confirmed by the results of similar studies in other subject areas, the other results require additional research in each individual subject area.

The use of an ontology-aware approach to zero-attempt LLM hints for processing Greek documents in the transportation industry is discussed in [8]. The solutions proposed in the study were found to be quite promising, as even small-sized LLM models showed very good results. However, such a semantic-oriented approach has several drawbacks, among which the authors of [8] highlight the following:

– extending ontologies to comply with evolving terminology and government policies in the transportation industry can be difficult;
– the proposed approach requires manual evaluation, so it is not easy to scale;
– some LLMs (e. g. Mistral 7B) had problems with irrelevant outputs and format inconsistencies.

The application of LLM to solve such a variant of the NER problem as the transformation of unstructured requirements texts in the aerospace industry into formal documents is discussed in [9]. This study proposed a novel approach to build aerospace-specific requirement knowledge graphs using LLM. The approach first uses the GPT model to augment data, followed by BERTScore filtering to ensure data quality and consistency. Then, efficient continuous learning based on token index encoding is implemented, which guides the model to focus on key information and improves domain adaptability by fine-tuning the Qwen2.5 model (7B). In addition, a chain of thought reasoning structure is established for improved entity and relationship recognition, combined with a dynamic multi-trial learning strategy that adaptively selects examples based on input characteristics. Experimental results confirm the effectiveness of the proposed method, achieving $F1$ scores of 88.75% in NER and 89.48% in relationship extraction tasks [9].

However, the results obtained by the authors of the study [9] are recognized as requiring further development in order to improve the performance of LLM in order to fully utilize its potential in the context of aerospace requirements engineering. In addition, the need for additional verification of the obtained results on larger and more diverse aerospace requirements datasets was recognized to comprehensively assess the generalizability and scalability of the obtained results in different systems and application areas [9].

A very large field for research in the field of LLM application to solve the NER problem is the medical field. It can be argued that the main volume of scientific research is devoted to the application of LLM to solve individual applied and scientific-applied problems in various fields of medicine. Thus, in [10] the application of LLM (in particular, GPT-4) to extract information from histopathological reports was considered, focusing on two large sets of pathological reports on colorectal cancer and glioblastoma. The study found a high correspondence between the structured data generated by LLM and the structured data generated by humans. However, the disadvantage of the study [10] is that it focuses on testing the hypothesis only about the possibility of using LLM to extract basic data for machine learning from unstructured pathological reports in the future. The issue of the application of LLM to solve the NER problem is not considered in [10].

In [11] examined the potential of LLM (in particular GPT-3.5 and GPT-4) in processing complex clinical data and extracting meaningful information with minimal training data. LLM was evaluated in [11] on two clinical NER tasks:

– extracting medical problems, treatments, and tests from clinical notes in the MTSamples corpus, according to the i2b2 2010 joint concept extraction task;
– detecting adverse events related to nervous system disorders from safety reports in the Vaccine Adverse Event Reporting System (VAERS).

Using the basic prompts, GPT-3.5 and GPT-4 achieved relaxed $F1$ scores of 0.634, 0.804 for MTSamples and 0.301, 0.593 for VAERS. Additional prompt components consistently improved the model performance. When all 4 components were used, GPT-3.5 and GPT-4 achieved relaxed $F1$ scores of 0.794, 0.861 for MTSamples and 0.676, 0.736 for VAERS, demonstrating the effectiveness of our cueing framework. Although these results lag behind BioClinicalBERT ($F1$ 0.901 for the MTSamples dataset and 0.802 for VAERS), they are very promising considering that few training samples are required [11]. However, the authors of the study [11] themselves acknowledge that while GPT-4 shows the potential to achieve performance close to that of the specialized LLM of BioClinicalBERT, GPT-4 still requires careful design of operational cues and understanding of task-specific knowledge. The study also highlights the importance of scoring schemes that accurately reflect the capabilities and performance of LLMs in clinical settings. The authors of [11] also recognized that direct application of GPT models to clinical NER tasks does not provide optimal performance.

In [12], a comparative evaluation of the use of different fine-tuned variations of generative LLMs in a NER task with a zero score for the clinical domain was considered. The Llama 2 and Mistral models were considered, including the base versions and those that were fine-tuned for code, chat, and instruction execution tasks. It was found that the fine-tuned instruction models performed better than the fine-tuned chat and base models in entity recognition. It was also shown that the models performed better when simple output structures were queried [12]. These findings require further testing and can be considered as hypotheses for further research in other subject areas.

The problem of developing a specialized NER tool based on deep learning and a lexicon for medical texts in Spanish is discussed in [13]. This tool uses a specialized lexicon and rules adapted from NegEx and HeidelTime. To train the tool, an annotated corpus of 1200 texts with high inter-annotation consistency was created (average $F1 = 0.841\% \pm 0.045$ for entities and average $F1 = 0.881\% \pm 0.032$ for attributes). This corpus was used to train models based on RoBERTa, mBERT and mDeBERTa. These models were integrated into the tool together with a dictionary-based system. During internal validation, the models gave $F1$ values up to 0.915. During external validation with 100 clinical trials, the tool achieved an average $F1$ score of 0.858 ($\pm$ 0.032); and in 100 anonymized clinical cases it reached an average $F1$ score of 0.910 ($\pm$ 0.019) [13]. The obtained results confirm the decrease in the values of LLM estimates when moving from verification and validation by developers to operation in the processes of customer organizations of such tools. Unfortunately, the problem of long-term effective operation of the developed tool is not considered in [13]. In addition, the results obtained in [13] indicate the feasibility of using not only the "LLMs" paradigm for solving applied NER problems, but hybrid solutions based on a combination of LLM and other paradigms. An interesting study of solving the NER problem in the field of classical literature text processing using LLM is considered in [14]. The study compared the Xunzi-Baichuan, Baichuan2-7B-Base, Xunzi-GLM and ChatGLM3-6B models. Experimental results showed that the fine-tuned LLMs achieved high scores on several indicators, demonstrating high performance in text generation. According to the researchers, the use of such LLMs can define new approaches to digital research of classical literature resources, interlinguistic understanding, textual knowledge extraction, and the promotion and preservation of cultural heritage. However, to obtain the published results, the considered LLMs were refined using supervised fine-tuning methods and tested on NER tasks using zero-, single-, and multiple-attempt hinting methods [14]. Such refinement significantly limits the possibilities of applied application of LLMs in solving NER tasks. It should be recognized that the application of LLMs to solve applied NER tasks faces significant limitations in the performance of LLMs compared to other NER modeling paradigms. Therefore, assessments of the performance of LLM-based NER tools and technologies obtained in laboratory studies require additional verification. One example of the validity of this statement is the aforementioned study [13]. Another similar example is the work [15], which investigated the performance of different encoder and decoder models trained for NER of clinical parameters in pathology and radiology reports. Three NER methods were evaluated: flat NER using transformer-based models; nested NER with a multi-task training system; and instruction-based NER using LLM. A dataset of pathology reports from 2013 and 413 radiology reports annotated by medical students were used for training and testing. The high-performance flat NER models achieved $F1$ scores of 0.87–0.88 in pathology reports and up to 0.78 in radiology reports, while nested NER models showed slightly lower results. Multilevel LLMs, despite achieving high accuracy, yielded significantly lower $F1$ scores (0.18 to 0.30) due to poor recall. One contributing factor is that these LLMs produce fewer but more accurate entities, suggesting that they become overly conservative in their output generation.

Therefore, [15] recognized that:

– LLMs in their current form are not suitable for complex entity extraction tasks in clinical domains, especially when faced with a large number of entity types per document;

– the computational cost of LLMs does not provide a proportional performance increase;

– encoder-based NER models, especially those pre-trained on biomedical data, remain the best choice for extracting information from unstructured medical documents.

Based on the results of the analysis of modern scientific research, the following conclusions can be drawn:

a) the "LLMs" paradigm really stands out among other paradigms in the versatility of NER models created on its basis (confirmed by the successful application of LLM to solve NER problems in various subject areas);

b) studies of the possibilities of using LLM to solve NER problems are mainly laboratory and do not take into account the applied aspects of the operation of relevant IT products and technologies;

c) the statement about the better scalability of the "LLMs" paradigm than other NER modeling paradigms requires additional research for individual subject areas;

d) for industrial operation, IT products for solving specific NER problems, which are based on LLMs, require significant refinement of these LLMs.

Therefore, from a scientific and applied point of view, it is relevant to study the features of using LLM to solve specific NER problems for each individual subject area. The results of such a study can not only contribute to the selection of a specific LLM for the development of a corresponding IT product, but also recognize the practical feasibility of using LLM to solve the selected NER problem.

The object of research is the arrays of unstructured documents located on public websites of rural and urban communities of Ukraine.

The subject of research is LLMs that can be used to solve the NER problem on these arrays.

The aim of research is a comprehensive assessment of the possibilities of using modern LLMs to solve the NER problem from unstructured legal texts in Ukrainian. This will allow justifying the choice of a specific LLM to solve the formulated problem according to the following criteria:

a) accuracy in performing the transformation of unstructured text into rigidly typed entities;

b) balance between the cost of processing and the accuracy of the information obtained.

To achieve this aim, it is proposed to solve the following objectives:

– develop a method for recognizing selected varieties of legal unstructured texts in Ukrainian;

– develop a methodology for assessing the cost of using LLM;

– conduct experimental studies and formulate recommendations on the possibilities of using different LLMs to solve the NER problem on a selected variety of legal unstructured text.

2. Materials and Methods

Achieving the stated aim of research requires conducting experiments on the use of different LLMs to process the same reference dataset to obtain results that can be compared with each other. To form such a reference dataset, data was collected from public web resources of rural and urban communities. Using a specialized program, 25,565 documents containing decisions made by local government bodies were downloaded. The documents were presented in various formats: pdf, doc, docx, rtf, jpg and png.

For further analysis, all downloaded document files were processed in a special way in order to create a text representation of these documents. Thus, for graphic formats (jpg, png) optical character recognition (OCR) was used, for text formats (docx, doc, some pdf) – appropriate libraries for conversion to text without using OCR. This made it possible to bring the array of primary representations of the collected documents to a unified text format.

From the resulting unified array, 350 documents containing the key phrase "cadastral number" and related to land relations were randomly selected. 150 random documents that do not contain information about cadastral numbers and are not related to land plots were additionally added to this sample. This approach allowed not only to test the ability of the studied LLM to find relevant information, but also to assess the stability of this LLM to false positive results.

The selected 500 texts were manually annotated by experts and converted into a machine-readable JSON format. This conversion allowed to create a structured data set necessary for testing algorithms for automatic information recognition. An example of the selected text and its annotation is presented in Fig. 3 and 4, respectively.

Each document annotation is an array of objects, the attributes of which contain detailed information about the land plot or related documentation. Information about the land plot is represented by the following attributes: cadastral number, area, purpose, form of ownership, numerical identifier. Information about the documentation related to the land plot is represented by the following attributes: document type, link to the corresponding cadastral plots through the "id" object. Assigning identifiers to both land plots and documents ensures the establishment of unambiguous relationships between entities.

Source text

CHORNOBAYIVKA VILLAGE COUNCIL

DECISION

dated 24.03.2022 No. 23 – 10/VIII

On approval of land management projects for the allocation of land plots from communally owned lands and transfer to private ownership

According to ..., the village council

DETERMINED:

1. To approve land management projects for the allocation of land plots from communally owned lands and transfer land plots to private ownership to citizens:

1 Kharchenko Denys Viacheslavovych for running a personal peasant farm (Code in Classifier of Types of Land Use – A, subdivision – 01.03) with an area of 2.0000 hectares (cadastral number of the land plot according to the Extract from the State Land Cadaster 7125185200:03:000:0302), located at the address: Novoselytsia village, within the administrative boundaries of Chornobayivka settlement council, Zolotonosha district, Cherkasy region;

2. Control over the implementation of the decision shall be entrusted to the permanent commission of the settlement council on issues of the agro-industrial complex, land relations, environmental protection, construction, transport, communications and housing and communal services.

Settlement head Serhiy LIUBYVYI

Fig. 3. An example of a village council decision text selected for further annotation (with abbreviations marked with an ellipsis)

Source text

```
[
{
"number": "7125185200:03:000:0302",
"area": 2.0,
"area_unit": "ha",
"purpose_code": "01.03",
"category": " Agricultural land ",
"ownership": " Private property ",
"id": 0,
"type": "parcel"
},
{
"documentation_type": "LAND_PLOT_ALLOCATION_PROJECT",
"involved_parcels": [
0
],
"id": 0,
"type": "documentation"
}
]
```

Fig. 4. Example of annotation of the text of the selected document in JSON format

The resulting dataset and the documents on the basis of which this dataset was created are available in an open form in a public repository [15].

For the experimental study, LLMs from four leading companies in the field of artificial intelligence were selected: OpenAI, DeepSeek, Google, and xAI. The sample included models that were relevant as of the end of spring 2025. The key criteria for selecting these LLMs were the possibility of their use via API and the technical ability to generate data in a structured format. According to these criteria, the following LLMs were selected for further research: deepseek-chat, gemini-2.5-flash, gpt-4.1-2025, gpt-4.1-mini, gpt-4o-mini, gpt-4o, grok-3 and grok-3-mini.

To assess the accuracy of each of the LLMs used in experimental studies, it was proposed to apply the $F1$ metric – the harmonic mean between precision and recall. The $F1$ metric was calculated separately for the values of each of the following attributes:

– "Cadastral number" ("number");
– "Area of the plot" ("area");
– "Unit of area measurement" ("area_unit");
– "Purpose code" ("purpose_code");
– "Land category" ("category");
– "Ownership" ("ownership");
– "Document type, the action on which is performed in the solution" (document_type);
– identifiers of the plots used by the found document.

This application of the $F1$ metric allowed to analyze in detail the quality of extraction of different types of information in the process of solving the NER problem.

As a result of experimental studies, each of the used LLMs formed document annotations, the structure of which is similar to the structure of the annotation shown in Fig. 4. As a result of such annotation, arrays of links to the corresponding cadastral plots were formed. Comparison of these arrays with the array of links to the corresponding cadastral plots in the annotations created by experts allowed to determine the following fundamental parameters: True Positives (TP), False Positives (FP), False Negatives (FN):

– TP: the number of detected matches of identifier values detected by experts and using LLM (cases when the model correctly extracted information);
– FP: the number of identifier values detected using LLM that did not coincide with the identifier values detected by experts (cases when the model made an erroneous extraction);
– FN: the number of identifier values detected by experts that are missing from the array of identifier values detected by the LLM (cases where the model missed relevant information).

These parameters are used to determine the "Precision" and "Completeness" metrics of each of the studied LLMs.

The "Precision" metric characterizes the proportion of all correct classifications, both positive and negative, out of the total number of classifications. The value of this metric should be calculated using the following formula

$$\text{Precision} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TN – the True Negatives parameter (calculated as the difference between the total number of negative classification results and the value of the FP parameter).

In this study, the value of the "Precision" metric means the proportion of plots for which the studied LLM correctly determined the presence or absence of a certain attribute (for example, purpose or ownership).

However, the Precision metric can be misleadingly high if the data source is unbalanced. For example, if most plots do not have a given land category, a model that always returns "no data" will have high accuracy, but will be of little use for practical applications. Therefore, in this study, the Precision metric will only be useful in combination with other metrics.

Another such metric is the Recall metric, which characterizes the proportion of actual positive cases that the model correctly

ISSN-L 2664-9969; E-ISSN 2706-5448

classified. The value of this metric should be calculated using the following formula

$$\text{Recall} = \frac{TP}{TP + FN}. \qquad (2)$$

In the context of this study, the "Recall" metric characterizes the proportion of important information (e. g., target designation) that was present in the annotations of a person that the LLM under study was able to find. For example, if among 100 sites 60 have a target designation code, and using the LLM under study, only 45 of them were found, the value of the "Recall" metric will be 45/60 = 0.75. This metric is especially important for cases where the omission of relevant information is critical. The $F1$, "Accuracy" and "Recall" metrics, as indicated in [4], are the most frequently used metrics for quantifying the results of using LLM to solve the NER problem. However, these metrics are insufficient to highlight the feasibility of applying the LLM under study to solve specific NER problems. Therefore, there is a need to develop an additional methodology for evaluating the applied aspects of using LLM to solve a specific NER problem.

3. Results and Discussion

3.1. Method for recognizing legal unstructured documents in Ukrainian

To solve NER tasks in texts with a complex structure, where there are several types of interconnected entities, it was proposed to apply a multi-stage approach. The basis is the principle of decomposition – breaking a single complex task into a sequence of simpler, highly specialized subtasks. This approach allows to minimize the task set for LLM at each stage, which contributes to increasing the accuracy and reliability of recognition.

The architecture of the method is a pipeline, where the result of the previous stage serves as additional input data for the next. The number of pipeline stages corresponds to the number of entity types that need to be recognized. Each pipeline stage has a unified structure of three steps:

– *Step 1*: preparation of input parameters for the model (formation of a query that includes both the source text and structured data obtained at the previous stages).

– *Step 2*: execution of the NER task by the selected LLM model (recognition of target entities of the current stage).

– *Step 3*: post-processing of the data received from the LLM (structuring the data received from the model and preparing it for the next stage or for the final output).

The ultimate goal of the pipeline is to transform the unstructured input text into a single structured array of unified records that unites all recognized entities. The goal of each of the stages of the proposed method is to form a structured array of unified records from fragments of unstructured text containing key attributes of the entity, the type of which is recognized at the current stage of the pipeline.

The scheme of the method on the example of three stages for recognizing three types of entities is shown in Fig. 5.

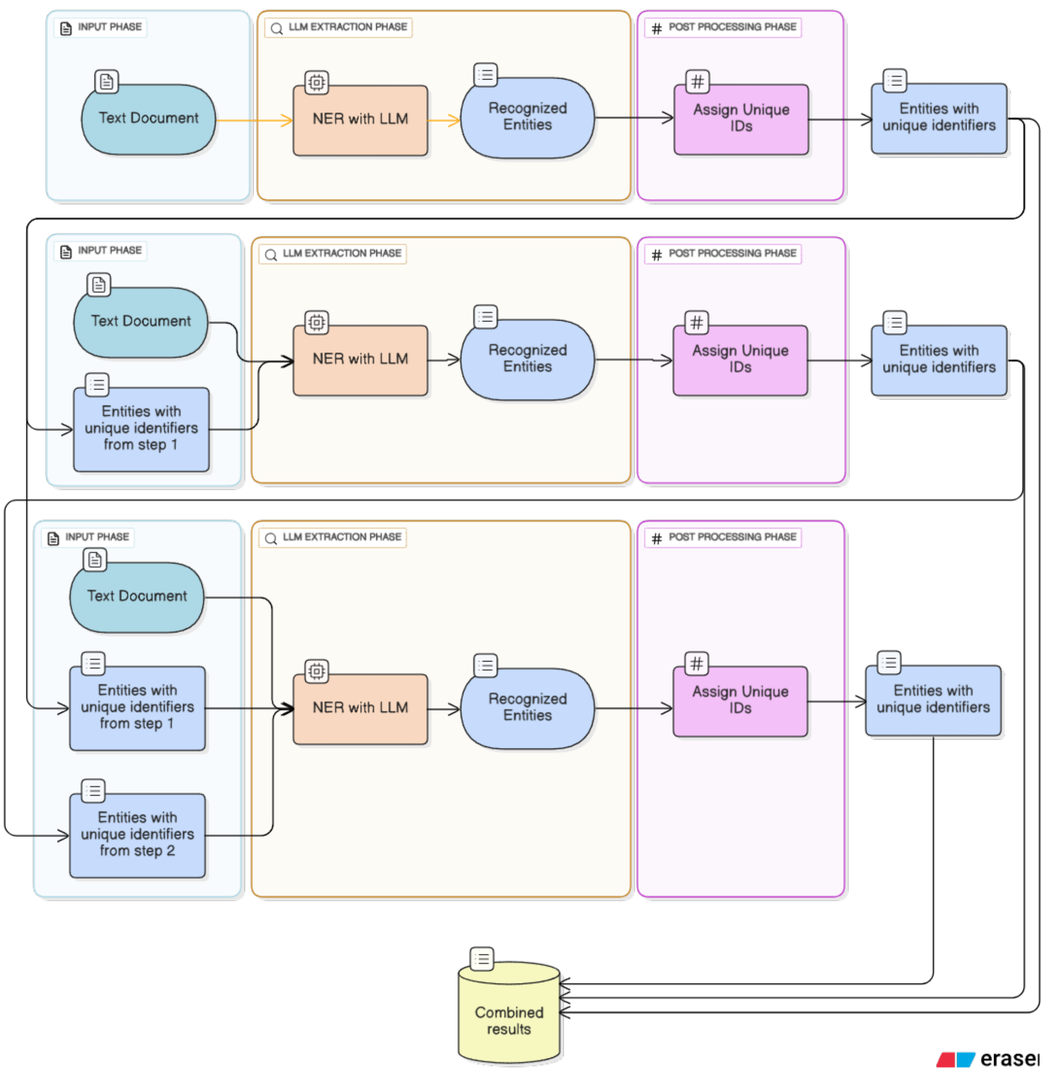

Fig. 5. Generalized block diagram of the proposed method (using the example of recognizing 3 different entities)

Let's consider the application of the method using the example of legal documents from the experimental dataset. The structure of these texts assumes the presence of two key types of entities: land plots and land management documents, which are the grounds for actions regarding these plots. Accordingly, the experimental pipeline consists of two stages.

At the first stage, the NER task is performed using LLM to extract information about land plots. The input data is the text of the legal document, and the result of the LLM operation is an array of structured data regarding the plots. At the post-processing stage, each recognized plot is assigned an identifier (id).

This allows:

– to focus the model on extracting attributes, without overloading it with the task of generating unique links;

– to process texts where the cadastral number is missing (for example, for plots that have not yet been registered).

For each plot, the following attributes are highlighted from the text:

– "Cadastral number" ("number");

– "Area of plot" ("area");

– "Area unit" ("area_unit");

– "Purpose code" ("purpose_code") – if available;

– "Land category" ("category") – if available;

– "Ownership" ("ownership");

– "Automatically assigned identifier" (id).

If certain information is missing from the text of the document (for example, the ownership form or land category is not specified), the corresponding attribute should be left empty.

The scheme of the first stage is shown in Fig. 6.

An example of the unstructured text of a document that is fed to the LLM input at the first stage of the proposed method is shown in Fig. 7.

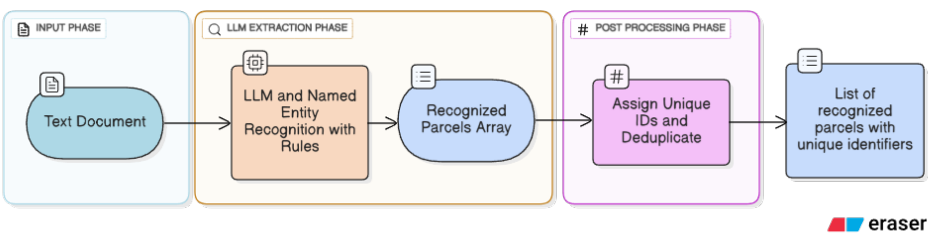

Fig. 6. Scheme of the first stage of the legal document recognition pipeline (using land management documents as an example)

An example of solving the NER problem using the LLM under study in the form of a structured annotation of the submitted document with the allocation of the land plot identifier is shown in Fig. 8.

At the second stage, a more complex NER task is performed: mentions of land management documents are detected in the text and are associated with previously identified plots. The input data for the model is the text of the legal document and an array of structured data on plots obtained at the previous stage. The result of its work is an array of structured data on land management documents, including identifiers of land plots to which the recognized document belongs.

DECISION No. 1676 from February 23, 2022

On approval of the land management project for the allocation of a land plot into ownership.

Having considered the application of the citizen Shafranska Halyna Mykhailivna for the approval of the land management project for the allocation of a land plot into ownership with an area of 0.3000 hectares for conducting a personal peasant farm in the territory of the Ostashivka village council.

D E C I D E D :

1. To approve the gr. Shafranska Halyna Mykhailivna a land management project for the allocation of a land plot into ownership with an area of ***0.3000 hectares****, cadastral number of the land plot* ***6122687300:01:002:0091*** *for conducting a personal peasant farm in the territory of the Ostashiv village council.*

2. ***To transfer the ownership*** *of the citizen Shafranska Halyna Mykhailivna a land plot with an area of 0.3000 hectares, cadastral number of the land plot* ***6122687300:01:002:0091****, for running a personal farm on the territory of the Ostashivka Council.*

Fig. 7. An example of an unstructured text of a document that is fed to the input of the first stage of the pipeline of the proposed method


```
[
{
"number": "6122687300:01:002:0091",
"area": "0.3000",
"area_unit": "ha",
"purpose_code": null,
"category": " Agricultural land ",
"ownership": " Private property "
}
]
```


Fig. 8. An example of solving the NER problem using LLM as a result of using the first stage of the pipeline of the proposed method

For each land management document, the following attributes are extracted from the text:

– "Documentation type" ("documentation_type");

– "Plot identifiers" ("involved_parcels");

– "Automatically assigned identifier" (id).

The scheme of using the proposed method is shown in Fig. 9.

Unstructured texts of documents that are fed to the input of the LLM at the second stage of the pipeline have a form similar to the example shown in Fig. 7.

An example of solving the NER problem using the LLM under study in the form of a structured annotation of the submitted document with the selection of the identified land management document is shown in Fig. 10.

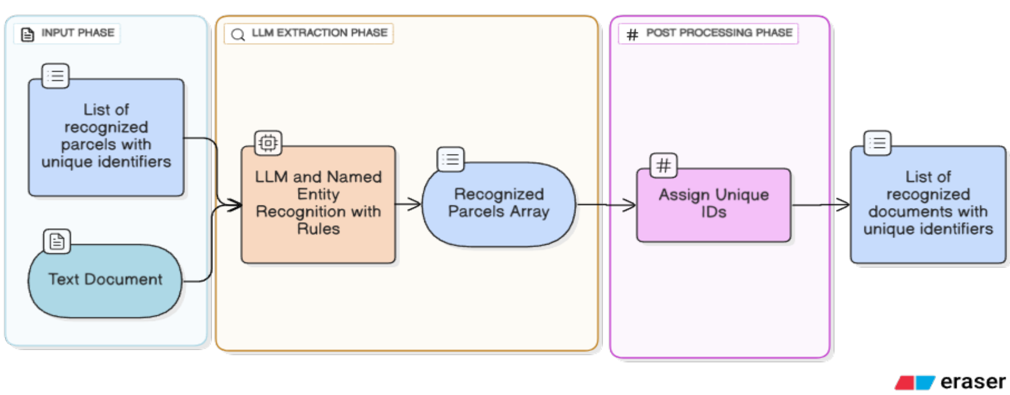

Fig. 9. Scheme of the second stage of the method for recognizing legal documents (using the example of land management documents)

ISSN-L 2664-9969; E-ISSN 2706-5448

After completing all stages of the pipeline, their results are combined into a single structured array containing complete information about all recognized entities and their relationships (Fig. 11).

[
{
"documentation_type": "LAND_PLOT_ALLOCATION_PROJECT",
"involved_parcels": [
"e769b73628834aaeaa2c6f13e2a7f794"
],
"id": "3e82c76efc464434808c946989a328bd"
}
]

Fig. 10. An example of solving the NER problem using LLM as a result of using the second stage of the pipeline of the proposed method

[
{
"number": "6122687300:01:002:0091",
"area": "0.3000",
"area_unit": "ha",
"purpose_code": null,
"category": " Agricultural land ",
"ownership": " Private property ",
"id": "e769b73628834aaeaa2c6f13e2a7f794"
},
{
"documentation_type": "LAND_PLOT_ALLOCATION_PROJECT",
"involved_parcels": [
"e769b73628834aaeaa2c6f13e2a7f794"
],
"id": "3e82c76efc464434808c946989a328bd"
}
]

Fig. 11. An example of a structured array obtained as a result of the proposed method

3.2. Applied methodology for estimating the cost of using LLM based on the results of experimental studies

A corresponding methodology was proposed to estimate the cost of using different LLMs. This methodology was based on the facts of publication by LLM supplier companies of tariffs for the cost of processing 1 million tokens. The values of these tariffs are established in the official documents of such companies (OpenAI, Google Gemini, XAI Grok, DeepSeek, etc.). In addition, the official documents also establish the sizes of possible discounts of these tariffs for individual tariff plans.

To calculate the cost of using LLM for processing incoming (prompt) tokens, in this study it is proposed to use the metric of the cost of processing incoming (prompt) tokens C_{In}. The value of this parameter is calculated by the formula

$$C_{In} = \frac{Am_{PT}}{1000000} \cdot \left(Tar_{PT} \cdot D_{PT}\right), \tag{3}$$

where Am_{PT} – the number of incoming (prompt) tokens of the document processed by LLM; Tap_{PT} – the value of the tariff for processing one million prompt tokens using the LLM under study; D_{PT} – the value of the discount valid for the selected tariff plan for using the LLM under study during the experimental studies.

Similarly, it was proposed to use the C_{Out} metric for calculating the cost of using the LLM for processing completion tokens in this study. The value of this parameter is calculated by the formula

$$C_{Out} = \frac{Am_{CT}}{1000000} \cdot \left(Tar_{CT} \cdot D_{CT}\right), \tag{4}$$

where Am_{CT} – the number of output (completion) tokens of the document processed by the LLM; Tap_{CT} – the value of the tariff for processing one million output (completion) tokens using the LLM under study; D_{CT} – the value of the discount valid for the selected tariff plan for using the LLM under study during the experimental studies.

In the absence of discounts for the selected tariff plan for using the LLM under study, during the experimental studies, DPT and DCT take the value 1.

Based on parameters (3) and (4), the proposed methodology for estimating the cost of using the LLM was presented as a sequence of the following steps:

Step 1. Collect information about tariff plans and discounts for using the LLM under study that will be valid during the experimental studies.

Step 2. In the process of conducting experimental research during the extraction, for each document, generate a separate file with information about the use of the model ('_usage.json'), which contains the number of input (prompt) and output (completion) tokens.

Step 3. For each of the values of the number of input (prompt) and output (completion) tokens obtained in Step 2, calculate the cost of their processing by the LLM under study using expressions (3) and (4).

Step 4. Obtain the value of the total cost of using the LLM under study as the sum of the calculation results obtained in Step 3. Complete the use of the methodology.

The application of the proposed methodology allowed both to compare the economic efficiency of the LLMs under study and to estimate the individual cost of processing the LLM of individual documents.

3.3. Conducting experimental studies

To conduct a comparative evaluation, it was decided to investigate the capabilities of the following LLMs when using the proposed method of recognizing individual types of legal unstructured documents in Ukrainian: deepseek-chat, gemini-2.5-flash, gpt-4.1-2025, gpt-4.1-mini, gpt-4o-mini, gpt-4o, grok-3 and grok-3-mini. As a reference data set, it was proposed to use a dataset collected by one of the authors of this study, available for public use [15].

The values of the cost of tariff plans and current discounts were obtained from the offers provided by Google [16], OpenAI [17] and DeepSeek [18] under the condition of batch processing or at certain hours.

The values of the metrics "Precision", "Recall" and $F1$ for the results of experimental studies of the selected LLMs are shown in the form of a heat map in Fig. 12.

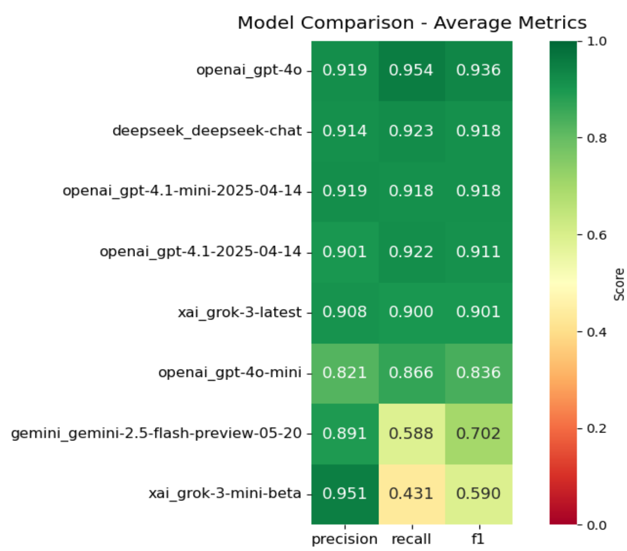

Fig. 12. Model Comparison: Average Precision, Recall, and $F1$ for each model (more is better)

The results demonstrate the variability in the accuracy of solving the NER problem by different models. GPT-4o showed the best result with an $F1$ metric value of 0.936. It is worth noting that the more compact version of GPT-4.1-mini achieved almost similar results with an $F1$ metric value of 0.918.

In second place with an $F1$ metric value of 0.918 is the DeepSeek Chat model, which is only 0.018 less than the result of the GPT-4o model.

The gemini-2.5-flash and grok-3-mini models demonstrated significantly lower indicators, especially in terms of the "Recall" metric value: 0.588 and 0.431, respectively. This may indicate their tendency to miss relevant information, which is a significant drawback.

The economic efficiency of the models was assessed by calculating the average cost of processing one document from the test dataset using the proposed methodology.

The calculation results are shown in Fig. 13 (hereafter, discounts are marked as Discounted).

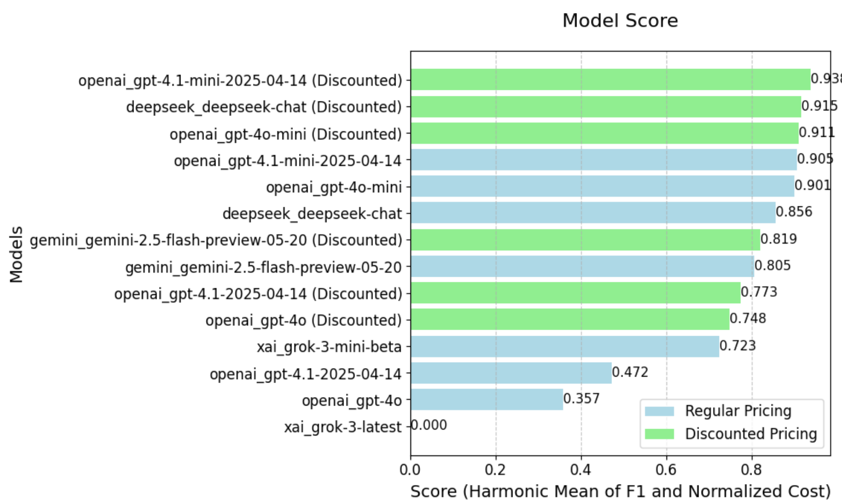

Fig. 14. Model Score: Harmonic mean of $F1$ and normalized cost (more is better)

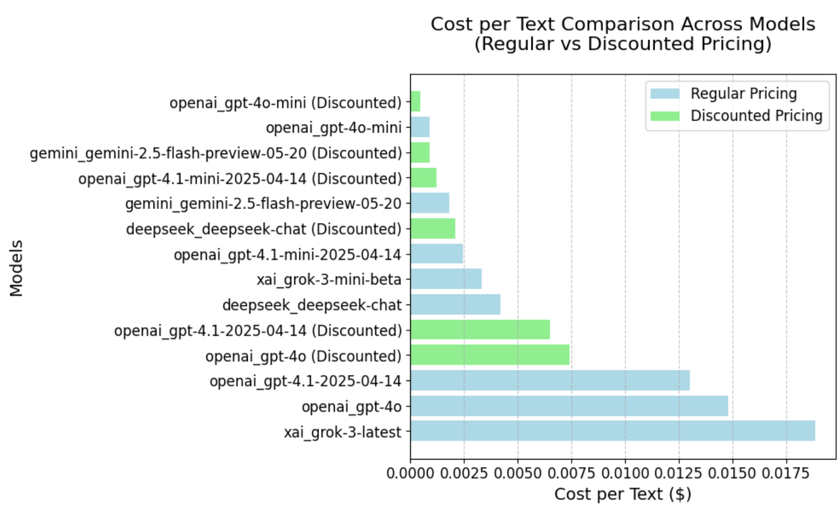

Fig. 13. Cost per Text Comparison Across Models (less is better)

Based on the assessments shown in Fig. 14, the following recommendations were proposed for choosing an LLM for practical application when solving the NER problem for unstructured documents in the Ukrainian language (Table 2).

Table 2

Recommendations for choosing an LLM for practical application when solving the NER problem for unstructured documents in the Ukrainian language

LLM place among the researched	LLM title without discounts	LLM title with discounts
1st place	GPT-4.1-mini	GPT-4.1-mini
2nd place	GPT-4o-mini	deepseek-chat
3rd place	deepseek-chat	GPT-4o-mini

The analysis of processing costs shows a significant difference between the models. The most expensive, as expected, are the full versions of the models, in particular Grok-3 (0.019 USD per document) and GPT-4o (0.015 USD per document). In contrast, their mini versions offer significantly lower costs: GPT-4o-mini costs only 0.00089 USD per document, which is 16.8 times less than the full version.

Special attention is paid to the discount system, which significantly affects the economic feasibility of using different models. For example, the application of discounts for GPT-4o-mini reduces the cost to 0.00045 USD per document, which makes this model the most economical choice among all the studied options (Fig. 13). To determine the optimal ratio between the quality and cost of data processing, it was proposed to calculate the value of the harmonic mean between the values of the $F1$ metric and the normalized values of the cost of information processing for the corresponding LLMs. This indicator would allow to identify those LLMs that provide the best price-quality ratio during their application. The results of calculating the values of this harmonic mean are shown in Fig. 14.

The values of the harmonic mean between the $F1$ metric values and the normalized information processing cost values for the corresponding LLMs shown in Fig. 13 are better than the most common metrics in terms of the practical feasibility of using individual LLMs to solve a specific NER problem. For example, the grok-3 model, which was ranked 5th in terms of accuracy (Fig. 12), received 0 points according to the assessment shown in Fig. 14, since its cost is the highest, and the results are not much better than GPT-4.1, the price of which is 30% lower.

3.4. Discussion of the research results

It was conducted a comprehensive assessment of the possibilities of using modern LLMs to solve the NER problem of unstructured legal texts in Ukrainian. Such a comprehensive assessment became necessary to understand the feasibility of applying modern LLMs in information systems and technologies used to process unstructured or weakly structured documents. To conduct such an assessment, it was developed a method for recognizing selected varieties of legal unstructured texts in Ukrainian. The developed method, unlike common methods of classifying or categorizing documents, at its first stage solves the NER problem for those unstructured documents that are subject to recognition (i. e. classification). Such an improvement allowed for the formation of a rigidly structured annotated description for each document. This, in turn, allowed for a significant simplification of the solution of the recognition (classification) problem in the second stage of the method.

Since the most common metrics for quantitative evaluation of LLMs do not allow to evaluate these models from the point of view of their applied application, the study proposed metrics for the cost of processing input (prompt) tokens C_{In} and output (completion) tokens C_{Out}. Based on these metrics, a methodology for assessing the cost of using LLMs was developed. This methodology allowed to assess both the economic efficiency of the LLMs under study and the individual cost of processing individual documents by the LLMs under study.

A special dataset was developed to conduct experimental research and develop recommendations on the practical feasibility of using individual LLMs to solve the NER problem on a selected type of legal unstructured text. This dataset is based on 25,565 documents containing decisions made by local government bodies. Of these, 350 documents containing the key phrase "cadastral number" were randomly selected for further experiments. This sample was supplemented with 150 randomly selected documents from the dataset that do not contain information about cadastral numbers and are not related to land plots. The selected 500 texts were manually annotated by experts and converted into a machine-readable JSON format. This dataset is freely available [15] and can be used for further experimental research.

For experimental research, it was recommended to choose LLM deepseek-chat, gemini-2.5-flash, gpt-4.1-2025, gpt-4.1-mini, gpt-4o-mini, gpt-4o, grok-3 and grok-3-mini. As a result of using the selected LLMs, it was recognized that according to the evaluations of the metrics "Precision", "Recall" and $F1$, the best for solving the NER problem is LLM GPT-4o. As a result of using the developed methodology, it was recognized that the best LLM from an economic point of view for applied application is the GPT-4o-mini LLM under the conditions of discounts. But this model is only in fifth place in terms of the values of the "Precision", "Recall" and $F1$ metrics. Therefore, an evaluation was carried out based on the values of the harmonic mean between the values of the $F1$ metric and the normalized values of the cost of information processing for the corresponding LLMs. This indicator made it possible to identify those LLMs that provide the best "price-quality" ratio during their applied application. The evaluation results made it possible to select the three best LLMs according to this indicator under the conditions of the presence and absence of discounts. In both cases, the first place was taken by the GPT-4.1-mini LLM.

It should be recognized that the evaluation results for the data processing quality metrics ("Precision", "Recall" and $F1$) obtained in this study generally correlate with the results presented in [6, 7]. This may indicate the reliability of the conducted studies and the obtained results. However, the results of the evaluation of economic characteristics are somewhat different from the results presented in [7]. This difference is explained by the fact that in [7] even the simplest operational characteristics of LLM were not taken into account during the evaluation. It is worth noting that the choice of a specific LLM should take into account the specific requirements as put forward for the IT project (the required amount of data for processing, the permissible level of errors, budget constraints, etc.). The results of this study can serve as a guideline for making informed decisions regarding the choice of a model for specific application cases. The study has a number of limitations. First, this is the limited data. The evaluation in this study was conducted exclusively on Ukrainian legal documents. The influence of the language of the text on the results and effectiveness of the models in Ukrainian texts from other fields was not investigated. Secondly, this is the limitation of technical aspects. In particular, when assessing the feasibility of the applied application, the parameters of the LLM speed were not taken into account. Thirdly, this is the limitation of security and ethics issues. In particular, this study did not consider the issues of data protection and the ethics of using LLM.

Based on these limitations, it is proposed to consider the following promising areas of further research:

- research on expanding the domain (in particular, conducting a comprehensive assessment of the feasibility of the applied application of LLM on other types of documents from various fields; evaluating the results of applying LLM to processing documents created in different languages);
- further technical improvements to the obtained scientific results (in particular, studying the impact of fine-tuning on the accuracy of extraction).

4. Conclusions

1. A method for recognizing selected varieties of legal unstructured texts in Ukrainian has been developed. Unlike common methods of document classification or rubrication, the developed method at its first stage solves the NER problem for those unstructured documents that are subject to recognition/classification. Such an improvement allowed to form a rigidly structured annotated description for each document. This, in turn, allowed to significantly simplify the solution of the recognition/classification problem in the second stage of the method.

2. Metrics for the cost of processing input (prompt) tokens C_{In} and output (completion) C_{Out} have been proposed. Based on these metrics, a methodology for estimating the cost of using LLMs was developed. This methodology allowed to estimate both the economic efficiency of the LLMs under study and the individual cost of processing individual documents by the LLMs under study.

3. Using the obtained results, a comparative evaluation of the application of common LLMs to solve the NER problem of creating a structured annotation of texts in the Ukrainian language that need to be recognized was carried out. The experiments used LLMs deepseek-chat, gemini-2.5-flash, gpt-4.1-2025, gpt-4.1-mini, gpt-4o-mini, gpt-4o, grok-3 and grok-3-mini. According to the evaluation results, it was recognized that:

a) in terms of accuracy and quality of processing, the best is LLM GPT-4o (metric values: Precision = 0.919; Recall = 0.954; $F1$ = 0.936);

b) in terms of the average cost of processing one document from the test dataset, the best is LLM GPT-4o-mini, subject to the application of discounts (the total cost of processing is 0.00045 USD per document);

c) according to the harmonic mean between the $F1$ metric values and the normalized information processing cost values, the best LLM is GPT-4.1-mini under the condition of applying discounts (the value of the indicator is 0.938).

According to the results of the comprehensive evaluation, it was recommended to use the three best LLMs for applied application to solve the NER problem considered in this study: GPT-4.1-mini; deepseek-chat and GPT-4o-mini.

Conflict of interest

The authors declare that they have no conflict of interest regarding this research, including financial, personal, authorship or other nature, which could affect the research and its results presented in this article.

Financing

The research was conducted without financial support.

Data availability

The manuscript has related data in the data repository.

Use of artificial intelligence

The authors used artificial intelligence technologies within the permissible framework to provide their own verified data, which is described in the research methodology section.

Authors' contributions

Oleksandr Shyshatskyi: Software, Validation, Formal analysis, Resources, Data Curation, Writing – original draft; ***Borys Moroz***: Conceptualization, Methodology, Writing – original draft, Supervision, Project administration; ***Maksym Ievlanov***: Methodology, Investigation, Resources, Writing – review & editing; ***Ihor Levykin***: Methodology, Investigation, Validation, Resources, Funding acquisition; ***Dmytro Moroz***: Software, Validation, Writing – original draft, Funding acquisition.

References

1. Jonker, A., Gomstyn, A. (2025). *Structured vs. unstructured data: What's the difference?* IBM. Available at: https://www.ibm.com/think/topics/structured-vs-unstructured-data Last accessed: 26.08.2025
2. *What is text mining?* IBM. Available at: https://www.ibm.com/think/topics/text-mining Last accessed: 26.08.2025
3. *What Percentage of Data is Unstructured?* 3 Must-Know Statistics (2024). Edge Delta. Available at: https://edgedelta.com/company/blog/what-percentage-of-data-is-unstructured Last accessed: 26.08.2025
4. *Shcho take rozpiznavannia imenovanykh sutnostei (NER) – pryklad, vypadky vykorystannia, perevahy ta problemy* (2025). Shaip. Available at: https://uk.shaip.com/blog/named-entity-recognition-and-its-types/ Last accessed: 26.08.2025
5. Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., Cambria, E. (2025). A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review, 58 (10).* https://doi.org/10.1007/s10462-025-11321-8
6. Pitsilou, V., Papadakis, G., Skoutas, D. (2024). Using LLMs to Extract Food Entities from Cooking Recipes. *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW).* Utrecht, 21–28. https://doi.org/10.1109/icdew61823.2024.00008
7. Brach, W., Košťál, K., Ries, M. (2025). The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats. *IEEE Access, 13,* 91808–91825. https://doi.org/10.1109/access.2025.3573030
8. Zeginis, D., Kalampokis, E., Tarabanis, K. (2024). Applying an ontology-aware zero-shot LLM prompting approach for information extraction in Greek: the case of DIAVGEIA gov gr. *Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics.* New York, 324–330. https://doi.org/10.1145/3716554.3716603
9. Liu, Y., Hou, J., Chen, Y., Jin, J., Wang, W. (2025). LLM-ACNC: Aerospace Requirement Texts Knowledge Graph Construction Utilizing Large Language Model. *Aerospace, 12 (6),* 463. https://doi.org/10.3390/aerospace12060463
10. Truhn, D., Loeffler, C. M., Müller-Franzes, G., Nebelung, S., Hewitt, K. J., Brandner, S. et al. (2023). Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). *The Journal of Pathology, 262 (3),* 310–319. https://doi.org/10.1002/path.6232
11. Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X. et al. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association, 31 (9),* 1812–1820. https://doi.org/10.1093/jamia/ocad259
12. del Moral-González, R., Gómez-Adorno, H., Ramos-Flores, O. (2025). Comparative analysis of generative LLMs for labeling entities in clinical notes. *Genomics & Informatics, 23 (1).* https://doi.org/10.1186/s44342-024-00036-x
13. Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A. (2025). Hybrid natural language processing tool for semantic annotation of medical texts in Spanish. *BMC Bioinformatics, 26 (1).* https://doi.org/10.1186/s12859-024-05949-6
14. Xu, Q., Liu, Y., Wang, D., Huang, S. (2025). Automatic recognition of cross-language classic entities based on large language models. *Npj Heritage Science, 13 (1).* https://doi.org/10.1038/s40494-025-01624-y
15. Shyshatskyi, O. (2025). *Dataset and additional materials.* GitHub. Available at: https://github.com/oshyshatskyi-phd/public-docs-processing Last accessed: 26.08.2025
16. *Gemini models that support batch predictions.* Google Cloud. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/batch-prediction-gemini#models_that_support_batch_predictions Last accessed: 21.06.2025
17. *Pricing.* OpenAI platform. Available at: https://platform.openai.com/docs/pricing Last accessed: 21.06.2025
18. *Models & Pricing.* Deepseek API Docs. Available at: https://api-docs.deepseek.com/quick_start/pricing Last accessed: 21.06.2025

Oleksandr Shyshatskyi, *PhD Student, Department of Software Engineering, Dnipro University of Technology, Dnipro, Ukraine, ORCID: https://orcid.org/0009-0008-6008-7079*

Borys Moroz, *Doctor of Technical Sciences, Department of Software Engineering, Dnipro University of Technology, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0002-5625-0864*

✉***Maksym Ievlanov,*** *Doctor of Technical Sciences, Department of Information Control Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, e-mail: maksym.ievlanov@nure.ua, ORCID: https://orcid.org/0000-0002-6703-5166*

Ihor Levykin, *Doctor of Technical Sciences, Department of Media Systems and Technologies, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0001-8086-237X*

Dmytro Moroz, *PhD, Department of Software Engineering, Dnipro University of Technology, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0003-2577-3352*

✉*Corresponding author*