Fine-Tuned Transformers and Large Language Models for Entity Recognition in Complex Eligibility Criteria for Clinical Trials

Klaudia Kantor
Roche Poland & Doctoral School of Poznan University of Technology
Warsaw, Poland *klaudia.kantor@roche.pl*

Mikołaj Morzy
Poznan University of Technology
Poznan, Poland *mikolaj.morzy@put.poznan.pl*

Abstract

This paper evaluates the `gpt-4-turbo` model's proficiency in recognizing named entities within the clinical trial eligibility criteria. We employ prompt learning to a dataset comprising 49 903 criteria from 3 314 trials, with 120 906 annotated entities in 15 classes. We compare the performance of `gpt-4-turbo` to state-of-the-art BERT-based Transformer models[1]. Contrary to expectations, BERT-based models outperform `gpt-4-turbo` after moderate fine-tuning, in particular in low-resource settings. The `CODER` model consistently surpasses others in both low- and high-resource environments, likely due to term normalization and extensive pre-training on the UMLS thesaurus. However, it is important to recognize that traditional NER evaluation metrics, such as precision, recall, and the F_1 score, can unfairly penalize generative language models, even if they correctly identify entities.

Keywords: prompt learning, LLM, clinical trial eligibility criteria, named entity recognition.

1. Introduction

Eligibility criteria ensure that participants meet the necessary characteristics for safe participation in the trial and unbiased results. These criteria, usually based on age, health status, and other relevant factors, make study results generalizable to the target population. They are classified into inclusion criteria, which define required traits, and exclusion criteria, which list disqualifying traits. Complex criteria often hinder the achievement of recruitment quotas and the progress of trial phases [11]. Named entity recognition (NER) is crucial in parsing these criteria, allowing quick extraction of demographic and medical information from protocols. This accelerates patient recruitment and improves data accuracy.

NER is commonly performed as a sequence tagging task. Each token in an input sequence is labeled either as `O` (outside of an entity), `B-entity` (beginning of an entity), or `I-entity` (inside a multi-token entity). Entity categories can vary, but traditionally include persons, organizations, numbers, locations, and dates. Advanced models can also tag entities such as works of art, events, and currencies. During the past decade, significant effort has been dedicated to developing models for extracting information from biomedical texts. These models identify and classify entities such as diseases, drugs, genes, and proteins and recognize complex relationships and patterns in the data. NER models such as `BioBERT`, `ClinicalBERT`, and `BioMedicalRoBERTa` have been successfully applied to tasks such as literature-based discovery, clinical information extraction, and drug development. The past four years have seen rapid advances in large language models (LLMs). Starting with BERT, which achieved state-of-

[1] Due to page limits, detailed results and code listings are presented in the supplementary material available at https://github.com/megaduks/isd24

the-art results in many NLP tasks, subsequent models like T5, GPT-3, and GPT-4 have further improved performance. These models introduced prompt engineering, a technique for crafting specific inputs to guide model outputs. Prompts can be used to extract instances of named entity classes from text.

This paper examines the utility of prompt engineering for extracting named entities from clinical trial eligibility criteria. Although generative LLMs have shown promise in biomedical NER, we focus on their use for eligibility criteria. Our study differs by evaluating generative models against state-of-the-art BERT models, fine-tuned on a very limited dataset using few examples for in-context learning. We use a simple prompt without additional guidelines, simulating a scenario with limited annotated data and minimal input from domain experts. The research hypothesis posits that large language models can recognize domain-specific biomedical vocabulary and differentiate between various classes of eligibility criteria. To test this, we compare the effectiveness of a prompt engineering-based model with state-of-the-art NER models. Contrary to expectations, we find that pretrained BERT-based models outperform gpt-4-turbo in NER tasks, even in low-resource scenarios. When ample annotated data are available, the performance gap widens further. Models like CODER [20] or SciBERT [3] also exceed gpt-4-turbo even with minimal annotations.

2. Few-shot Prompt Engineering for Entity Recognition

Prompt engineering involves creating token sequences to enhance the accuracy and generalization of LLMs. By presenting a specific prompt, the model can focus on particular language patterns or contexts. This technique aligns well with transfer learning, where a model trained on one task adapts to another. In our study, the aim is to adapt a general-purpose language model trained on masked token prediction to the sequence tagging task, marking tokens as belonging to an entity class or not. Individual eligibility criteria are often long, and entity spans can be short, so a prompt like "*[CRITERION]. List examples of drugs in this text*" typically yields no useful results. However, LLMs respond well to atypical text patterns. By using a template that includes several examples of eligibility criteria with marked entities, we guide the large language model to produce the desired output.

After experimentation, we selected a template for few-shot learning within the prompt. For each entity class, we randomly chose five eligibility criteria and listed the annotated entities. We included both positive examples (criteria where the entity appears) and negative examples (criteria with no instances of the entity) to guide the model. An example prompt template for the entity class *cancer* is presented in the supplementary material. This template, used as input to the LLM, consists of the same few-shot examples for all classes, changing only the criterion being analyzed. The model is expected to generate a list of cancers mentioned, such as *[medullary thyroid cancer (MTC), RET-altered solid tumor].*

3. Data set and metrics

In our experiments, we used an annotated data set from *Clinical Trial Parser* [18], which contains eligibility criteria for 3314 interventional trials in the U.S. The sample was downloaded from the AACT Database using the 2020-04-16 copy. The criteria were split into 49 903 samples and annotated by professionals, producing 120 906 labeled entities. The distribution of labels and examples of entity spans annotated in the *Clinical Trial Parser* data set are presented in the supplementary material. Generally, the annotated text is highly specialized, with abbreviations, domain-specific terms, and proper names, making it a challenging data set for any NER model.

In our experiments, we focus on five medical entities relevant to parsing eligibility criteria: *treatment*, *chronic disease*, *clinical variable*, *cancer*, and *allergy name*. For the gpt-4-turbo prompts, we selected 22 samples from 17 trials. To simulate low availability of annotated data, we randomly selected two subsets: a larger set of 100 trials and a smaller set of 27 trials, used for Transformer fine-tuning. We fine-tuned BERT-based models in two scenarios: high-resource with 80 trials in training data set (1243 samples) and 20 trials in validation data set (376 samples), and low-resource with 17 trials in training data set (448 samples) and 10 trials in validation data set (213 samples). The 17 trials used for a few-shot prompt were included in the training data sets only. All models (BERTs and GPT) were evaluated on a hold-out test set that contains 663 randomly selected trials, with 1 106 samples in total.

We evaluated NER models using precision, recall, and F_1 scores. The eligibility criteria are transformed into the BIO format: `B-entity` for the beginning of an entity span, `I-entity` for inside an entity span, and `O` for outside any entity span. This evaluation follows standard sequence-to-sequence learning metrics, but is challenging for generative LLMs. For example, for the criterion "*Histologically or cytologically confirmed diagnosis of gastric, lung, colorectal or breast cancer on file*", `gpt-4-turbo` generated "`[gastric cancer, lung cancer, breast cancer]`", a good extraction, but not aligned with the input sequence. Similarly, for "*cancer of the prostate*", it generated "`[prostate cancer]`", accurate yet misaligned answer according to BIO evaluation.

The problem of annotating complex, overlapping, and disjoint entity spans is typically addressed by Discontinuous Named Entity Recognition (DNER). DNER identifies and categorizes noncontiguous yet semantically linked entities, crucial in complex domains such as biomedicine, where entities such as symptoms or drug effects are often described in fragmented sentences. Unlike traditional NER, DNER labels entities across non-adjacent segments. Various methods address disjoint entity spans: some use relation extraction techniques to combine spans into disjoint, nested, or overlapping sequences [13]; others extend entity tags (BIO, IOBES) with H and D tags for shared and unshared parts of mentions (BIOHD); and some introduce uncertainty with FuzzyBIO labeling [6]. Recently, end-to-end neural models have also been proposed to discover discontinuous entities [4]. Unfortunately, existing methods only partially address the problem. They can annotate a span like "*lung or breast cancer*" as two entities (*lung cancer, breast cancer*), but do not handle LLM responses that do not align with the source text. The sequence tagging paradigm is not well suited for evaluating LLM responses. Although it is possible to prompt an LLM to generate output identical to the input, this is not a reliable solution. Switching to BIOHD evaluation would require costly reannotation of the entire dataset. In this paper, we use the traditional BIO labeling scheme, acknowledging that it underestimates the LLM's performance. The exact proportion of `gpt-4-turbo` predictions that are correct but misaligned remains unclear.

4. Experiments

For prompting, we select the `gpt-4-turbo` model [1]. We compare this LLM with the following BERT models: `BERT uncased` [5], `Biomedical BERT NER` [16], `BioBERT` [12], `SciBERT` [3], `PubMedBERT` [8], `BlueBERT` [14], `ClinicalBERT` [2], and `CODER` [20]. All layers of the BERT models were unfrozen. An additional linear layer was added on top for token classification. Early Stopping was used in training with patience set to 5. The number of epochs was set at 30, but all training processes were completed before the 15th epoch. The training arguments are as follows: the learning rate was set at $\eta = 1\text{e}{-5}$, the batch size $bs = 8$ for training and evaluation, and the weight decay was set at $\gamma = 0.01$. The learning rate scheduler used cosine with restarts, with 50 warm-up steps. Due to the lack of space, we present the results only for `CODER`, the results for other models are presented in the supplementary material.

Table 1 compares the `gpt-4-turbo` and `CODER` models. `CODER-27` represents the low-resource scenario, using eligibility criteria from only 27 clinical trials for fine-tuning, while `CODER-100` uses 100 trials. `gpt-4-turbo` is outperformed on every metric and BIO tag. Even a small fine-tuning data set yields significant benefits, highlighting the value of additional

Table 1. gpt-4-turbo vs. CODER on BIO NER (p-precision, r-recall, f-F_1 score

	gpt-4-turbo			CODER-27			CODER-100			
	p	r	f	p	r	f	p	r	f	support
B-CANCER	0,30	0,35	0,32	0,71	0,46	0,56	0,76	0,66	**0,71**	2093
I-CANCER	0,33	0,39	0,36	0,74	0,51	0,60	0,78	0,73	**0,75**	3073
B-TREATMENT	0,30	0,26	0,28	0,64	0,76	0,69	0,70	0,77	**0,73**	6209
I-TREATMENT	0,28	0,35	0,31	0,66	0,73	0,70	0,75	0,69	**0,72**	7332
B-CLINICAL_VARIABLE	0,32	0,47	0,38	0,73	0,72	0,72	0,84	0,68	**0,75**	2435
I-CLINICAL_VARIABLE	0,32	0,45	0,37	0,72	0,82	0,77	0,85	0,75	**0,80**	4770
B-ALLERGY_NAME	0,05	0,74	0,10	0,00	0,00	0,00	1,00	0,08	**0,14**	323
I-ALLERGY_NAME	0,02	0,35	0,03	0,00	0,00	0,00	0,89	0,06	**0,12**	265
B-CHRONIC_DISEASE	0,37	0,32	0,34	0,66	0,76	0,71	0,77	0,76	**0,76**	5115
I-CHRONIC_DISEASE	0,42	0,34	0,37	0,69	0,82	0,75	0,80	0,79	**0,80**	6247
micro avg	0,27	0,34	0,30	0,68	0,72	0,70	0,77	0,73	**0,75**	37862
macro avg	0,27	0,40	0,29	0,55	0,56	0,55	0,81	0,60	**0,63**	37862
weighted avg	0,33	0,34	0,33	0,67	0,72	0,69	0,78	0,73	**0,75**	37862

annotation despite the high cost of medical text annotation. A similar comparison for predicting entity spans without distinguishing between beginning and inside tokens (i.e. IO scheme) is presented in the supplement. The results mirror the BIO scheme, with CODER outperforming gpt-4-turbo, and more data leading to better performance.

5. Conclusions

In this paper, we demonstrate that a few-shot prompting is a viable solution when no fine-tuning data are available, but in the presence of even limited annotated data, BERT-based pre-trained models perform better, especially with fine-tuning. BERT-based models perform well in NER tasks due to their ability to capture relationships between closely placed tokens, aligning well with the BIO evaluation scheme. The eligibility criteria represent a specialized medical argot, which BERT-based models effectively handle. CODER excels because it is pre-trained on the Unified Medical Language System (UMLS) ontology, enhancing its recognition of specialized terms. In contrast, a broad understanding of the language of models such as gpt-4-turbo does not contribute as effectively to NER tasks in this specific domain.

It is important to note that our current prompting scheme relies on the model's ability to recognize hard-coded patterns. It is possible that gpt-4-turbo would perform better with more elaborate prompt engineering or a larger number of varying few-shot examples. In addition, the evaluation scheme does not align with the output of a generative language model and may underestimate the true efficiency of the model. Recent work on fine-tuning of GPT models for clinical trial analysis, such as the introduction of TrialGPT [10], or patient-trial matching using LLMs [19], clearly demonstrate the usefulness of LLMs in the analysis of medical information. However, the results presented in this paper show that smaller, domain-aligned and fine-tuned models are still a viable alternative to LLMs for difficult and linguistically narrow tasks, such as the extraction of medical argot.

Future work could include creating an ensemble of prompts and aggregating outputs, or transitioning from hard prompts (textual inputs) to soft prompts (dense numerical embeddings) [15]. Trainable soft prompts could potentially enhance the precision of information extraction from clinical trial protocols. Another interesting research question is the feasibility of retrieval-augmented models for NER in biomedical texts. Biomedical ontologies, such as the NCI Thesaurus [17] and SNOMED CT [7], provide structured vocabularies of entities and their relationships. Retrieval-augmented models [9] combine a parametric language model with a neural retriever that matches inputs with data from external ontologies or knowledge bases. Retrieval-augmented NER models could be a viable alternative to prompting large language models.

Acknowledgements

The research was carried out as part of the Polish Ministry of Education and Science's Industrial Doctorate Program implemented from 2020 to 2024 (Contract No. DWD/4/24/2020).

References

1. Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv:1904.03323 (2019)
3. Beltagy, I., *et al.*: Scibert: A pretrained language model for scientific text. arXiv:1903.10676 (2019)
4. Dai, X., Karimi, S., Hachey, B., Paris, C.: An effective transition-based model for discontinuous ner. arXiv:2004.13454 (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
6. Dirkson, A., Verberne, S., Kraaij, W.: Fuzzybio: A proposal for fuzzy representation of discontinuous entities. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. pp. 77–82 (2021)
7. Donnelly, K., et al.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics 121, pp. 279 (2006)
8. Gu, Y., *et al.*: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. on Computing for Healthcare 3(1), pp. 1–23 (2021)
9. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: Int. conf. on machine learning. pp. 3929–3938 (2020)
10. Jin, Q., Wang, Z., Floudas, C.S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J.: Matching patients to clinical trials with large language models. ArXiv (2023)
11. Kola, I., Landis, J.: Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery 3(8), pp. 711–716 (2004)
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), pp. 1234–1240 (2020)
13. Li, F., Lin, Z., Zhang, M., Ji, D.: A span-based model for joint overlapped and discontinuous named entity recognition. arXiv:2106.14373 (2021)
14. Peng, Y., *et al.*: Transfer learning in biomedical natural language processing. arXiv:1906.05474 (2019)
15. Qin, G., Eisner, J.: Learning how to ask: Querying lms with mixtures of soft prompts. arXiv:2104.06599 (2021)
16. Raza, S., Reji, D.J., Shajan, F., Bashir, S.R.: Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digital Health 1(12) (2022)
17. Sioutos, N., de Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics 40(1), pp. 30–43 (2007)
18. Tseo, Y., Salkola, M., Mohamed, A., Kumar, A., Abnousi, F.: Information extraction of clinical trial eligibility criteria. arXiv:2006.07296 (2020)
19. Yuan, J., Tang, R., Jiang, X., Hu, X.: Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. In: American Medical Informatics Association (AMIA) Annual Symposium (2023)
20. Yuan, Z., *et al.*: Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. Journal of biomedical informatics p. 103983 (2022)