dr hab. Konrad Furmańczyk, prof. SGGW
Institute of Information Technology, Warsaw
University of Life Sciences
konrad_furmanczyk@sggw.edu.pl

07 January 2025

**Reviewer's opinion**
**on PhD dissertation authored by**
*Klaudia Kantor*
**entitled:**
*Neural language models for clinical trial eligibility criteria*

## 1. Problem and its impact

The main problem of the dissertation is to investigate the application of neural language models in automating the parsing of eligibility criteria for clinical trials. The second problem considered in this dissertation is to provide a comprehensive evaluation of neural language models in biomedical applications. Both problems are of a scientific nature and are very important for applications in the pharmaceutical industry and for medical applications in clinical trials. Finally, based on this dissertation, a prototype of the eligibility criteria parsing tool was created to meet the needs of Roche Pharmaceutical Corporation. This prototype was designed in collaboration with domain experts. The research presented in the dissertation is challenging as there are no reliable comparative datasets to evaluate models that process eligibility criteria and no reliable methods to evaluate the application of the LLM in this context.

## 2. Contribution

The main original contribution of the dissertation is the development of a prototype tool that translates eligibility criteria into a machine-readable format that can support a trial screening algorithm for patient-trial matching. Another important contribution of this dissertation is the evaluation of different Natural Language Processing (NLP) models in biomedical applications in a pharmaceutical context.
The dissertation carried out as part of an industrial PhD programme, includes a comprehensive scoping review of NLP and Machine Learning (ML) solutions for parsing eligibility criteria.
In particular, in Chapter 3, a study was carried out on the prediction of several operational efficiency metrics using ML models, with the identification of the trial design features that influence these metrics. This chapter is based on the PhD candidate's research paper [18].
In Chapter 6, based on the PhD candidate's conference proceedings [100], different techniques for sentence embedding and their effectiveness in semantic textual similarity search for the biomedical domain were evaluated. In this context, static embeddings, transformers-based representations and sentence transformers are considered. The results of an experiment on benchmark data sets showed that the CODER and BERT transformers were more effective than competitors.

In Chapter 7, generative language models with prompt engineering are investigated in the named entity recognition task in clinical trials. In conclusion, BERT-based pre-trained models perform better than the GPT-4 model.
The implementation of the eligibility criteria parsing tool is presented in Chapter 10. This chapter presents the technical implementation of the proof of concept for the eligibility criteria parsing tool, focusing on breast cancer trials. This AI tool is an assistant rather than an autonomous solution with expert domain supervision. The proposed criteria parsing tool is evaluated in a specially designed experiment. LLM-assisted parsing is based on the GPT-4o model with advanced prompt engineering techniques (chain-of-though and few-shot prompting with negative examples). In conclusion, the LLM-assisted parsing tool performs well, has high accuracy and reduces the time required compared to the manual approach.
In addition, the PhD candidate presented two papers as part of the dissertation. The first paper [18] was published in a good international journal, AAPS, with a 5-year Journal Impact Factor of 4.1; the second paper [100] is a conference proceedings of rank B International Conference on Artificial Intelligence in Medicine.

## 3. Correctness

The quality of the prototype tool for parsing clinical trial eligibility criteria is high. All descriptions and explanations are well written and easy to read. Most of the empirical experiments are well designed and analysed in appropriate ML tools. I have only a few comments and questions:

1. On page 16, Subsection 3.4.2, we have the following sentences: 'This uncertainty is quantified by producing predictive intervals using a quantile loss function, trained at quantiles 0.05 and 0.95 to achieve a 90% predictive interval. Point estimates are derived from the 0.5 quantile, representing the median.' but the prediction intervals are not given in this chapter.

2. In Chapter 7, it is not clear whether BERT models use the prompts in the experiment in this chapter.

3. In Chapter 10, on page 116, we have the following sentences: 'A partial narrowing of the trial set is beneficial, as the goal is to reduce the pool of trials, without being overly restrictive. It is preferable to have broader criteria that might include some irrelevant trials than to risk excluding relevant ones, which could prevent a patient from accessing a potentially suitable treatment. '
I do not understand how we can strictly guarantee that we do not lose relevant trials in this work. Also, the high probability of 'not losing relevant trials' is a challenging task in practice.

4. In Chapter 10, it would be worthwhile to select a larger sample of experts and a larger sample of tasks.

5. Inclusion criteria for clinical trials can be very complicated and written in very hermetic language. This adds additional complications to the use of a fully or partially automated tool for assigning patients to a trial.

6. The results obtained should be treated as preliminary results for further research, which may reduce the cost and time of allocating patients to clinical trials in the future. However, more clinical trials and more possible inclusion and exclusion criteria should be considered.

7. It is not clear to me whether large language models supported by prompt engineering will perform very well in other diseases and in clinical trials dedicated to them. Perhaps some variant of reinforcement learning could be used in future work.

## 4. Knowledge of the candidate

Chapters 2-6 and 7 are similar to a tutorial and thus confirm a general knowledge of the candidate in the discipline of **Information and Communication Technology**. Those chapters covered machine learning, natural language processing, generative AI, and applications in the biomedical domain. These chapters are of high quality. The reference list in the dissertation is almost complete and includes classic and recent conference papers. The dissertation is written at a high level, demonstrating the candidate's ability to use modern methods of machine learning, generative AI and large language models. The application aspect of the work and the doctoral candidate's experience in this area are also very important. All this proves that the PhD candidate has a general knowledge and understanding of the discipline of **Information and Communication Technology**.

## 5. Other remarks

The dissertation is well written in good English. All chapters contain an introduction and conclusions. The editing of the thesis is of a high standard and appropriate for a doctoral thesis.

## 6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the *Act of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

**A.** Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

| [X] | [ ] | [ ] | [ ] | [ ] |
|---|---|---|---|---|
| *Definitely YES* | *Rather yes* | *Hard to say* | *Rather no* | *Definitely NO* |

**B.** After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of **biomedical engineering.**

| [X] | [ ] | [ ] | [ ] | [ ] |
|---|---|---|---|---|
| *Definitely YES* | *Rather yes* | *Hard to say* | *Rather no* | *Definitely NO* |

**C.** Does the dissertation support the claim that the candidate is able to conduct scientific work?

| [X] | [ ] | [ ] | [ ] | [ ] |
|---|---|---|---|---|
| *Definitely YES* | *Rather yes* | *Hard to say* | *Rather no* | *Definitely NO* |

Konrad Fuksiewicz
*Signature*

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276