Paper IC01

Large Language Models to support RBQM and Clinical Trial Developments

Laura Trotta, CluePoints, Louvain-la-Neuve, Belgium
Sebastiaan Höppner, CluePoints, Louvain-la-Neuve, Belgium
Nicolas Huet, CluePoints, Louvain-la-Neuve, Belgium

ABSTRACT

The release of GPT-4 has highlighted the capabilities of generative large language models in mimicking human language. Large language models are deep learning models trained to solve various natural language processing tasks. While these models are extremely powerful, they also come with limitations: they are trained on large volumes of openly available text data, such as the English Wikipedia corpus. While these sources of data are perfect to acquire language processing skills, they may fail to produce the specific, accurate, and reliable content needed to support the clinical trial process. In this paper, we show how our research team fine-tuned large languages models on trusted clinical databases to produce large language models that produce accurate clinical content. These models are used as building blocks of more complex deep learning architectures developed to support Risk-Based Monitoring and clinical trial developments.

INTRODUCTION

Risk-Based Quality Management (RBQM) is an approach to managing risk throughout the life cycle of clinical trials, ensuring continuous risk assessment from study design to planning and execution. Previous research has demonstrated the effectiveness of an RBQM implementation powered by advanced statistical and machine learning algorithms [1]. Targeted analyses based on Key Risks Indicators (KRIs) and holistic approaches based on the unsupervised analysis of the entire clinical trial data are effective methods for early risk detection [2,3]. This enables study teams to mitigate risks during study conduct and correct issues as diverse as missing data, miscalibration of equipment, protocol misunderstandings, and failure to comply with Good Clinical Practice (GCP) [4].

Previous work has highlighted the potential of ML in improving clinical research with applications in drug target identification, clinical trial protocol optimization, clinical trial participant management, data collection and data management [4]. ML is the field that gives computers the ability to learn without being explicitly programmed. It covers a wide variety of models ranging from simple linear regressions and decision trees to advanced deep neural networks. The past decade has seen major progress in Deep Learning (DL), a family of advanced ML models that rely on deep neural networks. Among the DL models, Large Languages Models (LLMs) have significantly improved Natural Language Processing (NLP) [6,7]. Generative LLMs are LLMs used to generate data such as new text. The recent release of ChatGPT (further enhanced by the release of GPT-4 has highlighted the capabilities of generative LLMs in mimicking human language and has led many industries to assess the benefits of these most recent technologies in supporting their daily operations.

This paper discusses the use of LLMs to support the clinical trial process. We share our experience with Large Language Models acquired in the context of three use cases from RBQM and clinical trial development. We believe that the insights gained from these use cases naturally extend to any application designed to support RBQM and clinical trial development. Our results highlight the benefits of developing specific LLMs targeted to the needs of our industry and emphasize the importance of scientific validation of LLMs for clinical use.

METHODS AND RESULTS

A series of experiments were conducted in the context of hackathons, co-development partnerships with industrial sponsors and internal research.

DEFINITIONS

Large Language Models (LLMs): LLMs are deep neural networks that rely largely on the Transformer architecture and trained on large volumes of text data. They are used as foundation models for multiple NLP tasks, including text translation, text summarization, question answering, etc. As an example, BERT (the Bidirectional Encoder Representations from Transformers) was trained on Wikipedia (~2.5B words) and Google's BooksCorpus (~800M words) and used as the underlying model for more than eleven NLP tasks [7]. Table 1 provides an overview of some of the most famous LLMs together with their number of parameters, owner and release date. When those models are proprietary, a profusion of open-source LLMs have been released recently. Table 2 provides an overview of the most used open source LLMs.

Table 1. Proprietary LLMs – Overview.

Name	Number of parameters	Owner	Release date
LaMDA	137 billions	Google;	Jan 2022
Chinchilla;	70 billions	DeepMind	March 2022
GPT-4	~1.7 trillion	OpenAI	March 2023
: Gemini	Unknown	Google	December 2023

Table 2. Open-source LLMs – Overview.

Name	Number of parameters	Owner	Release date
StableLM	7 billions	Stability AI	April 2023
Llama 2	70 billions	Meta	July 2023
Falcon 180B	180 billions	Technology Innovation Institute	September 2023
Mistral 8X7B	45 billions;	Mistral AI	December 2023

Fine tuning of LLMs: Fine-tuning an LLM is the process of creating a model that acquires specific knowledge about the task and the domain. If we compare a pre-trained LLM with a person with a good understanding of the language and a wide knowledge, the process of fine-tuning consists of providing this person with the explanations necessary to perform the specific task assigned to them. As the model already understands natural language and has reasoning capabilities, the model understands what is expected of it much faster. With the advent of LLMs with vast general knowledge and exceptional comprehension, fine-tuning has been increasingly replaced by prompt engineering, which is the process of correctly explaining (with possibly some examples) the task to the model, but without any modification of the internal parameters of the model. Although very efficient, this technique also has limitations [8].

Generative LLM: Generative LLMs are LLMs generating data. These LLMs are trained to generate text content, most of the time by predicting the next word of a sentence. On the architecture side, they are usually created based on the decoder part of the transformer architecture, while non-generative LLMs are usually created using the encoder part.

EXPERIMENT 1: Improve documentation of risk signals and promote active risk follow-up – Fine-tuned LLM

Model developments were performed with the objective of enabling a more active follow-up and documentation of risk signals detected by central statistical monitoring activities. Over the years, study teams from 111 organizations have analyzed a database of 1417 trials, generating more than 178,270 risk signals. A risk signal is a set of statistical anomalies that are grouped and presented together to the study team. A risk signal contains a text description used to describe the set of anomalies detected at the site/patient/region/country level. When a risk signal is created, the study team follows up and creates a series of actions to mitigate the risk signal. The study team documents follow-up activities by adding text comments and mitigation rationales to the risk signal. Ensuring active follow-up is critical to ensure that risks signals are mitigated early during study conduct.

However, study teams may have difficulty actively monitoring and documenting risk signals. To assist study teams in these activities, our team developed a deep learning model that screens the signal documentation in real time and detects risk signals with poor documentation and lack of follow-up activities. The model is designed not only to detect a lack of documentation, but also to detect inconsistent classification of risk signals into non-issues when the documentation points to issues at the site. The model was developed by using a version of the BERT language model that was fined-tuned on a database of reviewed risk signals. The risk signals were reviewed together with comments and rationales by Subject Matter Experts (SMEs) to create a trusted database for learning. Performance was assessed by using an independent test set from the same database. An overall accuracy of 70% was reached in the classification of signal documentation.

The feature was released into production and an analysis was performed after active usage of the feature by 35 organizations on 279 studies and 19,404 risk signals. The results showed that the proportion of signals without documentation decreased by 25% after the feature was introduced [9]. By prompting information to the user, the feature promotes the active follow-up of risk signals and enables more reliable documentation of risk follow-up activities.

EXPERIMENT 2: Enhance risk detection with information from study protocol – Use of GPT-4

Risk detection would benefit from better accounting for study parameters such as the schedule of assessments or

protocol specific requirements. Information extracted from the study protocol could be used to enable a study-specific detection of risk.

We performed a series of experiments to assess whether information could be easily extracted from study protocols using Large Language Models. We connected to the clinicalTrial.gov database and downloaded a series of study protocols. We used the GPT-4 API to ask a series of questions and extract information from the study protocol. Examples of questions included “Can you list all the inclusion criteria for this study?” or “Can you tell the number of treatment arms of this study?”. The team assessed the capabilities of the tool in extracting the right information from study documents.

Our analysis showed that GPT-4 is powerful in extracting protocol information. In particular, the model was almost always able to provide the correct answer when presented with the right part of the protocol. However, our analysis revealed one limitation of the current solution that must be overcome to automatically feed information from the study protocol. GPT-4 has a limited entry size for text: the text size that GPT-4 can accept as input is limited. It is not possible at this stage to inject the entire study protocol into the GPT-4 prompt at once. Instead, only a limited number of pages can be processed at once.

The way this limitation is usually overcome is by using a technique like semantic search, where the documents are segmented in chunks and associated to a semantic index. This index is then used to search for the most relevant chunk of document for a given question [10]. Although powerful, this technique has several potential limitations like the use of a universal semantic index [11].

EXPERIMENT 3: Medical coding of AEs - GPT-4 versus fined-tuned LLM

In a clinical trial, medical coders map verbatim entries for adverse events, drugs and medical histories to their dictionary entries in MedDRA and WHODrug. To understand the capabilities of GPT-4 in medical coding, we used GPT-4 to map twenty adverse events to their dictionary entry in MedDRA. In the prompt, the task of medical coding was described together with an example. Figure 1 shows both the input and output of GPT-4 for the medical coding task.

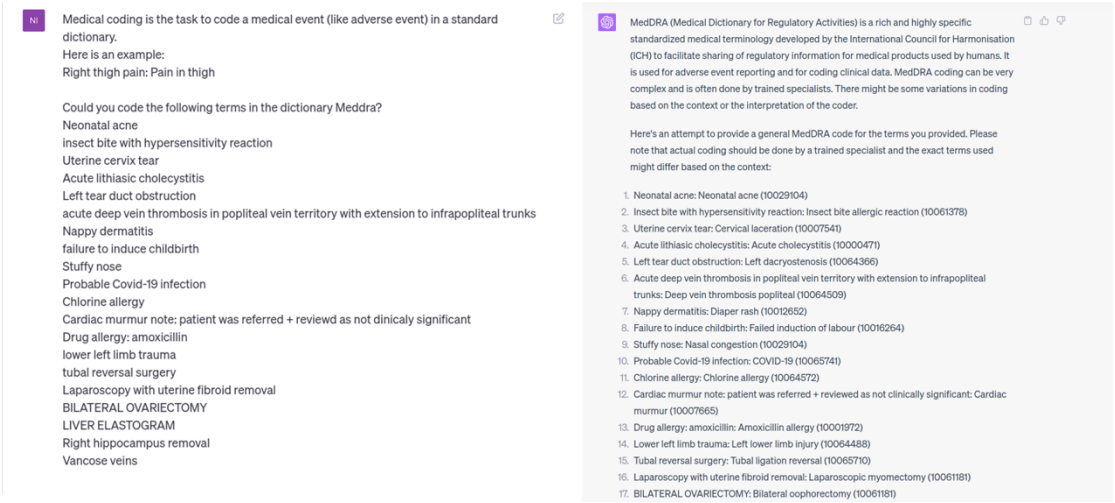

Figure 1. GPT-4 for medical coding. Input provided to GPT-4 (left) and output from the model (right).

Although GPT-4 was asked to code adverse events to their dictionary entry in MedDRA, it was able to understand the tasks and demonstrated that it has knowledge of the MedDRA dictionary. The answer was nicely formatted and presented to the user. However, a comparison of the coded terms with terms coded by Subject Matter Experts revealed that only 8 out of 20 terms were coded correctly yielding an accuracy of 40% on the task. In addition, the model produced hallucinations. Indeed, the model produced 9 coded terms that did not even exist in the MedDRA dictionary. This result is not surprising as GPT-4 does not have direct access to MedDRA and relies solely on its memory to perform medical coding. This is inefficient and prone to errors. When used directly, GPT-4 does not allow to reach the accuracy required for clinical use.

To achieve accurate clinical predictions, our team tested the potential of a smaller open-source LLM (BERT-like) fine-tuned on a database of coded terms together with their mapped dictionary entries. Top 1 and Top 5 suggestions for coded terms were presented to SMEs for review. The accuracy, i.e. the number of correctly coded terms divided by the total number of terms, was computed. Results showed that both the Top-1 and Top-5 accuracy reached the sponsor-set performance goal (>90%).

DISCUSSION

I. General purpose LLMs (such as GPT-4) demonstrate impressive capabilities but still present difficulties in producing accurate clinical content.

Our experiments with GPT-4 have shown that general purpose LLMs (such as GPT-4) offer impressive capabilities. Results from Experiment 2 showed that LLMs can extract information from study documents and format answers in standard formats. However, Experiment 3 revealed one of the key limitations of general-purpose LLMs in supporting RBQM or clinical trial development activities. These solutions may fail to produce accurate clinical content. Our experience with medical coding has shown that these generic solutions perform poorly when addressing specific clinical questions, such as mapping adverse event terms to their MedDRA dictionaries. Further literature review revealed there have been recent attempts to quantify the performance of general-purpose LLMs in encoding clinical knowledge. A series of general-purpose LLMs were tested in the context of answering medical questions. The authors showed that general-purpose LLMs were unable to compete with clinicians (SMEs) in addressing medical questions. The authors have shown that performances can be improved by designing instruction prompt tuning. However, the authors called for further model development in creating safe and useful LLMs for clinical applications and emphasized the importance of defining evaluation frameworks and addressing important gaps in current predictions.

II. Small fined-tuned LLMs can produce accurate predictions

Recent research has shown that general-purpose LLMs (like GPT-4) often perform worse than LLMs that are trained specifically on clinical data [12]. This is further illustrated by our experiments.

In Experiment 1, fine-tuning a small LLM (BERT-like model) was sufficient to achieve good capabilities for understanding the documentation from risk signals generated by central statistical monitoring activities. The model was able to perform a specific task that required training and could only be performed by Subject Matter Experts (SME). The good model performances achieved in this use case demonstrated that the model understood the clinical trial lexical space (e.g. AE, site, …) and the task of risk signal documentation.

In Experiment 3, fine-tuning a small LLM (BERT-like model) on a historical database of adverse events mapped to their dictionary entries demonstrated a much higher accuracy than coded terms generated by GPT-4.

III. LLMs used in a non-generative mode produce reliable content

Experiments 1 and 3 show that LLMs used in a non-generative mode produce reliable content. Experiment 3 further highlights one major issue of LLMs when used in a generative mode. One identified issue is the inclinication of generative models to generate hallucinations. In Experiment 3, GPT-4 invented coded terms that were not included in the MedDRA dictionary entries. Hallucinations can have serious consequences in a clinical environment where users are presented with answers that seem realistic but are completely irrelevant from a clinical point of view. The potential harm of hallucinations for clinical applications was discussed in [13].

CONCLUSION

The clinical trial landscape is rapidly evolving and bringing new challenges. Clinical trial protocols are becoming more complex and collect larger volumes of data. The typical Phase III protocol collects an average of 3.5 million data points, seven times more than fifteen years ago [14]. In that context, effectively monitoring clinical trial risks becomes even more challenging. Recent developments in Machine Learning (ML) provide great opportunities to overcome those challenges.

The development of LLMs opens up new opportunities to further support RBQM and clinical trial development activities. In this paper, we share our experiences with LLMs applied to clinical trial data. Our results show that general purpose LLMs (such as GPT-4) offer impressive capabilities but may fail to make accurate predictions. In contrast, smaller LLMs fined-tuned on trusted clinical database can achieve the accuracy needed to support clinical trial activities. Finally, LLMs use in non-generative mode are more reliable and overcome the challenge of model hallucinations.

Our results were illustrated through three specific use cases, enabling highly effective documentation and follow-up of risk signals, improving risk detection using study-specific information from the protocol, and automating resource-intensive tasks such as the mapping of adverse events. These are just three potential applications of LLMs to support clinical trial oversight activities. We believe these models offer many more opportunities for our industry. This paper is a first attempt to develop a deeper understanding of LLMs applied to clinical trial data. We call on our industry to openly share results from their experiments on LLMs. The benefits and risks of these models need to be discussed, and significant effort should be made to define a common framework for testing and validating models.

REFERENCES

1. Venet, D., Doffagne, E., Burzykowski, T., Beckers, F., Tellier, Y., Genevois-Marlin, E., Becker, U., Bee, V., Wilson, V., Legrand, C., & Buyse, M. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical trials*, 9(6), 705–713.
2. de Viron, S., Trotta, L., Steijn, W., Young, S., & Buyse, M. (2023). Does Central Monitoring Lead to Higher Quality? An Analysis of Key Risk Indicator Outcomes. *Therapeutic innovation & regulatory science*, *57*(2), 295–303.
3. de Viron, S., Trotta, L., Steijn, W., Young, S., & Buyse, M. (2023). Does Central Statistical Monitoring Improve Data Quality? An Analysis of 1,111 Sites in 159 Clinical Trials. (Manuscript accepted for publication - TIRS)
4. de Viron S, Trotta L, Schumacher H, Lomp HJ, Höppner S, Young S, & Buyse, M. (2022). Detection of Fraud in a Clinical Trial Using Unsupervised Statistical Monitoring. *Therapeutic innovation & regulatory science*, *56*(1), 130–136
5. Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., Slater, P., Shameer, K., Roe, M., Hutchison, E., Kollins, S. H., Broedl, U., Meng, Z., Wong, J. L., Curtis, L., Huang, E., … Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1), 537.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaeser, L., et al. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17*).
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (1)*: 4171–4186, 2019.
8. Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., et al. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
9. Young, S., & de Viron, S. (2023). Using machine learning and NLP to improve central monitoring documentation. *Applied Clinical Trials*, 32(12), 8-8.
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
11. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110.*
12. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1), 135.
13. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.
14. Getz, K., Smith, Z., & Kravet, M. (2023). Protocol design and performance benchmarks by phase and by oncology and rare disease subgroups. *Therapeutic Innovation & Regulatory Science*, 57(1), 49-56.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Laura Trotta
Company: CluePoints
Address: CluePoints SA, New Tech Centre, Avenue Albert Einstein 2A, 1348 Louvain-La-Neuve, Bel
Work Phone: +32 10 77 89 10
Email: laura.trotta@cluepoints.com
Website: https://cluepoints.com/