UDC 004.89
DOI: 10.15587/1729-4061.2025.337915

AUTOMATED EXTRACTION OF KEY PARAMETERS AND DETECTION OF INCONSISTENCIES IN CLINICAL DOCUMENTATION USING LARGE LANGUAGE MODELS

Vitaliy Horlatch
PhD*
Vasyl Pasichnyk*
Corresponding author
*Department of Information Systems
Ivan Franko National University of Lviv
Universytetska str., 1, Lviv, Ukraine, 79000
E-mail: vasyl.pasichnyk.apmi@lnu.edu.ua

This study investigates unstructured text data on clinical trials. The task addressed relates to the fact that analyzing such data involves a laborious and error-prone process, hard-to-tackle even for specialists. In turn, this leads to an increase in the duration of studies and delays in the release of new drugs to the market.

This work reports an approach to constructing a dataset on clinical trials, as well as subsequent extraction of key information using state-of-the-art large language models. A study was conducted on extracting such indicators as the eligible gender of participants, a research phase, as well as the study's therapeutic area. A total of 11,703 experiments were performed, most of which achieved high results. In particular, the average values when using the GPT-4o-mini model were as follows: F1-measure – 0.92; accuracy – 0.98; recall – 0.99; precision – 0.87.

Extraction of information from clinical documentation in Ukrainian demonstrated similar results compared to English-language counterparts. In some cases, a significant number of false positives were observed, and the indicators were significantly lower (the lowest recorded values: F1-measure – 0.52; accuracy – 0.82; recall – 0.78; precision – 0.35). For such cases, the reasons were analyzed, and the corresponding conclusions and recommendations were formulated.

In addition, the results of the experiments helped identify a number of discrepancies and errors in official registries, which is a vivid example of practical application. Other examples of using the result are the possibility of scaling the technology to additional data types, as well as supporting digital transformation in the medical field. Such results are prerequisites for automating the clinical trial process and accelerating the release of new drugs to the market

Keywords: clinical trials, large language models, clinical documentation, data mining

Received 01.08.2025
Received in revised form 02.10.2025
Accepted 13.10.2025
Published 30.12.2025

How to Cite: *Horlatch, V., Pasichnyk, V. (2025). Automated extraction of key parameters and detection of inconsistencies in clinical documentation using large language models. Eastern-European Journal of Enterprise Technologies, 6 (2 (138)), 6–14. https://doi.org/10.15587/1729-4061.2025.337915*

1. Introduction

Typically, the clinical trial process takes six to seven years, and sometimes more than ten. A significant part of this process is documentation. In particular, clinical trial protocols are complex documents (often more than 100–200 pages) that outline the trial design, study drugs, objectives, patient inclusion criteria, possible side effects, and other aspects. Reviewing protocols to obtain structured data is a time-consuming and error-prone process, even for experts in the field.

Artificial intelligence (AI), including machine learning (ML), is helping to significantly improve the efficiency of clinical trials, helping to reduce development time and cost [1, 2]. Natural language processing (NLP) tools can process thousands of previous protocols or trial results in seconds. This speeds up the design phase and helps avoid errors. AI can also quickly scan protocol requirements and compare them with the capabilities of clinical centers to select the best sites for trials and to assist with patient recruitment [3].

In recent years, neural network models – especially language-based transformer models – have significantly improved the accuracy of extraction. Instead of manually developed features, these models learn from large text structures. For example, Novartis researchers proposed CT-BERT, a BERT-based model tuned to recognize entities in clinical trial text [4]. This model was trained to extract different entities from inclusion and exclusion criteria and related sections. In general, BERT-style models (and their biomedical variants such as BioBERT, ClinicalBERT, or PubMedBERT) have become the standard for protocol text analysis. These models can handle a variety of phrases and contextual nuances better than previous methods, especially for complex criteria [5]. On the other hand, studies conducted in 2023–2024 demonstrate that in certain cases general generative models could perform as well as their specialized counterparts [6].

The latest trend is to use very large models (GPT-3, GPT-4, etc.), which can perform complex information extraction with minimal training using prompts or rapid learning [7]. These models, pre-trained on a huge amount of text data, can retrieve instructions, parse protocol text into structured JSON, or populate a table with defined fields. Similarly, other studies have used GPT-4 to extract results from cancer trial reports. It has been illustrated that large language models (LLMs) can perform well in a scalable manner [8–10]. However, the biggest caveat is that LLMs sometimes have a tendency to generate spurious fragments or omit information, so careful evaluation and analysis of the results is a pressing task.

2. Literature review and problem statement

The first systems related to the extraction of structured data from clinical texts focused on the manual construction of rules and dictionaries for recognizing medical entities and facts in the text. For example, EliIE [11] became one of the first open tools for parsing inclusion and exclusion criteria – descriptions of patient characteristics that determine the possibility of participating in the study. Although such approaches demonstrated the ability to process simple sentences, their scalability was limited because new formulations of criteria required the addition of rules, and generalization to different diseases was difficult. An option to overcome this is machine learning (ML).

In [12], a combination of several ML models (BERT, RoBERTa, XLNet, ELECTRA, ERNIE) was used to automatically categorize inclusion criteria sentences by type (category). This model achieved an accuracy of ~0.85 and an F1-measure of 0.8169 on a large Chinese-English set of criteria. But this model solves only the classification problem and does not provide a complete extraction of the content of the criteria. The reason is the narrow specialization of the models.

A separate area of research is the formalization of criteria in the form of structures or queries. In [13], the Chia system is described, where each criterion is represented as a dependence graph that can be converted into a Boolean query to the database. Thus, the problem of information extraction is considered as the construction of a formal representation suitable for algorithmic verification. This method has become the basis for the emergence of new approaches – both fully automatic and hybrid (a combination of rules and machine learning), but the issues of completeness and universality remain unresolved. The reason is the narrow specialization of the system.

LLMs may be an option to overcome these difficulties. In the 2020s, researchers began to experiment with applying these universal models to medical domain tasks. Despite the general nature of training, LLMs were able to answer clinical questions, solve diagnostic problems, and analyze medical texts, even if the model was not specifically trained on them [14, 15]. In [16], it is noted that large language models can significantly speed up the routine stages of preparing systematic reviews, while increasing reproducibility and reducing the number of human errors. However, the issues of quantifying the effectiveness of such models for specific tasks and languages (in particular, Ukrainian) and their reliability remain unresolved. The main reason is the lack of targeted research. The paper does not provide data that would quantitatively assess the effectiveness of automated information extraction from clinical texts or investigate the reliability of the results obtained. The lack of such focused studies makes it difficult to assess the effectiveness of large language models (LLMs) and their level of trust.

One way to eliminate the need for manual formalization has been to use generative LLMs directly. For example, AutoCriteria [17] is a relatively new system that uses GPT-based query modeling to obtain detailed elements of inclusion and exclusion criteria. When evaluating 180 trial protocols for 9 diseases, an F1 score of ~ 89% was achieved for identifying criteria elements and ~ 79% accuracy for extracting all context attributes. This demonstrates the possibility of using generative LLMs for similar purposes even without additional training. However, some attributes still remained unallocated, indicating the model's limitations in understanding context.

LLMs are especially actively used in the task of finding patients for trials, which is closely related to obtaining information from criteria [18]. Work [19] demonstrated that GPT-4 can analyze the text of the inclusion criteria in detail and compare it with the text description of the patient's medical diagnosis, additionally providing explanations at the level of individual criteria. This is significantly different from conventional systems, where they initially tried to transform the criteria into formal rules or queries. LLMs practically remove this limitation by transferring the work of interpreting the text to a universal model. However, it is not entirely clear how to ensure stable accuracy and trust in such results.

Based on the analysis of the literature, one can conclude that the problem of quantitatively assessing the effectiveness of using universal LLMs for processing clinical documentation, in particular in the Ukrainian language, remains unsolved. All this gives grounds to argue that it is advisable to conduct a study that would allow a reasonable assessment of the capabilities of LLMs for extracting key indicators from clinical texts.

3. The aim and objectives of the study

The aim of our research is to devise an approach for automated extraction of key structured data from text records of clinical trials based on LLM. This will make it possible to analyze and check data from official registries for errors, as well as to form prerequisites for building "smart assistants" in the field of clinical research.

To achieve this goal, the following tasks were set:

– to build datasets for experiments based on records from the registries at the National Institutes of Health (NIH), USA (ClinicalTrials.gov), and data from the State Expert Center (SEC) at the Ministry of Health (MOH) of Ukraine (clinicaltrials.dec.gov.ua);

– to apply LLM to automatically extract such elements as the eligible gender of participants and the phase of the clinical trial and to quantify the results using classification metrics (accuracy, recall, precision, and F1-measure);

– to apply LLM to automatically extract the research therapeutic area and generate a confusion matrix for visualizing the results;

– to check the input data for inaccuracies and errors.

4. The study materials and methods

4. 1. The object and hypothesis of the study

The object of our study is unstructured text data on clinical trials in Ukraine and the world, in particular obtained from the registries of the US NIH and the SEC of the Ministry of Health of Ukraine. Before the start of the study, the following hypotheses were formulated:

– hypothesis 1. LLMs, even without specialized fine-tuning and additional training, are capable of automatically extracting clearly defined structured fields from unstructured texts of clinical protocols with high accuracy;

– hypothesis 2. The accuracy of data extraction depends on the context in which the relevant data is mentioned, which may affect the quality of the results;

– hypothesis 3. Large language models are able to correctly determine the underlying disease or condition from the name of a clinical trial, even in the absence of its explicit mention, using general knowledge about the structure of clinical documentation and the medical context.

The following assumptions were also adopted:

– structured fields in official registries are reference and contain correct information about the phase, gender of participants, criteria, etc.;

– the text part (titles, inclusion and exclusion criteria, etc.) contains all the necessary data that can be logically linked to structured fields;

– data from different registries have common unique identifiers, which makes it possible to compare the extraction results.

To simplify assessing the quality of the results, experiments were conducted on those fields that have structured counterparts in the official registries, namely: eligible gender of participants, study phase, and study therapeutic area.

4. 2. Data mining models and methods

This study used the principle of "zero-shot prompting", which is justified by the need to demonstrate the capabilities of publicly available LLMs for clinical professionals without complex settings. GPT-4o-mini was chosen as the main model since it is one of the most powerful available LLMs developed by OpenAI. GPT-3.5 Turbo, one of the previous models, was also used for comparison. GPT works under a "black box" mode via API but makes it possible to customize the prompts.

The experiments were performed in several stages. The first stage was research related to the extraction of information about the eligible gender of patients based on the inclusion criteria in the study in the Ukrainian language. Within this stage, several experiments were conducted using different models and queries to find out which one demonstrates the best and most stable results. The experiments were conducted on a sample of 1700 different clinical studies. Details of the experimental parameters are described in Table 1.

One can see that experiments 1, 2, as well as 3, 4, are identical. This was done deliberately in order to compare the "robustness" of the model and the predictability of the results at repeated use. Next, similar experiments were conducted related to the definition of eligible gender based on the inclusion and exclusion criteria in the study, but in English, as well as the definition of the phase and therapeutic area of the study. A detailed description of the parameters is given in Table 2.

Table 1

Detailed description of experimental parameters related to obtaining the eligible gender of patients based on study inclusion criteria

Experiment ID	Model	Input language	Prompt	Number of experiments
G3.5UA1.1	GPT-3.5. Turbo	Ukrainian	Get eligible sex (gender) from the following inclusion criteria of a clinical trial. The answer should be strictly one of the following: ‘жіноча’, ‘чоловіча’, ‘чоловіча, жіноча’, ‘не вказано’. Don’t provide any other comments. Here is the criteria:	1700
G3.5UA1.2	GPT-3.5. Turbo	Ukrainian	Get eligible sex (gender) from the following inclusion criteria of a clinical trial. The answer should be strictly one of the following: ‘жіноча’, ‘чоловіча’, ‘чоловіча, жіноча’, ‘не вказано’. Don’t provide any other comments. Here is the criteria:	1700
G4oUA1.1	GPT-4o-mini	Ukrainian	Get eligible sex (gender) from the following inclusion criteria of a clinical trial. The answer should be strictly one of the following: ‘жіноча’, ‘чоловіча’, ‘чоловіча, жіноча’, ‘не вказано’. Don’t provide any other comments. Here is the criteria:	1700
G4oUA1.2	GPT-4o-mini	Ukrainian	Get eligible sex (gender) from the following inclusion criteria of a clinical trial. The answer should be strictly one of the following: ‘жіноча’, ‘чоловіча’, ‘чоловіча, жіноча’, ‘не вказано’. Don’t provide any other comments. Here is the criteria:	1700
G4oUA2	GPT-4o-mini	Ukrainian	Get eligible sex (gender) from the following inclusion criteria of a clinical trial. The answer should be strictly one of the following: ‘жіноча’, ‘чоловіча’, ‘чоловіча, жіноча’, ‘не вказано’. If the inclusion criteria doesn’t contain clear gender criteria, return ‘не вказано’. Don’t guess. Don’t provide any other comments. Here is the criteria:	1700

Table 2

Detailed description of experimental parameters related to extracting the eligible gender of patients based on inclusion and exclusion criteria in English, as well as the phase and therapeutic area of the study (in Ukrainian)

Experiment ID	Model	Input language	Prompt	Number of experiments
G4oEN2	GPT-4o-mini	English	Your task is to get eligible sex (gender) from the eligibility criteria of a clinical trial that will be provided to you in a prompt. The answer should be strictly one of the following: ‘MALE’, ‘FEMALE’, ‘ALL’, ‘Not mentioned’. If the eligibility criteria doesn’t contain clear gender criteria, return ‘Not mentioned’. Don’t guess. Don’t provide any other comments.	1017
Ph4oUA1	GPT-4o-mini	Ukrainian	You will get full name of a clinical trial. Extract the phase of the study. Your answer should be a single number (e.g., 1, 2, 3, or 4). If there is no explicit mention of the phase or it is ambiguous, answer ‘N/A’.	927
Pr4oUA1	GPT-4o-mini	Ukrainian	You will get full name of a clinical trial. Your goal is to understand the profile of the study. Your answer should be one of the following: ‘Онкоурологія’, ‘Офтальмологія’, ‘Гінекологія’, ‘Дерматовенерологія’, ‘Онкогематологія’, ‘Урологія’, ‘Нефрологія’, ‘Хірургія’, ‘Дерматологія’, ‘COVID-19/Інфекційні хвороби’, ‘Інфекційні хвороби’, ‘Гематологія’, ‘Кардіологія’, ‘Гастроентерологія’, ‘Ендокринологія’, ‘Пульмонологія’, ‘Неврологія’, ‘Ревматологія’, ‘Психіатрія’, ‘Онкологія’. If it is not clear or ambiguous, answer ‘N/A	1259

Importantly, no examples ("few-shot") were added to the queries – the model performed the task under a "zero-shot" mode, i.e., relying only on its general "understanding" of the task. The "temperature" of the model was set to 0 to avoid randomness and increase focus and predictability.

4. 3. Post-processing of results

After receiving the responses from the LLM, the results require further post-processing to ensure their compatibility with the reference structured data, correct interpretation and unified presentation. The main tasks of this stage are normalization of formats, reduction of synonyms, elimination of redundant information, as well as preparation of data for calculation of accuracy metrics.

In particular, the values of fields with a limited set (for example, gender) were reduced to a standardized form. For example, despite the fact that the model received clear instructions on the format of representation of the results ("female", "male", "male, female"), the results could have minor differences ("female", "women", "female gender"). Therefore, it was important to bring them to a single form before evaluating the results. For the phase of the study (e.g., "Phase II", "2", "II/III"), rules were devised to convert to uniform values (1, 2, 3, 4, N/A, 1/2, etc.).

These data were used to further calculate metrics and construct confusion matrices.

4. 4. Evaluation methods

The results were evaluated using standard data mining and classification metrics. Precision, Recall, Accuracy, and F1-measure were calculated for each class. These metrics are best suited for evaluating text mining tasks, as they provide a comprehensive characteristic of the model quality. They allow us to determine how well the system is able to correctly recognize the desired elements and avoid false positive results. Due to this, such metrics have become a standard in research related to classification and information mining. Let us consider each of them in more detail.

Precision determines the proportion of correctly predicted positive cases among all predicted positive results and is calculated from formula (1)

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

where:

– TP is the number of true positive predictions (for example, in experiments related to the extraction of eligible gender, the model returned the result ‘female' and the reference value was also ‘female');

– FP is the number of false positive predictions (for example, the model determined that the eligible gender of the study is ‘female', although the reference value is ‘male' or ‘male, female').

Using formula (2), recall is calculated – a metric that shows the proportion of correctly found positive cases among all expected positive cases

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

where FN is the number of false negative cases (for example, for studies where the reference value is "female", the model returned any other value, but not "female").

Formula (3) demonstrates the calculation of the F1-measure – the harmonic mean between precision and recall

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

This metric is particularly useful because it considers both precision and recall at the same time.

Accuracy is the total proportion of correct predictions among all cases; it is given by formula (4)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

where TN is the number of true negative cases (for example, for studies where the reference value is different from "female", the model also returned a result different from "female").

In addition, a confusion matrix was constructed for experiments related to determining the study therapeutic area. This allowed us to summarize typical errors and clearly visualize the frequency of their occurrence.

5. Results of extracting key parameters from clinical texts

5. 1. Data sets formed based on records from open registries

For the experiments, a data set was formed based on records of clinical trials from two sources:

1. Database at the State Expert Center of the Ministry of Health of Ukraine (clinicaltrials.dec.gov.ua). It contains general information about almost all clinical trials that have taken place, are taking place, or are planned in Ukraine.

2. Database at the US National Institutes of Health (ClinicalTrials.gov). It contains detailed information about almost all clinical trials from around the world.

Both registries contain a similar structure of records: short title of the study, description, inclusion and exclusion criteria, trial phase, requirements for gender and age of participants, status indicator, etc. At the same time, the mechanisms for accessing data, as well as their presentation, differ significantly.

The database at the State Expert Center (SEC) of the Ministry of Health of Ukraine does not have an open application programming interface for convenient interaction. Therefore, it was necessary to use Selenium WebDriver methods to read information about each study separately. After that, for each study, it was necessary to additionally obtain data from the US National Institutes of Health database. This database has a convenient API that simplifies work with it. However, an obstacle arose since clinical studies in the database at the State Expert Center of the Ministry of Health of Ukraine and in the database at the US National Institutes of Health had different identifiers.

To obtain data from the American database, it was necessary to use the so-called NCT ID, which was absent in the database at the State Expert Center of the Ministry of Health of Ukraine. This task was solved by executing an HTTP query in the following format "https://clinicaltrials.gov/api/int/studies?id={other_id}" for each study. This query returned basic information about the study, including the NCT ID. Further, having this identifier, it was possible to use the API to obtain all the necessary details about the study.

In total, 1,700 records of clinical trials of phases I–IV in various therapeutic areas that are ongoing, completed, or planned in Ukraine were selected for the experiments. For each record, the following data were obtained from both sources:

– structured data – that is, the values of the fields officially provided in the registry: gender of participants (male, female, or both), phase (I, II, III, IV), study therapeutic area, and others. These data were used as the correct reference values for evaluating the performance of the models;

– unstructured text data, namely the full title of the study and the inclusion and exclusion criteria section. It is these fields that contain the unstructured description from which structured facts for comparison should be extracted.

The next step was to clean and transform the data to ensure that all information was represented in a unified form. This included removing unnecessary spaces, eliminating other markup elements, and general standardization of coding and language. In particular, research phases are usually denoted by Roman numerals, but for comparison it is more convenient to use Arabic numerals. Some fields required translation from Ukrainian to English, or vice versa, to ensure consistency. Visually, the data processing pipeline is depicted in Fig. 1.

All collected and cleaned data sets are available and published at [20].

5. 2. Results of extracting information on the eligible gender of participants and the phase of the clinical study

Detailed research results are available at [21]. Tables 3–6 give the summarized results related to the extraction of information on the eligible gender of patients based on the inclusion criteria in the study. In particular, the values of precision, recall, accuracy, and F1-measure for each of the target elements are given.

One can seen from the results given in Tables 3–5 that both models (GPT-3.5 Turbo and GPT-4o-mini) demonstrated quite high performance in experiments related to the extraction of eligible gender. However, the results of GPT-4o-mini turned out to be better and more stable. Given that this model is also more modern and cheaper, it was used for further experiments. Table 6 gives the summarized results of experiments related to the extraction of information about the eligible gender of patients based on the inclusion and exclusion criteria in the study, this time in English.

Table 7 gives the summarized results of experiments related to extracting information about the study phase based on the full name of the clinical trial in Ukrainian.

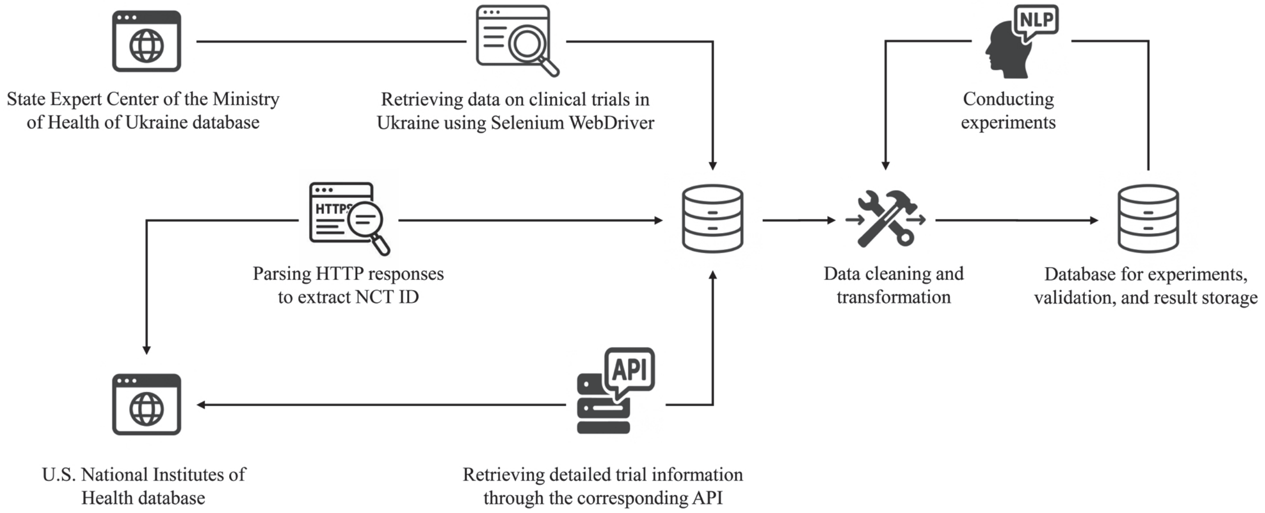
State Expert Center of the Ministry of Health of Ukraine database
Retrieving data on clinical trials in Ukraine using Selenium WebDriver
NLP
Conducting experiments
HTTPS
Parsing HTTP responses to extract NCT ID
Data cleaning and transformation
Database for experiments, validation, and result storage
API
U.S. National Institutes of Health database
Retrieving detailed trial information through the corresponding API

Fig. 1. Data processing pipeline

Table 3

Results of experiments related to determining the eligible gender based on inclusion criteria (in Ukrainian), where the reference gender is "male"

Experiment	AP	AN	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
G3.5UA1.1	49	1651	45	1644	7	4	0.99	0.87	0.92	0.89
G3.5UA1.2	49	1651	45	1646	5	4	0.99	0.90	0.92	0.91
G4oUA1.1	49	1651	49	1647	4	0	1.00	0.92	1.00	0.96
G4oUA1.2	49	1651	49	1647	4	0	1.00	0.92	1.00	0.96
G4oUA2	49	1651	49	1648	3	0	1.00	0.94	1.00	0.97

Table 4

Results of experiments related to determining the eligible gender based on inclusion criteria (in Ukrainian), where the reference gender is "female"

Experiment	AP	AN	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
G3.5UA1.1	94	1606	79	1516	90	15	0.94	0.47	0.84	0.60
G3.5UA1.2	94	1606	81	1510	96	13	0.94	0.46	0.86	0.60
G4oUA1.1	94	1606	94	1572	34	0	0.98	0.73	1.00	0.85
G4oUA1.2	94	1606	94	1572	34	0	0.98	0.73	1.00	0.85
G4oUA2	94	1606	93	1534	72	1	0.96	0.56	0.99	0.72

Table 5

Results of experiments related to determining eligible gender based on inclusion criteria in the Ukrainian language, where the reference gender is "male, female"

Experiment	AP	AN	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
G3.5UA1.1	1326	374	1250	170	204	76	0.84	0.86	0.94	0.90
G3.5UA1.2	1326	374	1248	175	199	78	0.84	0.86	0.94	0.90
G4oUA1.1	1326	374	1323	350	24	3	0.98	0.98	1.00	0.99
G4oUA1.2	1326	374	1321	349	25	5	0.98	0.98	1.00	0.99
G4oUA2	1326	374	1252	372	2	74	0.96	1.00	0.94	0.97

Table 6

Results of experiments related to determining eligible gender based on inclusion and exclusion criteria in English

Gender	AP	AN	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
Female	97	920	97	743	177	0	0.83	0.35	1.00	0.52
Male	61	956	60	950	6	1	0.99	0.91	0.98	0.94
Female, male	859	158	676	157	1	183	0.82	1.00	0.79	0.88

Table 7

Results of experiments related to phase determination based on the full title of the study in Ukrainian

Study phase	AP	AN	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
Phase I	23	904	23	893	11	0	0.99	0.68	1.00	0.81
Phase II	271	656	263	630	26	8	0.96	0.91	0.97	0.94
Phase III	624	303	593	303	0	31	0.97	1.00	0.95	0.97
Phase IV	9	918	7	917	1	2	1.00	0.88	0.78	0.82

It is worth noting that in most experiments the lowest indicators are observed for the metric reflecting precision.

5. 3. Results of extracting the research therapeutic area

Table 8 visualizes the results of experiments on determining the research therapeutic area based on the full name. The data are represented in the form of a confusion matrix.

In Table 8, the rows (marked with the letter "a") reflect the actual class, and the columns (marked with the letter "p") reflect the predicted class, i.e. the result of the model; the following diseases are marked with Latin letters: A – COVID-19, B – Gastroenterology, C – Hematology, D – Gynecology, E – Dermatovenereology, F – Dermatology, G – Endocrinology, H – Infectious diseases, I – Cardiology, J – Neurology, K – Nephrology, L – Oncohematology, M – Oncology, N – Oncourology, O – Ophthalmology, P – Psychiatry, Q – Pulmonology, R – Rheumatology, S – Urology, T – Surgery.

Additionally, the results can be displayed in the form of a diagram (Fig. 2).

Table 8

Confusion matrix, built on the basis of the results of experiments related to determining the therapeutic area based on the full title of the study in Ukrainian

p a	A	B	C	D	E	F	G	H	I	G	K	L	M	N	O	P	Q	R	S	T
A	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	91	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
C	0	0	5	0	0	0	0	0	1	0	1	63	2	0	0	0	0	0	0	0
D	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	2	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	1	0	0	81	0	3	0	8	0	0	0	0	0	0	0	0	0
H	19	6	0	0	1	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	2	0	58	2	0	0	0	0	0	0	9	0	0	1
J	0	1	0	1	0	0	1	0	0	18	1	0	0	0	2	0	0	0	0	0
K	0	0	0	0	0	0	0	0	1	0	24	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	22	1	0	0	0	0	0	0	0
M	0	5	0	72	0	2	0	0	0	0	0	23	137	13	2	0	29	0	2	1
N	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	1	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	128	0	0	0	0
Q	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	90	0	0	0
R	0	0	6	6	10	4	0	0	0	0	5	0	0	0	0	0	0	101	0	0
S	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	23	0
T	7	0	0	0	1	8	0	0	0	0	0	1	0	0	0	0	0	0	0	7

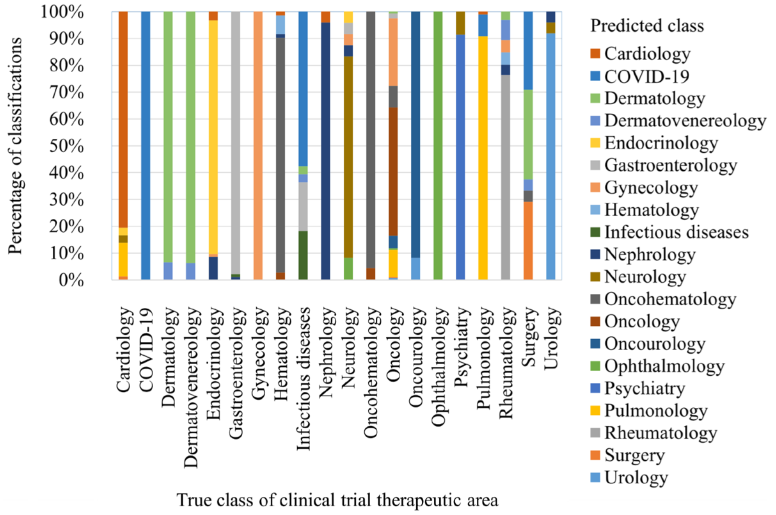

Fig. 2. Distribution of clinical trial therapeutic area classifications

The diagram illustrates how often the model correctly categorized the clinical trial therapeutic area (usually a dominant solid color in each column) and what typical errors occurred (admixtures of other colors). This makes it possible to clearly see both the strengths of the approach and the problem areas where the model tends to make mistakes.

5. 4. Found errors in official registries

As a result of our experiments, it was found that the reference data taken from the website of the SEC of the Ministry of Health of Ukraine c ontain certain inaccuracies and errors, in particular regarding the eligible gender of participants. Detailed examples of these inaccuracies can be found in Table 9.

Even non-specialists in the medical field can notice the obvious discrepancies.

Table 9

Examples of errors found in the SEC data of the Ministry of Health of Ukraine

CT code	Gender indicated on the SEC website	Inclusion criteria confirming the existence of an error
GS-US-259-0110	male	... Men and women aged 18 to 75... ... For women of childbearing age...
CNTO328MMY2001	female	... men or women over 18 years of age for the screening period...
SB3-G31-BC-E	male, female	... Patients who completed treatment in study SB3-G31-BC according to the protocol ...
LFB-FVIIa-007-14	male, female	... Male gender, diagnosis of congenital hemophilia A or B of any degree of expression....
P05691	female	... Men and women who are not capable of having children are allowed to participate in the study ...
NN7415-4159	male, female	Men diagnosed with hemophilia A without inhibitors...
ODO-TE-B301	male, female	Female patients at least 18 years of age Histologically or cytologically confirmed breast cancer ...
STM01-102	male, female	1. Women over 18 years of age with breast adenocarcinoma based on histological or cytological findings...
251001	male, female	In Ukraine, women are not participating in this study ... ... Hemophilia B
P06384	male, female	... The patient can be a man or a woman who is incapable of bearing children ...

6. Discussion of the results of extracting key parameters from clinical texts, as well as the errors found

One of the prerequisites of our study was the construction of a combined dataset [20] on clinical trials in Ukraine based on the registries at the US NIH and the SEC of the Ministry of Health of Ukraine. This is already a significant result because, unlike the existing ones, this dataset significantly expands the available information on trials, provides the opportunity to compare records, train models, and conduct other studies. This became possible owing to the designed data processing pipeline shown in Fig. 1. Another important achievement is the extraction of key parameters from clinical texts and subsequent quantitative assessment of quality. Unlike [11], this became possible owing to the use of LLM and the application of classification methods.

For experiments where the eligible gender is "male" and "male, female" (Tables 3, 5), we observe high rates of accuracy (from 0.82 to 0.99), recall (from 0.92 to 1), and F1-measure (from 0.88 to 0.99). In contrast, for studies with female gender as a criterion, the results were noticeably lower. For example, in Tables 4 and 6, we observe a large proportion of false-positive results and a low precision value (from 35% to 73%) specifically for female gender. This is due to the fact that the inclusion criteria often include additional restrictions for pregnant women (for example, "For women of childbearing age, a negative pregnancy test result must be obtained during screening"). The model often interpreted such texts not as an additional restriction for a certain group of women, but as the main criterion ("For women").

If we compare the results obtained on Ukrainian-language texts with their English counterparts (Table 6), the indicators on Ukrainian texts turned out to be even slightly higher. This may seem unexpected because LLMs usually work better with English-language texts. The explanation can be found in the differences in the structure and representation of inclusion and exclusion criteria on the websites of the SEC of the Ministry of Health of Ukraine and the US National Institutes of Health. In particular, in Ukrainian sources the model focused specifically on inclusion criteria, while American counterparts combine inclusion and exclusion criteria into one set of participation requirements. This combination could complicate the interpretation of the model since it perceived exclusion as part of inclusion.

If we analyze the results related to the extraction of the study phase (Table 7), here too we observe quite high rates. Only for the first phase of studies, the Precision value is quite low (68%). The main reason is that the model was explicitly instructed to return only one of the four options (1, 2, 3, or 4), while for some studies combined phases were specified (1/2, 1b/2, Ib/II, etc.). In such cases, the model usually returned the first phase, although the reference value could be "phase II".

As for the classification of the study therapeutic area (Table 8), we observe high rates in most cases. However, for individual therapeutic areas, the results turned out to be extremely low. For example, out of 16 studies that were labeled as "Dermatovenereology" in our dataset (row E), the model categorized 15 as "Dermatology" (column F). This may indicate errors in the "reference" data rather than poor model performance since these studies were related to diseases such as atopic dermatitis, acne vulgaris, and psoriasis. However, we leave the interpretation of these results to medical professionals.

As a result of our experiments, a number of inconsistencies were found in the official registries (Table 9). For example, studies "CNTO328MMY2001" and "P05691" are indicated as being for women only, although the inclusion criteria clearly state that both men and women are allowed to participate in these studies. Studies "LFB-FVIIa-007-14" and "251001" are marked as being available for both men and women. Although the criteria clearly state that these studies are exclusively for men since these are trials related to hemophilia, a disease that occurs only in men. Most likely, the main cause of such errors is the human factor.

In general, the results provide a reasonable assessment of the use of AI for extracting key parameters from clinical documentation, in particular in Ukrainian. In contrast to [17], the extraction of such fields as eligible gender, phase and study therapeutic area was demonstrated, in particular in Ukrainian-language texts. A quantitative assessment was conducted and high indicators were obtained. The model errors were of a template nature – for example, erroneous interpretation of combined phases (I/II) or incorrect interpretation of additional restrictions on gender. The results of the experiments allow us to state that artificial intelligence methods are able to effectively assist in working with clinical documentation, in particular in Ukrainian. This became possible due to the rapid evolution of modern AI.

It is expected that the application of the proposed approach in practice could significantly reduce the time required for manual verification of clinical documentation, increase the reliability of information in registers, and reduce the number of errors associated with the human factor. Clinical trials typically last years, sometimes decades. Automated analysis of unstructured text would help speed up the preparation of studies, reconcile data between registries, and increase the transparency of clinical trials. Even if this process can be reduced by a few weeks or months, it could have an impact on human lives.

Several limitations are worth noting. First, the model was evaluated on a limited set of relatively simple attributes. This is because there were "reference" values for them and it was possible to accurately calculate the result. Fields such as results, drug dosage, side effects were not considered – they can also be tried to be extracted from the text of protocols or reports, but this is a topic for a separate study. Second, the "zero-shot" approach relied on the quality of the instructions, and it is possible that there are better options for prompts. The "few-shot" option was also not investigated: perhaps by providing the model with 2–3 parsing examples before the task, some of the errors could be eliminated.

Among the disadvantages of the approach, one can single out the instability of the output data because even with the same input requests, LLMs are able to return different answers. This creates uncertainty in the functioning of the system and can become critical for tasks where full determinism of the result is required. Ethical aspects should also be taken into account. Models like GPT-4 can sometimes look convincing but produce incorrect results (so-called "hallucinations"). Even one error in a critical criterion can have consequences. Therefore, today it is most appropriate to use LLMs as effective assistants but not as a replacement for specialists.

Future studies may tackle the following tasks:

– scaling the process of extracting key information, by expanding the number of fields and involving a larger sample of clinical trial records, as this would help expand the scope of application;

– using additional methods for assessing the accuracy of the results, for example, by artificially and controlled introduction of errors into the input data, as this could help evaluate those experiments where there are no reference values for comparison;

– modifying prompts, adding examples, and retraining models to improve the accuracy of results;

– integrating LLM with the knowledge graph to clarify ambiguous terms or verify extracted facts.

7. Conclusions

1. A unique database has been built that contains detailed information on existing, completed, and planned clinical trials in Ukraine, as well as similar data on studies in the world. These data can be used for practical purposes – for example, for the selection of patients for clinical trials. They also form the basis for further analysis and scientific research, in particular for additional training of neural networks.

2. Experiments have been conducted that demonstrate the high efficiency of modern artificial intelligence methods in data mining and analysis of documents related to clinical trials. Large language models show high performance (average values: F1-measure – 0.92; accuracy – 0.98; recall – 0.99; precision – 0.87) when extracting clearly defined fields from clinical text. In cases where the metrics are lower (the lowest recorded values: F1-measure – 0.52; accuracy – 0.82; recall – 0.78; precision – 0.35), errors are predictable.

3. It has been confirmed that LLMs are able to process medical texts in Ukrainian, in particular, to determine the study therapeutic area based on the general title. This is important for application under local conditions where most of the clinical documentation is kept in Ukrainian. Such capability of the models opens up opportunities for the scaled use of artificial intelligence at Ukrainian medical and scientific institutions.

4. As a result of the experiments, inaccuracies were found in the official data. In particular, in some cases, the eligible gender of the study participants differed from the information specified in the inclusion criteria. Such errors may have significant consequences not only for the clinical trial process but also for further training of language models.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The manuscript has associated data in the data warehouse [20, 21].

Use of artificial intelligence

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

References

1. Cascini, F., Beccia, F., Causio, F. A., Melnyk, A., Zaino, A., Ricciardi, W. (2022). Scoping review of the current landscape of AI-based applications in clinical trials. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.949377
2. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F. et al. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics, 73, 14–29. https://doi.org/10.1016/j.jbi.2017.07.012
3. Alexander, M., Solomon, B., Ball, D. L., Sheerin, M., Dankwa-Mullan, I., Preininger, A. M. et al. (2020). Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA Open, 3 (2), 209–215. https://doi.org/10.1093/jamiaopen/ooaa002
4. Liu, X., Hersch, G. L., Khalil, I., Devarakonda, M. (2021). Clinical Trial Information Extraction with BERT. 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). IEEE, 505–506. https://doi.org/10.1109/ichi52183.2021.00092
5. Li, J., Wei, Q., Ghiasvand, O., Chen, M., Lobanov, V., Weng, C., Xu, H. (2022). A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Medical Informatics and Decision Making, 22 (S3). https://doi.org/10.1186/s12911-022-01967-7
6. Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N. et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv preprint. arXiv:2311.16452. https://doi.org/10.48550/arXiv.2311.16452
7. Lee, J.-M. (2024). Strategies for integrating ChatGPT and generative AI into clinical studies. Blood Research, 59 (1). https://doi.org/10.1007/s44313-024-00045-3
8. Lee, K., Paek, H., Huang, L.-C., Hilton, C. B., Datta, S., Higashi, J. et al. (2024). SEETrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. Informatics in Medicine Unlocked, 50, 101589. https://doi.org/10.1016/j.imu.2024.101589
9. Wong, A., Plasek, J. M., Montecalvo, S. P., Zhou, L. (2018). Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 38 (8), 822–841. https://doi.org/10.1002/phar.2151
10. Dyyak, I., Horlatch, V., Pasichnyk, T., Pasichnyk, V. (2024). Assessing generative pre-trained transformer 4 in clinical trial inclusion criteria matching. Proceedings of the CEUR Workshop, 3702, 305–316. Available at: https://ceur-ws.org/Vol-3702/paper25.pdf
11. Kang, T., Zhang, S., Tang, Y., Hruby, G. W., Rusanov, A., Elhadad, N., Weng, C. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association, 24 (6), 1062–1071. https://doi.org/10.1093/jamia/ocx019
12. Zeng, K., Xu, Y., Lin, G., Liang, L., Hao, T. (2021). Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning. BMC Medical Informatics and Decision Making, 21 (S2). https://doi.org/10.1186/s12911-021-01492-z
13. Kury, F., Butler, A., Yuan, C., Fu, L., Sun, Y., Liu, H. et al. (2020). Chia, a large annotated corpus of clinical trial eligibility criteria. Scientific Data, 7 (1). https://doi.org/10.1038/s41597-020-00620-0
14. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29 (8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8
15. Veen, D. V., Uden, C. V., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C. et al. (2023). Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. https://doi.org/10.21203/rs.3.rs-3483777/v1
16. Luo, X., Chen, F., Zhu, D., Wang, L., Wang, Z., Liu, H. et al. (2024). Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. Journal of Medical Internet Research, 26, e56780. https://doi.org/10.2196/56780
17. Datta, S., Lee, K., Paek, H., Manion, F. J., Ofoegbu, N., Du, J. et al. (2023). AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. Journal of the American Medical Informatics Association, 31 (2), 375–385. https://doi.org/10.1093/jamia/ocad218
18. Kim, J., Quintana, Y. (2022). Review of the Performance Metrics for Natural Language Systems for Clinical Trials Matching. MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation. IOS Press Ebooks, 641–644. https://doi.org/10.3233/shti220156
19. Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracken-Clarke, D. et al. (2024). Matching patients to clinical trials with large language models. Nature Communications, 15 (1). https://doi.org/10.1038/s41467-024-53081-z
20. Pasichnyk, V. (2025). Datasets of clinical trials in Ukraine. Figshare. Collection. https://doi.org/10.6084/m9.figshare.c.7887785.v1
21. Pasichnyk, V. (2025). Results of key data extraction from clinical trial documentation. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.29378450