Assessing Generative Pre-trained Transformer 4 in Clinical Trial Inclusion Criteria Matching

Ivan Dyyak[1], Vitaliy Horlatch[1], Tymofii Pasichnyk[1], and Vasyl Pasichnyk[1]

[1] *Ivan Franko National University of Lviv, 1 Universytetska St., Lviv, 79000, Ukraine*

Abstract

Clinical trials play a very important role in improving the quality and expectancy of life. At the same time, this is a complicated process which require a lot of time, effort, and costs. Effective interpretation and understanding of inclusion and exclusion criteria are key elements for streamlining trial documentation, enhancing patient engagement, and facilitating data analysis. The nuances in clinical criteria can significantly affect trial outcomes and participant selection across different linguistic and cultural backgrounds. The aim of this study is to assess the capability of the Generative Pre-trained Transformer 4 (GPT-4) model in comprehending and matching clinical trial inclusion criteria between English and Ukrainian, focusing on the model's potential to automate document processing and improve patient engagement in a multilingual context. The model was tested on 315 inclusion criteria from 15 clinical trials from ClinicalTrials.gov and their corresponding Ukrainian versions from ClinicalTrials.dec.gov.ua. The GPT-4 model accurately identified and matched 91% of the criteria pairs fully and demonstrated a detailed understanding of clinical language and nuances. An additional 4% of pairs were identified partially. All together it resulted in 95% accuracy for the tasks which accept partial matches. These findings indicate the model's substantial capacity for understanding complex medical and clinical terminologies and for facilitating the automation of clinical documents. Moreover, the experiment has already helped to identify several cases where the English and the Ukrainian versions had significant meaningful differences, most likely caused by human mistakes. The results of the study open a wide range of opportunities for automating clinical documentation processes and enhancing patient engagement through modern technologies. The scientific novelty is in applying Natural Language Processing (NLP) to understand and process clinical trial criteria in different languages, in this case Ukrainian, leading to more automated and patient-centered approaches in clinical research.

Keywords

Natural Language Processing, Generative Pre-trained Transformer, GPT-4, Clinical Trial Matching, Inclusion Criteria

1. Introduction

The increasing globalization of clinical trials presents significant challenges and opportunities in clinical research. One of the cases is related to the accessibility of trial participation in different regions and landscapes. Typically, clinical trials are conducted in many different countries. According to the regulations, all the related documentation should be translated into the corresponding languages. Also, in most cases, local Electronic Health Records (EHR) keep data in local languages too.

The use of automated tools and natural language processing (NLP) technologies is becoming increasingly common in addressing the complexities of multilingual documentation in clinical trials. Despite advancements, there remains a significant gap in the understanding and application of automated tools related to clinical trial criteria matching, especially across different languages. This gap not only prevents the efficiency of global trial operations but also impacts patient engagement and onboarding, which are crucial for the ethical and effective conduct of clinical research.

The object of this study is the process of interpreting and understanding clinical trial inclusion criteria in a multilingual context using computer sciences. The subject covers the methods and

CMIS-2024: Seventh International Workshop on Computer Modeling and Intelligent Systems, May 3, 2024, Zaporizhzhia, Ukraine

ivan.dyyak@lnu.edu.ua (I. I. Dyyak); vitaliy.horlatch@lnu.edu.ua (V. M. Horlatch); tymofiy.pasichnyk@lnu.edu.ua (T. V. Pasichnyk); vasyl.pasichnyk.apmi@lnu.edu.ua (V. T. Pasichnyk)

0000-0001-5841-2604 (I. I. Dyyak); 0000-0001-5401-1731 (V. M. Horlatch); 0000-0002-1495-7423 (T. V. Pasichnyk); 0009-0004-3738-6629 (V. T. Pasichnyk)

technologies which are used to understand and match these criteria across languages, with a particular focus on the English and Ukrainian language pair, assuming that it can be generalized and extended to other languages too.

The goal of this work is to analyze and validate the use of artificial intelligence and the Generative Pre-trained Transformer 4 (GPT-4) in particular to enhance the understanding and application of clinical trial inclusion criteria in different linguistic contexts. This involves identifying, comparing, and assessing the equivalence of clinical trial criteria between English and Ukrainian, aiming to improve document automation processes, facilitate more effective patient engagement, and streamline data analysis procedures.

The importance of this research originated from the ongoing need to improve inclusivity and accessibility in global clinical trials, ensuring that non-English speaking populations are adequately represented and can participate fully and safely in clinical research. By leveraging advanced NLP tools like GPT-4, this study addresses a critical gap in current research methodologies, offering a potential for significant advancements in clinical trial design, execution, and analysis. The appropriateness of this article within the relevant field is underscored by its contribution to improving clinical trial efficiency and participant comprehension, ultimately promoting more fair and effective global health research.

2. Problem Statement

The core problem addressed in this study involves the accurate understanding and matching of clinical trial inclusion criteria across different languages, specifically English and Ukrainian. This problem can be formalized mathematically as follows.

Input Variables:

- A set of inclusion criteria in English extracted from ClinicalTrials.gov, which can be represented by the formula

$$I_E = \{e_1, \quad e_2, \quad \dots \quad , e_n\} \tag{1}$$

  where each e_i represents a specific criterion.
- A corresponding set of inclusion criteria in Ukrainian obtained from ClinicalTrials.dec.gov.ua, which can be represented by the formula

$$I_U = \{u_1, \quad u_2, \quad \dots \quad , u_n\} \tag{2}$$

  where each u_i represents a specific criterion in Ukrainian.

Output Variables:

- A set of matched pairs between English and Ukrainian inclusion criteria:

$$M = \{(e_i, u_j) | e_i \in I_E, u_j \in I_U\} \tag{3}$$

  where each pair consists of one English criterion and one Ukrainian criterion.
- A set of scores indicating the degree of match between pairs in M:

$$S = \{s_1, \quad s_2, \quad \dots \quad , s_n\} \tag{4}$$

  where each s_i corresponds to the semantic similarity between e_i and u_i.

Results Quality Evaluation Criteria:

- Accuracy (Acc) – The proportion of correctly matched criteria pairs in M to the total number of criteria pairs. It can be calculated by the formula:

$$Acc = \frac{\left|\{(e_i, u_j) \in M |\ s_i\ \ indicates\ a\ correct\ match\}\right|}{n} \quad (5)$$

- Precision (P) – The proportion of true positive matches to the total number of positive matches identified by the model.
- Recall (R) - The proportion of true positive matches to the total number of actual positive matches across the datasets.

Necessary Definitions:

- Positive match identified by the model – a pair of criteria that the model identifies as a match, regardless of whether this match is correct.
- True positive match – a pair of criteria that are correctly identified as equivalent by the model.
- Actual positive match – a pair of criteria that an expert identifies as a match.
- False positive match – a pair of criteria that are identified as equivalent by the model but in fact is not a match.
- False negative match – a pair of criteria that are identified by the model as not a match but in fact is a match.

3. Review of the literature

The application of Artificial Intelligence (AI) and Natural Language Processing (NLP) in the field of clinical trials is a rapidly emerging area [1]. It is often being discussed in modern scientific and business conferences and a considerable number of scientific articles have been published. If properly used, AI can significantly enhance operational efficiencies, patient engagement, and the accuracy of data interpretation. Furthermore, the capacity of AI to innovate and refine the design of clinical trials is recognized as a potential catalyst for increasing the efficacy and efficiency of trial methodologies [2]. The advancements in these technologies underscore their potential to transform the landscape of clinical research by optimizing trial procedures and ensuring more targeted and effective patient engagement strategies.

A comprehensive overview of current capabilities and limitations of AI applications in clinical trials was provided, underscoring the transition from artificial to applied intelligence and highlighting the importance of these technologies in enhancing trial outcomes [3]. The transformative role of machine learning in clinical research focuses on the potential to automate and optimize various aspects of trial execution and evidence generation, aligning with findings that NLP can enhance the standardization and utilization of clinical data [4–10].

One of the key elements of each clinical trial is eligibility criteria. An accurate and comprehensive understanding of inclusion and exclusion criteria is fundamental to ensuring that the proper patients are enrolled. These criteria not only define the scope and structure of a trial but also ensure the safety and appropriateness of participants. Misinterpretations or miscommunications in these criteria can lead to incorrect selection of participants, regulatory issues, and even trial failure or health consequences for the participants [11, 12].

The need for precise matching of patients to trials based on inclusion and exclusion criteria is a complicated process. It requires a lot of manual effort and collaboration between different organizations and electronic systems. At the same time, considering that nowadays most of the information is tracked in electronic records, this process has many opportunities to be automated. The growing potential of NLP systems plays a critical role in streamlining the patient recruitment process [13].

However, a significant gap exists in the application of NLP technologies for understanding and matching clinical trial inclusion criteria across different languages. Research into NLP tasks in clinical settings has not specifically addressed the challenges of multilingual criteria interpretation, indicating a need in clinical research where language barriers can prevent participant enrollment and engagement [14].

Recent studies have explored the capabilities of GPT-4, a large language model, in clinical settings, demonstrating its potential in outperforming traditional methods and medical

professionals in certain tasks. These findings provide insights into its language understanding capabilities and suggest promising opportunities for utilizing GPT-4 in the interpretation and application of clinical trial criteria, potentially overcoming language and comprehension barriers [15–20].

In summary, while significant advancements have been made in AI and NLP applications within clinical research, a critical gap remains in the context of multilingual inclusion criteria interpretation. This study aims to bridge this gap by leveraging the capabilities of GPT-4, thus contributing to the existing body of knowledge and addressing the identified limitations. The research underscores the ongoing evolution of AI in medicine, aiming to enhance the accuracy, efficiency, and inclusivity of clinical trial processes [21].

4. Materials and Methods

The theoretical foundation of this research is grounded in artificial intelligence and machine learning, particularly in the context of natural language processing (NLP) and machine translation (MT). Even though, there are special models tuned for medicine (e.g. PubMedBERT, BioGPT, Med-PaLM, etc.), recent studies have shown that GPT-4 can outperform them [22]. Therefore, we chose the GPT-4 model to analyze and compare inclusion criteria across different languages in our study. The hypothesis posits that the GPT-4 model, trained on diverse linguistic datasets, can accurately match and assess the equivalence of clinical trial inclusion criteria in English and Ukrainian.

The research employs semantic analysis techniques inherent to GPT-4 to interpret and compare the meaning of text segments from the inclusion criteria. These techniques involve the decomposition of text into smaller, manageable units (bullets), enabling a detailed, point-by-point comparison. The theoretical significance lies in assessing the model's ability to understand and translate medical and clinical terminology accurately, a critical factor given the nuanced nature of clinical trial criteria.

The methodological approach encompasses several key components:

- **Data Preparation**: In order to get a proper data set which will be close to the real-world scenarios, the following subsets were prepared and mixed:
    1. Inclusion criteria of the same clinical trial in English and Ukrainian were taken from the corresponding sites without any changes. Even though they had a similar meaning, the text could be several pages long and semantically structured in a completely different way.
    2. The inclusion criteria were extracted and segregated into individual bullets. The number of bullets in the English and Ukrainian versions could differ. As well as their content.
    3. Similar to the above however English and Ukrainian criteria were intentionally taken from different clinical trials

  This step ensured that the model was evaluated on different conditions, enriching the usability and applicability of the study.
- **Model Selection**: We utilized the GPT-4 model for identifying and comparing criteria pairs. This involved configuring the model to process pairs of text bullets and assess their semantic equivalence across languages. The model's performance metrics were predetermined to evaluate its accuracy in identifying exact matches, partially correct matches, and discrepancies.
- **Analytical Techniques**: The study applied quantitative methods to analyze the model's output, focusing on the proportion of fully correct, partially correct, and incorrect matches. Statistical analysis was used to assess the model's accuracy and reliability in translating and matching criteria across languages.
- **Validation Process**: To ensure the validity of the research results, a subset of the GPT-4-identified pairs was manually reviewed by an expert. This step served to cross-verify the accuracy of the automated assessments and to identify potential areas for model improvement.

The research was conducted in a sequential manner, beginning with the collection and preparation of data, followed by the deployment of the GPT-4 model, preparation of

corresponding prompts, data processing, analysis, and validation of the results. The important step was related to the manual assessment and interpretation of the results. This sequence was designed to ensure a logical progression of tasks and to build a coherent dataset for analysis.

The selection of the GPT-4 model as the primary research tool was justified by its advanced language processing capabilities and its proven effectiveness in similar linguistic tasks. Nowadays, it is easily accessible by different companies and organizations. That provides a significant boost for further research and development in this area. The use of statistical analysis for evaluating the model's performance provided a quantitative basis for assessing accuracy, reliability, and validity.

5. Experiments

The experiment was started from the data collection. We created the dataset which contained 315 inclusion criteria extracted from 15 distinct clinical trials. Each criterion had up to 400 words. English criteria were collected from publicly available records on ClinicalTrials.gov [23], a database of privately and publicly funded clinical studies conducted around the world. The corresponding Ukrainian translations were obtained from ClinicalTrials.dec.gov.ua, the official Ukrainian counterpart hosting clinical trial information.

As described in the section above, several types of data collection had a place. It is worth mentioning that even though typically inclusion criteria in the English and Ukrainian versions should be the same, in reality, they might have many additional details related to local regulations, measurement details, etc. For example, the English version of a criterion might look like “Has adequate organ function within 7 days prior to the start of study treatment” while the Ukrainian version will contain a detailed list of laboratory values to be met in order to be evaluated as “adequate organ function”.

For each clinical trial, was created a separate but exactly the same (except the input parameters) prompt to establish proper isolation. The prompt contained the instructions, a numerated set of inclusion criteria in English, and a numerated set of inclusion criteria in Ukrainian. The preprocessed inclusion criteria were inputted into the OpenAI GPT-4 model hosted on an isolated Microsoft Azure instance. The model’s temperature was set to 0 to avoid randomness, and enhance the model's focus and predictability.

The model's task was to identify corresponding pairs of criteria between the English and Ukrainian datasets and then to assess whether the pairs matched completely or exhibited any meaningful differences. The outputs were systematically recorded, categorizing each pair into 'Full Match', 'Partial Match', ‘No Ukrainian equivalent’, ‘No English equivalent’.

All outputs were collected together in a single table. Each result was manually reviewed by experts to validate the accuracy of the model's assessments. Then the data was summarized in different views and the corresponding parameters were calculated by the formulas defined in the “Problem Statement” section.

This setup ensures that the experiments can be replicated by other researchers or automated tools, assuming they use the same datasets and GPT-4 model configuration. It provides a clear and systematic approach to evaluating NLP models’ accuracy in cross-linguistic analysis of clinical trial inclusion criteria.

6. Results

The summary of the results of the experiments is presented in the Table 1.

Table 1
Summary of the results

	GPT-4	Manual assessment
Full match	84	69
Partial match	30	38
Significant difference	4	12
No Ukrainian equivalent	16	15
No English equivalent	63	63
Total	197	197

Graphically they can be illustrated in the Fig. 1.

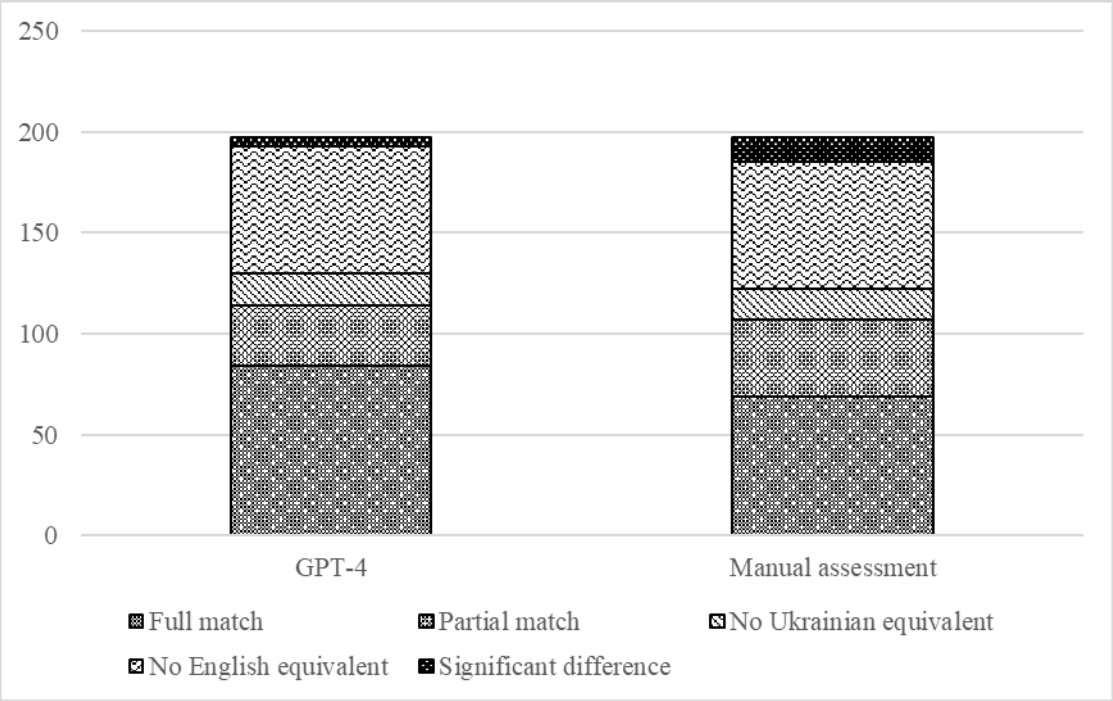

Figure 1: Graphical representation of the results.

Then we structured the data like in the Table 2 for proper accuracy, precision, and recall calculations. These metrics were selected for their relevance in evaluating the performance of NLP models, particularly in tasks involving text matching and semantic analysis.

Table 2
Summary of the results validation

Parameter	Value
True positive matches:	97
Total number of positive matches identified by the model:	114
Total number of actual positive matches across the datasets:	107

Based on that, we got the following results:

- **Accuracy (Acc).** The model achieved an overall accuracy rate of 91% in matching the correct pairs of inclusion criteria between the two languages. This measure indicates the proportion of all criteria pairs for which the model's predictions were correct to the total number of pairs.
- **Precision (P).** The model demonstrated a precision rate of 85%, reflecting the proportion of true positive matches to all matches identified by the model as correct.
- **Recall (R).** The recall rate obtained in the study was 91%. This metric measures the proportion of true positive matches to actual positive matches that the model successfully identified.

Since it is difficult to define the exact difference between “Full match” and “Partial match”, and since for some classes of tasks it is important just to identify whether there is a match or not, it is also worth looking at the results combining “Full match” and “Partial match” together and considering them as a “Match”. In this case, the data will look like the following.

Table 3
Summary of the results without the breakdown into “Full match” and “Partial match”

	GPT-4	Manual assessment
Match	114	107
Significant difference	4	12
No Ukrainian equivalent	16	15
No English equivalent	63	63
Total	197	197

Graphically this can be illustrated in the Fig. 2.

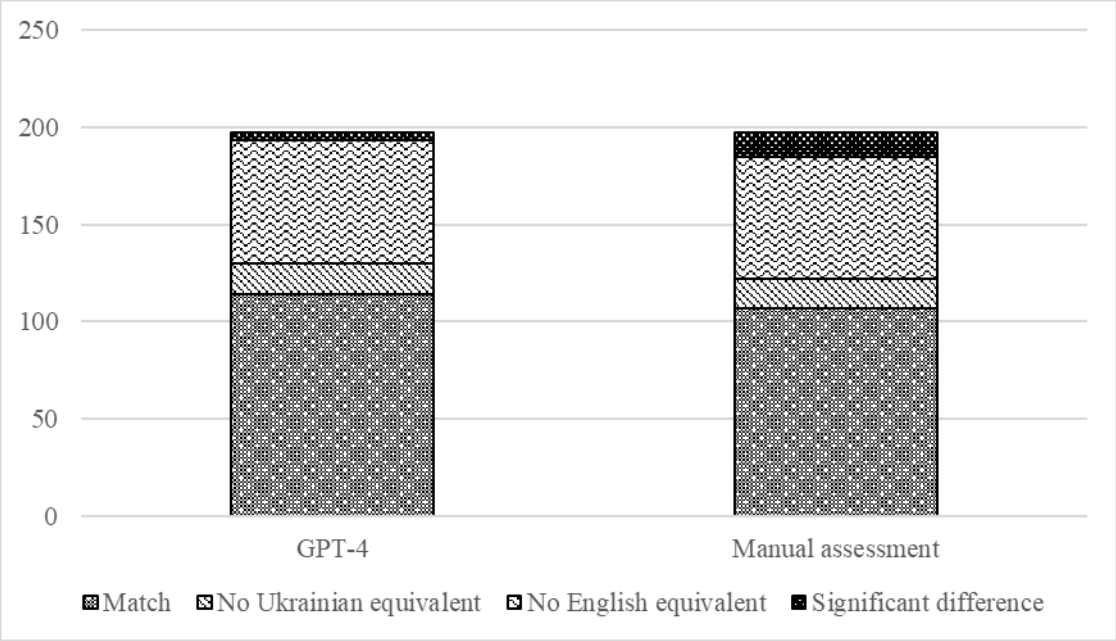

Figure 2: Graphical representation of the results without the breakdown into “Full match” and “Partial match”.

The summary of the resulting calculation is represented in the Table 4.

Table 4
Summary of the results validation without the breakdown into “Full match” and “Partial match”

Parameter	Value
True positive matches:	106
Total number of positive matches identified by the model:	114
Total number of actual positive matches across the datasets:	107

Accuracy (Acc), precision (P), and recall (R) in this case will be the following:

- Acc = 95%
- P = 93%
- R = 99%

Even though the analysis of the incorrect results is not in the scope of this research, we decided to attach the correlation between the length of a criterion in the English and the Ukrainian versions and the results. It is assumed that longer texts may contain more nuances, which could either increase or decrease the model's ability to find a correct match based on linguistic and semantic complexity. We calculated the length of each inclusion criterion by counting the words

in both the English and Ukrainian texts. This correlation is represented in Figure 3. Even though no correlation is visually noticed, this data can be used for further discussions and research.

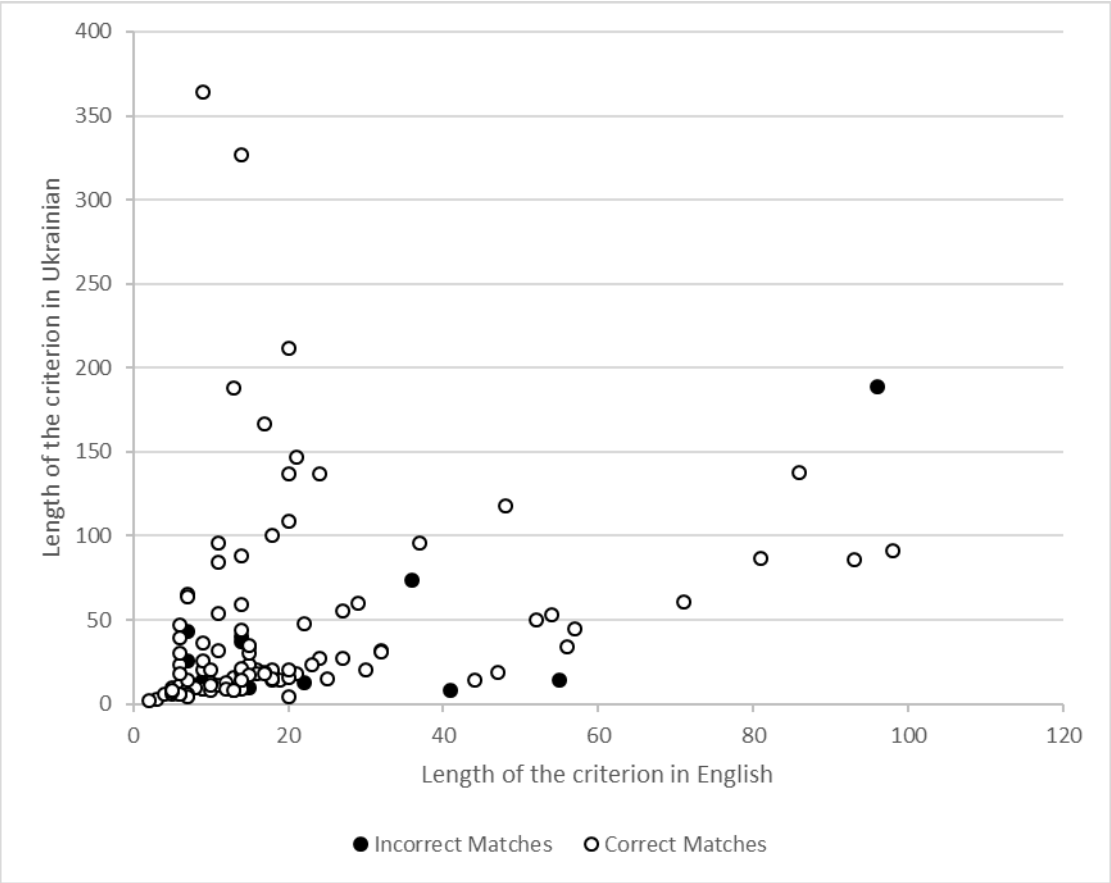

Figure 3: The correlation between the length of a criterion and the results.

In addition to the scientific observations, as a result of the experiments, the model identified four cases where unexpected significant meaningful differences had a place. In the first case, the English version of the inclusion criteria had the following text: "Full recovery from cystectomy and enrollment within 24 weeks following cystectomy.". While in the Ukrainian version, 14 weeks were mentioned. In the second example, patients with tuberculosis were specifically mentioned in the English version of the exclusion criteria but not in the corresponding Ukrainian version. In the third and the fourth, the Ukrainian version included additional screening requirements for hepatitis B and C, which were not mentioned in the English version provided.

7. Discussion

The findings from this study illustrate the advanced capabilities of the GPT-4 model in understanding and accurately matching clinical trial inclusion criteria across languages, specifically between English and Ukrainian. The high performance in terms of accuracy, precision, and recall indicates a notable advancement in the application of natural language processing technologies in clinical research. This opens a variety of opportunities for further research and development.

GPT-4 model's language understanding capabilities demonstrate not only the validity but also highlight its potential to mitigate linguistic barriers in global clinical trials. For the cases where a deep understanding of the eligibility criteria is required, the success rate of 91% in accurately matching criteria pairs between English and Ukrainian shows great potential in the model to be used. The precision rate of 85% indicates a high level of specificity in the model's output, reducing the likelihood of false positives and thereby enhancing the reliability of trial participant matching. Furthermore, the recall rate of 91% suggests that the model effectively captures almost all of the

relevant matches, i.e. minimizing the risk of overlooking suitable trial participants due to language variations.

For the cases which do not require a very deep understanding of the eligibility criteria, the results are even better. The accuracy of 95%, precision of 93%, and recall of 99% indicate the readiness of the model to be used in real-life cases. For instance, in finding discrepancies in clinical trials' documentation in different languages. Even within the scope of this research, several such cases have been identified. Another example is matching which will be followed by a human validation.

Additional practical implications of this research are diverse. It can be used for automation and validation of clinical documentation, clinical trial recruitment, patient comprehension and engagement, adherence to regulatory compliance, and others. Also, it can be used for integrations between digital health systems. By leveraging the capabilities of advanced NLP technologies, this research contributes to making clinical research more efficient, inclusive, and accessible.

The promising results of this study justify further investigation into the application of artificial intelligence in clinical research. Future studies could explore the model's applicability across a broader range of languages and medical specializations to understand its full potential and limitations. This can lead to more inclusive, diverse, and representative clinical research, contributing to the advancement of global health outcomes.

Additional analysis of the incorrect matches, explorations of correlation between different parameters and matching results, open new opportunities for further research. The complexity of trial inclusion and exclusion criteria could impact the effectiveness of AI tools in clinical settings. This observation could lead to future AI model training processes, emphasizing the need for diverse and comprehensive datasets that combine various clinical documents in different languages.

Even though the study has demonstrated great results even using a general GPT-4 model and it was confirmed by other studies, it is worth keeping in mind special models which are tuned for medicine (e.g. PubMedBERT, BioGPT, Med-PaLM 2, BioBERT, and BioMegatron, etc.). Obviously, for some of the cases, their usage will be more feasible and relevant. The upcoming release of GPT-5 is another intriguing and anticipated event which will be very interesting for further research.

In conclusion, it is important to keep in mind ethical considerations. Patient data, especially in clinical trials, is highly sensitive. Also, AI systems can make mistakes which potentially might affect humans and have a severe sequence. These ethical challenges are not unique to our study but reflect broader concerns within the AI research community, especially as AI tools become more integrated into healthcare and clinical research. Efforts must be made to ensure that the corresponding systems are trained on diverse datasets, regularly evaluated, and compliant. Qualified medical personnel should always verify the results.

8. Conclusions

This research addresses the challenges of understanding and matching clinical trial inclusion criteria across languages, specifically English and Ukrainian, utilizing the GPT-4 model. The main scientific task was to assess whether advanced natural language processing (NLP) technologies could enhance the accuracy and efficiency of cross-lingual analysis in clinical research settings. The GPT-4 model has demonstrated a robust capability to understand and match clinical trial inclusion criteria between English and Ukrainian versions, showcasing high accuracy, precision, and recall.

The application of GPT-4 represents a significant advancement in bridging language barriers in clinical research, contributing to more efficient, inclusive, and equitable clinical trial processes. The study's quantitative and qualitative results provide a strong basis for the practical application of artificial intelligence in enhancing global health research methodologies in multilingual environments.

Scientific Novelty. The use of GPT-4 for matching English and Ukrainian clinical trial inclusion criteria matching demonstrated a high degree of scientific novelty, marking the first instance of such an application in the field. By breaking down and analyzing inclusion criteria, the study demonstrates how deep learning models can analyze and understand complex medical

terminology and nuances across languages. Compared to previously known methods, the GPT-4 model has shown superior performance, which allows for a more accurate and inclusive participant selection process, decreasing linguistic biases and barriers in global clinical research, and provides other advantages.

Practical Significance. The findings offer a practical framework for applying AI and NLP technologies in improving global clinical trial processes, particularly in patients' recruitment and documentation. The research outcomes can be used by Clinical Research Organizations (CRO), researchers, trial coordinators, and regulatory bodies to enhance the inclusivity, efficiency, and accuracy of clinical trials across different linguistic regions. It can also be used by regular people and patients who are struggling with severe disease and desperately looking for the most recent medicine which can potentially help with their treatment and give hope for a better and healthier life.

Recommendations for Practical Use include integrating NLP tools like GPT-4 into clinical trial management systems to automate and optimize the screening and matching of participants. In particular, considering the positive results of Ukrainian clinical text interpretation, Clinical Research Organizations (CRO) should seriously consider integrations with E-health (a medical information system for the Ministry of Healthcare of Ukraine) or local Electronic Health Records (EHR) systems like HELSI, DrEleks, and others.

Another area of use can be related to centralized data validation and minimization of human errors. Even during the experiments of this study, we have identified four significant differences in comparison of English and Ukrainian inclusion criteria, most likely caused by a human factor. Such mistakes can potentially cause incorrect patients' selection and affect their safety as well as the clinical study's results.

Further research should explore the application of GPT-4 and similar models across a broader range of languages and clinical contexts. Investigating the integration of NLP tools with other digital health systems like Clinical Trial Management Systems (CTMS) and Electronic Health Records (EHR) could yield additional improvements in clinical trial design and execution. Longitudinal studies to track the impact of NLP integration on trial outcomes and patient diversity could provide deeper insights into the long-term benefits and challenges of AI in clinical research.

9. Acknowledgements

The authors would like to thank the anonymous reviewers for their detailed and constructive feedback.

References

[1] F. Cascini, F. Beccia, F.A. Causio, A. Melnyk, A. Zaino and W. Ricciardi. Scoping review of the current landscape of AI-based applications in clinical trials. Frontiers Public Health. 10:949377. (2022). doi:10.3389/fpubh.2022.949377

[2] Bin Zhang, Lu Zhang, Q. Chen, Zhe Jin, S. Liu, S. Zhang. Harnessing artificial intelligence to improve clinical trial design. Communication Medicine 3, 191 (2023). doi:10.1038/s43856-023-00425-3

[3] A.F. Hernandez, C.J. Lindsell. The Future of Clinical Trials: Artificial to Augmented to Applied Intelligence. JAMA. November 11. 330(21): (2023) 2061–2063. doi:10.1001/jama.2023.23822

[4] E. H. Weissler, T. Naumann, T. Andersson et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials 22, 537 (2021). doi:10.1186/s13063-021-05489-x

[5] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, T. Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J. Biomedical Informatics. V.73, (2017) 14-29. doi:10.1016/j.jbi.2017.07.012

[6] Y. Gao, D. Dligach, L. Christensen, S. Tesch, R. Laffin, D. Xu, T. Miller, O. Uzuner, M.M Churpek, M. Afshar. A scoping review of publicly available language tasks in clinical natural language

processing. J. of the American Medical Informatics Association. V.29. Issue 10. (2022) 1797-1806. doi:10.1093/jamia/ocac127
[7] Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, Jan A. Kors. Natural Language Processing in Radiology: A Systematic Review. Radiology. May;279(2). (2016) 329-43. doi:10.1148/radiol.16142770
[8] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, D. Mitsoura. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics. Jan-Feb;36(1). (2016). doi:10.1148/rg.2016150080
[9] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan et al. Large language models in medicine. Nature Medicine. 29, (2023) 1930–1940. doi:10.1038/s41591-023-02448-8
[10] I. J. B. Young, S. Luz, N. Lone. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. International Journal of Medical Informatics. V.132. (2019). doi:10.1016/j.ijmedinf.2019.103971
[11] A. Wong, J. M. Plasek, S. P. Montecalvo, L. Zhou. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy. Aug;38(8). (2018) 822-841. doi:10.1002/phar.2151
[12] L. P. Nijhawan, M. D. Janodia, B. S. Muddukrishna, K. M. Bhat, K. L. Bairy, N. Udupa, P. B. Musmade. Informed consent: Issues and challenges. J. of Advanced Pharmaceutical Technology & Research. Jul;4(3). (2013). doi:10.4103/2231-4040.116779
[13] J. Kim, Y. Quintana. Review of the Performance Metrics for Natural Language Systems for Clinical Trials Matching. Studies in Health Technology and Informatics. Jun 6;290. (2022) 641-644. doi:10.3233/SHTI220156
[14] C. Fang, Y. Wu, W. Fu, J. Ling, Y. Wang, X. Liu, Y. Jiang, Y. Wu, Y. Chen, J. Zhou, Z. Zhu, Z. Yan, P. Yu, X. Liu. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. Dec 1;2(12). (2023). doi:10.1371/journal.pdig.0000397
[15] S. Voloshyn, V. Vysotska, O. Markiv, I. Dyyak, I. Budz and V. Schuchmann, "Sentiment Analysis Technology of English Newspapers Quotes Based on Neural Network as Public Opinion Influences Identification Tool," 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2022, pp. 83-88, doi: 10.1109/CSIT56902.2022.10000627
[16] D. Ueda, S. L. Walston, T. Matsumoto et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health 2, 4 (2024). doi:10.1186/s44247-023-00058-5
[17] N. Oh, G.-S. Choi, W. Y. Lee. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Annals of Surgical Treatment and Research. May;104(5). (2023) 269-273. doi:10.4174/astr.2023.104.5.269
[18] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, N. Zaman, P. Sarker, A. G. Lee, A. Tavakkoli. GPT-4: a new era of artificial intelligence in medicine. Irish Journal of Medical Science. V.192(6). (2023) 3197-3200. doi:10.1007/s11845-023-03377-8
[19] G. A. Guerra, H. Hofmann, S. Sobhani, G. Hofmann, D. Gomez, D. Soroudi, B.S. Hopkins, J. Dallas, D. J. Pangal, S. Cheok, V. N. Nguyen, W. J. Mack, G. Zada. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. World Neurosurgery. V.179. (2023) e160-e165. doi:10.1016/j.wneu.2023.08.042
[20] Sang-Wook Lee, Woo-Jong Choi. Utilizing ChatGPT in clinical research related to anesthesiology: a comprehensive review of opportunities and limitations. Anesthesia and Pain Medicine. 18(3). (2023) 244-251. doi:10.17085/apm.23056
[21] M. Rosoł, J. S. Gąsior, J. Łaba, K. Korzeniewski & M.Mlynczak. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Scientific Reports. 13, 20512 (2023). doi:10.1038/s41598-023-46995-z
[22] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White & E. Horvitz. Can

Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. ArXiv. Computer Science. Computation and Language. (2023)
[23] A. Wang, X. Xiu, S. Liu, Q. Qian, S. Wu. Characteristics of Artificial Intelligence Clinical Trials in the Field of Healthcare: A Cross-Sectional Study on ClinicalTrials.gov. International Journal of Environmental Research and Public Health. 19(20). (2022). doi:10.3390/ijerph192013691