Comparison of Traditional NER and Generative Large Language Models for PICO Extraction in Medical Research

Mark Bartos

Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam
m.bartos@student.vu.nl

Abstract. This study compares the performance of Long Short-Term Memory (LSTM) networks and Generative Pre-trained Transformers (GPT) for extracting and annotating PICO (Population, Intervention, Comparison, Outcome) elements from medical research papers. Using the EBM-NLP dataset and a secondary dataset of full-length medical papers, the BiLSTM-CRF model was trained on tokenized text sequences, while the GPT-4o model utilized prompt engineering. Results indicate that LSTM models excel at token-level annotation with a weighted F1-score of 0.76, whereas GPT-4o achieves a better recall of 0.89 and an F1-score of 0.82 at the abstract level. The findings suggest that while LSTM offers superior explainability and token-level precision, GPT-4o provides greater contextual understanding and flexibility via prompting. Future research should explore hybrid models and validate findings on a larger dataset of full-length papers.

Keywords: PICO elements · LSTM · GPT-4o · Named Entity Recognition · Medical Text Mining · Evidence-Based Medicine

1 Introduction

Evidence-based practice in medical research relies extensively on the systematic extraction and analysis of PICO (Population, Intervention, Comparison, Outcome) elements [7]. PICO elements form the backbone of systematic reviews, creating a link between the synthesis of clinical evidence and the development of guidelines that influence patient care [6]. Historically, rule-based approaches and traditional Natural Language Processing (NLP) techniques have been used to extract PICO elements [17]. One NLP technique is Named Entity Recognition (NER), for which Long Short-Term Memory (LSTM) networks are commonly used [8]. Recently, complex machine learning models have also shown potential in this extraction process, particularly Generative Pre-trained Transformers (GPTs) [9], including OpenAI's GPT-3.5 and GPT-4 models [4]. Comprehensive studies comparing these models' performance in the context of PICO element extraction from medical literature remain rare despite their potential. The aim of this paper is to investigate how traditional Natural Language Processing methods - specifically Named Entity Recognition - compare with Generative Large

Language Models in the performance and explainability of tagging PICO elements in medical research papers. This study strives to inform the development of more efficient automated PICO extraction tools by offering a thorough comparison, ultimately promoting the growth of evidence-based medical research. The paper structured as follows: Section 2 reviews related work, Section 3 outlines the methodology, Section 4 presents the results, and Section 5 discusses these findings and future research directions.

2 Related Work

The systematic extraction of PICO (Population, Intervention, Control, Outcome) elements from medical literature forms the cornerstone of modern evidence-based medicine (EBM) [1]. By systematically identifying and analyzing these elements, researchers and clinicians can synthesize data more effectively, leading to the formulation and answering of specific clinical questions through systematic reviews. The accuracy and efficiency of medical research are significantly improved by this methodology, which additionally aids in the development of guidelines and policies that more accurately represent the state of medical best practices today. Healthcare practitioners may gain valuable insights that are directly applicable to patient care through a comprehensive review of PICO aspects, improving treatment quality and patient outcomes [11].

Traditionally, Named Entity Recognition (NER), a subset of Natural Language Processing (NLP), has been employed to automate the extraction of PICO elements [8]. NER involves the identification and classification of specific phrases within texts. A prominent approach to NER is the use of a Bidirectional Long Short-Term Memory (BiLSTM) model combined with a Conditional Random Field (CRF) layer (BiLSTM-CRF) [8, 18]. This technique accurately models dependencies and transition constraints between expected tags by making use of the sequential pattern of text. By taking into account the context of nearby tags, CRF helps to improve predictions by preventing illogical tag sequences and lowering transition errors [3].

The emergence of Large Language Models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers) has significantly shifted the NLP landscape [5]. These LLMs take advantage of extensive pre-training on diverse corpora to develop a deep, contextual understanding of language. For tasks like PICO extraction, LLMs, including BERT have been fine-tuned to outperform traditional models in both accuracy and efficiency [21]. Furthermore, domain-specific adaptations of BERT, such as BioBERT, have demonstrated superior performance in analyzing systematic reviews by leveraging biomedical text pre-training [18].

With the introduction of sophisticated generative models like GPT-4 (Generative Pre-trained Transformer 4), medical applications of artificial intelligence

have advanced even further. Unlike traditional models, GPT-4 can be utilized effectively with minimal domain-specific training, relying instead on sophisticated prompt engineering. Although this represents a shift from classification-based approaches, such as LSTM, using GPT-4 for PICO extraction and classification can be an innovative approach in the field, due to its deeper understanding of context and coherent text [10, 4].

Prompt engineering acts as a crucial technique for utilizing the capabilities of generative large language models. This method involves crafting detailed, context-appropriate prompts that can aid the model in producing accurate and relevant outputs. By enhancing the prompt, researchers can significantly improve the performance of the model. Techniques such as writing clear and detailed instructions, specifying the output format, or providing a sheet of references to guide the model's reasoning can yield significantly better results [13]. Furthermore, the model's output can be further enhanced through the inclusion of example-based prompts, also referred to as "n-shot learning", which provide the model with tasks that are contextually similar to infer from [19].

3 Methodology

3.1 Research Design

This paper aims to adopt a comparative approach to evaluate the capabilities of gold-standard Named Entity Recognition (NER) methods and Generative Large Language Models (LLMs) in extracting and annotating PICO elements from medical research papers. Specifically, the traditional method employs a Bidirectional Long Short-Term Memory (BiLSTM) model with BIO tagging as a Conditional Random Field (CRF) layer, a gold-standard in Natural Language Processing for NER tasks. Although BERT, and its medicine-focused variants are now used widely in the domain, LSTM still maintains a higher explainability than BERT [20], and therefore has been preferred as a baseline for this experiment. Conversely, the advanced LLM approach utilizes OpenAI's latest GPT-4o model through its Python API, allowing for reproducible interactions rather than a conversational format, like the popular ChatGPT. These models will be evaluated not only in terms of their performance but also with regard to their explainability, as the latter is a crucial aspect in the medical domain. In section 3.1, the datasets used for this experiment will be detailed. In section 3.2 and 3.3 the two models will be discussed in detail, while in section 3.4 the steps of the evaluation process will be outlined.

3.2 Datasets

The research utilizes two datasets. The first dataset, EBM-NLP (Evidence Based Medicine Natural Language Processing) [2, 12], is a collection of abstracts, that will be used both for training and evaluation purposes. The second dataset, a manually annotated Full-Length Papers dataset, contains medical research

papers in their whole length. Its aim is to gain insight into what the models are capable of in near-real-life scenarios. Therefore, this dataset is only used as a secondary evaluation dataset.

EBM-NLP: The EBM-NLP dataset is a diverse corpus sourced from GitHub [2, 12], which contains abstracts of 4.993 medical research papers rigorously annotated for (P)articipants, (I)nterventions, and (O)utcomes elements. The range of abstracts are sourced from PubMed and serves as an ideal basis for training and testing due to its diverse nature in topics. Annotation of these abstracts was done by medical professionals, and their responses have been aggregated to reduce noise and potential bias. An example of a PICO annotated abstract can be seen in Table 1.

Each document in the dataset is identified by its PubMed identification number (PMID) and includes a raw text file, as well as a tokenized version of the abstract. The tokenized texts align precisely with the annotated labels, which helps maintain the word positions during training, predicting, and testing, keeping both models as explainable as possible.

For the documents in the database, two types of annotations are available: individual annotations, which include all labels from each annotator for documents that are annotated multiple times, providing insights into the variability and potential noise in human annotations. Additionally, aggregated annotations, which are derived from multiple annotations per document, consolidated through a voting mechanism to enhance reliability and reduce noise. The latter set is used for model training and evaluation.

Furthermore, both starting spans and hierarchical labels are available, the latter one allowing for more detailed classifications within each starting span, providing a nuanced understanding and extraction of PICO elements. The specific labels for each P/I/O element are thoroughly defined, ranging from demographic details (e.g., age, sex) to types of interventions (e.g., surgical, drug) and outcomes (e.g., physical outcomes, mortality) [1]. A full list of annotations with their corresponding labels are stated in Table 2 below, additionally, in Table 3 an example for each PICO label is mentioned.

-	**I-Drug**	-
The effect of	vitamin A-fortified coconut cooking oil	on the
P-Condition	-	**P-Age**
serum retinol concentration	of Filipino children	4-7 years old.

Table 1. Example of a PICO annotated abstract (PMID: 18376682)

Label	P	I	O
0	No label	No label	No label
1	Age	Surgical	Physical
2	Sex	Physical	Pain
3	Sample size	Drug	Mortality
4	Condition	Educational	Adverse effects
5		Psychological	Mental
6		Other	Other
7		Control	

Table 2. Label mappings for each PIO element

PICO Label	Example
P-Age	8-12
P-Sex	women
P-Sample size	six
P-Condition	mildly dehydrated state
I-Surgical	conventional esophageal resection
I-Physical	professional supragingival plaque removal
I-Drug	metronidazole
I-Educational	group therapy
I-Psychological	contingently responsive
I-Other	family history questionnaires
I-Control	placebo
O-Physical	fluid retention
O-Pain	post-thoracotomy pain
O-Mortality	in-hospital mortality
O-Adverse effects	cancer death
O-Mental	drug abuse
O-Other	neuropsychological tasks

Table 3. An example for each PICO element

Full-Length Papers: The Full-Length Papers dataset includes 5 papers annotated in their full-length, in contrast to the EBM-NLP dataset, which only includes the abstracts of medical research papers. The five papers were randomly selected from the EBM-NLP's test set, to maintain consistency. Unfortunately, no similar dataset was publicly available at the time, and therefore these papers were annotated by the author, a non medical professional, and therefore, these PICO labels might be imperfect. The annotations can be found in the supplementary material for transparency.

3.3 Models

Traditional NER Model: A BiLSTM model is designed to capture contextual dependencies from both the previous and the subsequent tokens in the text. Additionally, a CRF layer was applied utilizing a BIO (Beginning, Inside, Outside) tagging scheme to determine the boundaries of classification within the text.

The BiLSTM-CRF model for this experiment was constructed using Google's TensorFlow, chosen for its robust support for deep learning tasks and its good reputation within the scientific community. In its input layer, the model accepted the tokenized sequences of the abstracts. The embedding layer converted these tokens into a fixed-size vectors, utilizing word representations for the model. The Bidirectional LSTM layer processed the sequences in both forward and backward directions, to capture contextual information from both subsequent and prior tokens respectively. The next layer of the model, the dropout layer applied dropout regularization to prevent overfitting. A time distributed dense layer was also applied at each time step, to output a sequence of predictions. Finally, a CRF layer ensured that the predicted labels were from valid sequences, improving the overall accuracy. This architecture of the model is illustrated in Figure 1.

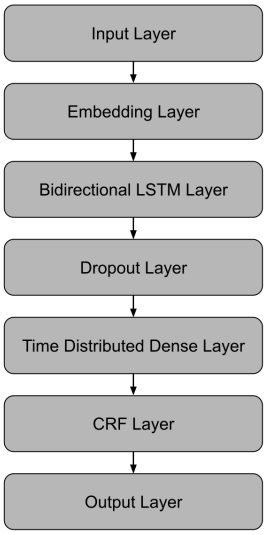

Fig. 1. Architecture of the BiLSTM-CRF model

In order to encode the data for model training, each token in the sentences was mapped to an integer index, generating a numerical representation of the

text. In a comparable manner, PICO labels were encoded as integers to initialize their use in the neural network. A maximum of 500 tokens per sequence was used for padding to guarantee consistent sequence lengths throughout the dataset.

After encoding both the X and Y values, the model was trained using the Adam optimizer and a sparse categorical cross-entropy loss function. The training included a train-test split, ensuring the model was evaluated on previously unseen data, and it was trained in batches of 32 samples, iterating over the dataset in 10 epochs. A validation split of 10% was employed to monitor the model's performance and prevent overfitting.

When training was completed, the model was used to predict PICO labels for the test split of the EBM-NLP dataset. The predicted labels were decoded from their integer-encoded representations back to their original PICO annotations.

Generative LLM model: GPT-4o, OpenAI's latest model released in May 2024, has the capability to accept input as combination of text, audio, image, and video and generates any combination of text, audio, and image as outputs based on user given prompts [14, 15]. However, it has been preferred over GPT-4 due to its main advantage of generating responses faster, being 50% more cost-effective while still matching the performance of GPT-4.

In contrary to the more traditional BiLSTM-CRF model, which requires training on the training set of the data, this model is utilized by crafting a prompt to extract PICO elements from a given text input. According to OpenAI's Prompt Engineering Guidelines [13, 15], the ideal prompt begins by simply denoting the task of the assistant. As a further criterion, and for the model to be able to understand the data structure, the separator values of the dataset were specified. To allow for comparability, and to improve the quality of the responses, a set of possible PICO labels was given to the model. These PICO labels correspond to the labels used by the EBM-NLP abstracts dataset mentioned above, in Table 2.

> "You are a PICO annotator tasked with labeling tokens. Each token is separated by a comma (,). Provide a label for every token from the following options: 0 for none, or P-Age, P-Sex, P-Sample size, P-Condition, I-Surgical, I-Physical, I-Drug, I-Educational, I-Psychological, I-Other, I-Control, O-Physical, O-Pain, O-Mortality, O-Adverse effects, O-Mental, O-Other."

To further improve response quality, a 1-shot learning approach was applied, providing the model with one example input and output. This example was randomly selected from the training set of the EBM-NLP database and adjusted to 75 tokens. This significantly improved the responses of the model, as it helped the model output the correct format, correctly separated by commas, and use the correct PICO labels, including the null value. Testing during development

reveled that a 0-shot approach produced highly unreliable output. Predictions were not in the desired format - e.g. values not separated by commas, which made the output hard to process further in the pipeline -, or the model returned only a short string consisting only of PICO labels, while skipping over "No Label" values, without indicating which tokens were skipped. On the contrary, a 2-shot or higher approach consumed unnecessary resources, making the model slower, and less cost-effective, as 2-shots or more did not improved the output quality further. The example input and output that was given to the model can be seen below:

> "Randomized , double-blind , placebo-controlled trial of oral sirolimus for restenosis prevention in patients with in-stent restenosis : the Oral Sirolimus to Inhibit Recurrent In-stent Stenosis (OSIRIS) trial . BACKGROUND Despite recent advances in interventional cardiology , including the introduction of drug-eluting stents for de novo coronary lesions , the treatment of in-stent restenosis (ISR) remains a challenging clinical issue . Given the efficacy of systemic sirolimus administration to prevent neointimal"
>
> "0 0 0 0 I-Control 0 0 0 0 0 O-Physical 0 0 0 0 P-Condition P-Condition 0 0 0 I-Drug 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 O-Physical"

To further ensure that all tokens were annotated, an extension was added to the prompt:

> "You are a PICO annotator. Each token is separated by a comma (,). Provide a label for every token from the following options: [...]. **It is crucial that you return exactly one label for each token, ensuring that the number of labels matches the number of tokens exactly.**"

However, this approach still had limitations, with GPT only annotating approximately the first 200 tokens before stopping. Using different n-shot examples and prompts did not seem to fix the issue. Since using variants of GPT-4 (GPT-4o, GPT-4 and GPT-4-Turbo) appeared to annotate slightly more tokens, while variant of GPT-3.5 (GPT-3.5, GPT-3.5-Turbo) quit earlier, this issue was denoted to the limitation of the model. To circumvent this, the textual inputs were split up into 75-token chunks, feeding them into the model separately, while maintaining the character positions per paper. The prompt was adjusted accordingly:

> "You are a PICO annotator tasked with labeling tokens. **For every chunk of 75 tokens** separated by commas (,), **return 75 PICO annotations**. Use one of the following options: [...]. It is crucial that you return **exactly 75 labels for each chunk**, ensuring that the number of labels matches the number of tokens exactly."

This approach resulted in the lowest bias per abstract, with occasional deviations of only ± 0-5 tokens from the chunk size. To prevent shifts in word positions, if an output was longer than number of tokens inputted, the end of the output was cut off. In a similar manner, to prevent missing values, for outputs shorter than the chunk size, tokens were extended with empty labels (0).

To conclude the experiment, the model was asked to predict a PICO label for each token in the EBM-NLP dataset's test set. The predictions were organized into a dataframe, similar to the gold data and the predictions of BiLSTM-CRF model, to make comparisons straightforward.

3.4 Evaluation

After the traditional NLP model, BiLSTM-CRF was trained on the training subset of the dataset, it made predictions on the test subset of the EBM-NLP dataset, and consequently, on the Full-Length Paper dataset. Conversely, the LLM model, GPT-4o requiring no traditional training, also made its predictions on the same test sets based on the engineered prompt, given in Section 3.3 above. Both models were tasked with predicting a suitable PICO label for each token in a test set containing 100 abstracts, and 5 full-length papers.

Two evaluation tactics were used in combination with each other, to obtain the most relevant assessment of both models' performance. First, a quantitative evaluation was performed to act as a baseline, making the two models comparable to each other. Additionally, a qualitative evaluation was performed by the author to gain a deeper understanding of the models' performance by observing regular patterns they exhibited.

These tactics were applied to the predictions made on the test set of the EBM-NLP dataset, which consists of abstracts only. Additionally, to evaluate the models in a real-life scenario, the same evaluation tactics were performed on five full-length evidence-based medical research papers from PubMed. These papers were annotated manually by the author, a non-medical professional, and might contain slight biases.

Quantitative Evaluation: Standard machine learning evaluation methods were utilized through the calculation of performance metrics, based on the confusion matrix of each experiment. The examined metrics include:

- *Accuracy*: Measures the overall correctness of the model's predictions, which is crucial for understanding the general reliability of the model.

- *Precision*: Indicates the proportion of true positive predictions among all positive predictions, which is important for minimizing false positives in medical applications.
- *Recall*: Reflects the model's ability to identify all relevant instances, ensuring that no critical information is missed.
- *F1-Score*: Provides a balance between precision and recall, offering a single metric for model comparison that accounts for both false positives and false negatives.

These metrics were examined both at the abstract level and at the token level for the EBM-NLP dataset. An abstract-level evaluation check if the model was able to find the gold PICO label anywhere in the abstract, and therefore uses a broader criteria, reflecting on the general understanding of context and thematic relevance of the text (illustrated in Table 4). In use cases where no word-position identification is required - e.g. searching based on available PICOs - an abstract level evaluation provides a more accurate measure. Conversely, a token-level evaluation proposes a more strict criteria, requiring the model to precisely identify and label individual tokens where the relevant information can be located (illustrated in Table 5). This level of precision is particularly useful in automated information extraction pipelines, where finding the precise location of the token is crucial for subsequent processing steps without requiring human input.

I-Drug, P-Condition, P-Age
The effect of vitamin A-fortified coconut cooking oil on the serum retinol concentration of Filipino children 4-7 years old.

Table 4. Example of the abstract-level evaluation technique (PMID: 18376682)

-	**-**	**-**	**I-Drug**	**I-Drug**	**I-Drug**	**I-Drug**	**I-Drug**	-
The	effect	of	vitamin	A-fortified	coconut	cooking	oil	on
-	**P-Condition**	**P-Condition**	**P-Condition**	**-**	**-**	**-**	**P-Age**	-
the	serum	retinol	concentration	of	Filipino	children	4-7 years old	.

Table 5. Example of the token-level evaluation technique (PMID: 18376682)

Qualitative Evaluation: To gain a deeper insight beyond numerical metrics, the predictions of both models were manually assessed. Twenty-five random

abstracts were selected from the EBM-NLP dataset's test set and the predictions of BiLSTM-CRF and GPT-4o were manually compared to each other, as well as to the gold annotations. Criteria for this evaluation included identifying regular false negative or false positive patterns occurring for certain PICO labels. This manual review helped identify frequently occurring patterns in the models' labeling behaviors.

4 Results

4.1 Quantitative Evaluation

A comprehensive overview of the key metrics — accuracy, precision, recall, and F1-score — across different datasets and evaluation levels is summarized in Table 6 and Table 7, which report the weighted and macro averages, respectively.

Since in medical texts, there is a natural class imbalance - mostly containing "No Label" values - a weighted average gives a good overall measure, providing each class a weight proportional to its occurrence. On the other hand, macro-average gives each class equal consideration. If the goal is to examine the models via the strictest criteria, and check whether all PICO labels are found with the highest precision possible, the macro-average can provide that.

Table 6. Performance of GPT-4 and LSTM on EBM-NLP Database and Full-Length Papers (calculated with weighted average)

Model	**Dataset**	**Level**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	EBM-NLP	Token	0.77	0.76	0.77	0.76
GPT-4o	EBM-NLP	Token	0.67	0.67	0.67	0.67
LSTM	EBM-NLP	Abstract	0.00	0.50	0.51	0.48
GPT-4o	EBM-NLP	Abstract	0.00	0.79	0.89	0.82
LSTM	Full-Length Papers	Token	0.67	0.86	0.67	0.75
GPT-4o	Full-Length Papers	Token	0.75	0.70	0.75	0.73

Table 7. Performance of GPT-4 and LSTM on EBM-NLP Database and Full-Length Papers (calculated with macro average)

Model	**Dataset**	**Level**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	EBM-NLP	Token	0.77	0.10	0.11	0.10
GPT-4o	EBM-NLP	Token	0.67	0.06	0.06	0.06
LSTM	EBM-NLP	Abstract	0.00	0.29	0.32	0.28
GPT-4o	EBM-NLP	Abstract	0.00	0.65	0.82	0.71
LSTM	Full-Length Papers	Token	0.67	0.05	0.04	0.04
GPT-4o	Full-Length Papers	Token	0.75	0.09	0.09	0.09

The detailed performance metrics, that illustrates performance per PICO label, both for the traditional BiLSTM-CRF model and for the GPT-4o large language model are presented in Table 8 & 9 & 13 and Table 10 & 11 & 12 respectively.

4.2 Qualitative Evaluation

A manual examination of the predicted labels across twenty-five randomly selected abstracts revealed several patters of repetitions. The most frequently encountered error was the failure of both BiLSTM-CRF and GPT-4o to assign any labels to tokens that were annotated according to the gold data. Instances involving "P-Age" and "P-Sample size" labels were rarely identified precisely.

Occasionally, either GPT-4o, BiLSTM-CRF, or both models were able to successfully identify a token, but the model(s) misclassified it (e.g. labeling *"vomiting"* as P-Condition by GPT, while it was O-Adverse Effects according to the gold label, or *"body mass index"* as I-Drug by LSTM, P-Condition by GPT, while it was P-Condition in gold label). This occurred in a total of 122 times for BiLSTM-CRF, and 192 times for GPT-4o.

In other instances, both models occasionally assigned labels to medical terms that were not labeled in the gold-standard dataset. Examples include *"chemotherapy"* being labeled as I-Drug by GPT, *"formula"* as I-Drug by LSTM or *"surgical technique"* as I-Surgical by GPT. For BiLSTM-CRF, this happened 32 times overall, and for GPT-4o, it happened 47 times.

Generally, the GPT-4o model demonstrated a tendency to predict labels for individual tokens, whereas the BiLSTM-CRF model was more adept at predicting chains of labels for sequence of tokens. Consequently, BiLSTM-CRF also periodically recognized causative phrases, such as *"indicated that"*, *"agreement between both methods remained good"*, or *"information is not available for many early"*, although these were often mislabeled with Patient or Intervention tags. BiLSTM-CRF presented this pattern 20 times.

When mislabeling tokens, the GPT model exhibited a bias against stop words (e.g. *"with"*, *"of"*, and *"the"*), whereas the LSTM model repeatedly mislabeled tokens containing verbs (e.g. *"obtained"*, *"indicates"*, and *"avoid"*). This occurred 48 times in total for GPT-4o and 45 times for BiLSTM-CRF. The false positive labels for these misclassifications varied, and no strong pattern was recognised. However, the BiLSTM-CRF model showed a slight bias towards I-Drug labels (110 times), and GPT towards P-Condition (92 times).

5 Discussion

The GPT-4o large language model and the traditional BiLSTM-CRF model were compared for PICO extraction from medical research papers abstracts and from

full-length research papers. This assessment generated significant insight into the benefits and drawbacks of each model. The subsections below analyze the performance of the models both at the token and at the abstract levels to shows the advantages and drawbacks of each model. Additionally, the flexibility and explainability of each approach will be explored, and directions for future work will be suggested.

5.1 Overall Performance

Recall is a crucial metric in this paper's context, since it measures the model's ability to identify all relevant instances of PICO elements, ensuring that no critical information is missed. This is important in medical contexts, where missing key information can lead to incomplete or inaccurate evidence synthesis. Emphasizing recall is also beneficial for this evaluation, since both LSTM and GPT may recognise more neighbouring tokens due to their sequence dependent and LLM nature. Recognising more PICO labels around the gold-label (to a healthy degree) is not considered an error when annotating PICOs. Additionally, the F1-score provides a harmonic mean between precision and recall, providing a comprehensive view of a model's performance. High F1-scores indicate that the model is not only capturing most of the relevant elements (high recall) but is also doing so with minimal false positives (high precision).

Overall, when examining the weighted-average of the EBM-NLP dataset, BiLSTM-CRF reached a recall of 0.77 and an F1-score of 0.76, outperforming GPT-4o at the token-level, which was only able to reach a recall and F1-score of 0.67. Conversely, GPT-4o reached a higher recall and precision of 0.89 and 0.82 respectively, when examined at the abstract-level, outperforming BiLSTM-CRF, which achieved only 0.51 and 0.48. For the Full-Length Papers, which was evaluated only at the token-label, BiLSTM-CRF reached a lower recall of 0.67, than GPT-4o did with 0.75, but it reached an ever-so-slightly higher F1-score of 0.75, than GPT-4o's 0.73. However, when examining the macro-average metrics, where all classes of PICO labels and the class of "No Label" are treated equally, both models perform poorly, only able to reach a recall or an F1-score of 0.10 at the abstract-levels. There, GPT-4o outperformed the BiLSTM-CRF model as expected by the weighted-average results, being able to reach a recall of 0.82, and a F1-score of 0.71. In all abstract-level evaluations, the accuracy could not be calculated, and therefore is 0.00.

In contrast to the research conducted on the EBM-NLP dataset, where the authors of the dataset achieved an F1-score ranging between 90 and 91 using a NER model implemented with TensorFlow, incorporating a combination of LSTM, CRF, and character embeddings [2, 12], the model discussed in this paper, which employed a similar Bidirectional Long Short-Term Memory architecture, reached a moderately lower F1-score of 0.77. It is important to note, however, that the evaluation methodology in their research was not clearly detailed, and

no additional performance metrics were provided, as the primary focus of the paper was on the development and curation of the database.

Performance at abstract level At the abstract level, the GPT-4o model significantly outperformed the BiLSTM-CRF model, achieving higher accuracy, precision, recall, and F1-score. The improvements were 0.02, 0.36, 0.50, and 0.43 for weighted-average scores, and 0.02, 0.29, 0.38, and 0.34 for macro-average scores, respectively. (See Table 6 and 7) This contrasting difference suggests that GPT-4o offers significant advantages when interpreting a larger, document-level context due to its contextual understanding and capacity to leverage its large-scale pre-trained knowledge. The GPT-4o model's capability to consider the entire context of an abstract allows it to make more informed predictions about PICO elements, even when specific tokens are not explicitly tagged. This ability is particularly useful, in scenarios where automated extraction is not performed, but rather for retrieving papers via specific PICO tags during a search, or when manual human review is expected.

Performance at token level Conversely, the BiLSTM-CRF model showed a slight advantage over GPT-4o at the token level, achieving higher accuracy, precision, recall, and F1-score. The improvements were 0.10, 0.09, 0.10, and 0.09 for weighted-average scores, and 0.10, 0.04, 0.05, and 0.04 for macro-average scores, respectively. (See Table 6 and 7) This result illustrates how effective the BiLSTM-CRF model is at handling the classification of individual tokens. This likely occurs due to the models robust training on the particular PICO tagging task and explicit design to manage sequence dependencies. The use of a Conditional Random Field (CRF) layer in the BiLSTM-CRF model helps enforce logical label sequences, thereby resulting in fewer transition error and an improved overall token-level performance. This capability is particularly beneficial in automated machine learning pipelines, where precise token identification is critical for subsequent autonomous extraction and application in further algorithms.

However, these results change when we examine the Full-Length dataset at the token level. Here, GPT-4o outperforms BiLSTM-CRF both in terms of accuracy and recall, indicating its strong ability to identify relevant PICO tokens. This suggests that GPT-4o's contextual understanding allows for accurate labeling at the token-level as well, when provided with the broad context of full-length papers. Nevertheless, the BiLSTM-CRF model's higher precision significantly boosts its F1-score, enabling it to outperform GPT-4o in these metric. This benefit is attributed again to BiLSTM-CRF's vigorous sequence modeling capabilities, which can efficiently manage dependencies between tokens and maintain logical label sequences, ensuring high precision in token-level classifications.

5.2 Flexibility and Explainability

Flexibility and explainability are crucial factors in the adoption of NLP models in medical applications. The traditional BiLSTM-CRF model offers high explainability due to its structured approach and transparency of its sequential tagging process. This makes it easier for medical professionals to understand and trust the model's prediction. GPT-4o offers flexibility through prompt engineering as one of its main benefits. With well-constructed prompts, GPT-4o can be refined, as opposed to traditional models that require notable retraining in order to adjust to new tasks or datasets. This flexibility is particularly useful in dynamic research settings where needs can change rapidly. For instance, altering the prompt can help GPT-4o extract different PICO labels without the need for re-training. However, LLMs pose explainability issues - due to their "black-box" nature - which are hard to overcome.

5.3 Statistical Analysis

A Wilcoxon signed-rank test was carried out to test if there is a significant difference between the two models. This test is able to compare the micro/macro averages of performance metrics of multi-class classification models. [16] When comparing the accuracy scores of GPT-4o and BiLSTM-CRF, there was no statistically significant difference found between the models, based on the p-value of 0.3916. Comparing the F1-scores revealed a p-value of 0.3916, thus the difference was once again not statistically significant. Unfortunately, a thorough statistical analysis could not be carried out correctly. The Large Language Model, GPT, would produce a suitable statistical distribution if it was run again; however the Long Short-Term Memory model, BiLSTM-CRF, would always generate the same metrics because of its mathematical nature. As a result, each model only has one performance metric available, and this lack of data renders the results of the Wilcoxon signed-rank test insignificant.

5.4 Future Work

While the performance and capabilities of GPT-4o and the BiLSTM-CRF model for PICO extraction have been highlighted in our current analysis, there are a number of potential directions for future research that could improve and hone these techniques even further.

Layered Algorithmic Method: The creation of a layered algorithmic approach is one of the interesting possibilities. A two-stage LLM could potentially be used for this, with the first layer being entirely focused on deciding if a token is a PICO element. The specific PICO element that each token represents would then be categorized by a subsequent layer. By taking advantage of sequential modeling and contextual knowledge, this layered approach may overcome the issue of GPT-4o labeling tokens as false negative or false positive.

Integration of Hybrid Models: Another appealing method is to integrate a hybrid model in which the token-level predictions of the BiLSTM-CRF are fed into the GPT-4o model, alongside the tokens themselves. Through the use of BiLSTM-CRF's powerful sequence modeling capabilities, this hybrid model would make it possible to cross-validate predictions via GPT-4o's contextual understanding. This way, we could take advantage of the benefits of both models, and potentially achieve the highest performance metrics.

Evaluation on an Extensive Full-Length Dataset: It is essential that these suggested methods undergo evaluation on a dataset with at least 1.000 medical research papers annotated in their full-length in order to provide an accurate basis for their performance. A dataset of this kind would guarantee that the models are examined against a wide variety of medical literature, and therefore offer a better knowledge into their performance in actual situations.

6 Conclusion

In conclusion, this comparison of Natural Language Processing methods, specifically a BiLSTM-CRF model, and Generative Large Language Models, specifically GPT-4o, for the extraction of PICO elements from medical research publications revealed distinctive advantages and disadvantages of each method. When it comes to labeling abstracts, the LSTM model has advantages in explainability and precise token-level annotation, which makes it an excellent match for applications where the specific location of each token is important. However, its flexibility is limited by the requirement to retrain when datasets or labeling requirements change.

On the other hand, GPT-4o performs better in situations where a more comprehensive understanding of the whole papers is required, as it is particularly good at handling abstract-level context, or labeling full papers token-wise. Its adaptability to change tasks without requiring significant retraining is made possible by its underlying prompt engineering based approach. Therefore, GPT-4o is the best option for dynamic environments where requirements could change regularly. However, the "black-box" nature of the model presents significant barriers to comprehensibility, which is crucial in the medical field where transparency and trust in model predictions are necessary.

Future research should investigate hybrid models, which may potentially improve performance at both the token and the abstract levels by fusing the advantages of GPT-4o's contextual understanding with LSTM's sequence modeling. Layering LLMs, where one model is identifying PICO elements and then the other one is labeling them could increase GPT-4o's performance, while keeping the benefit of its flexibility and contextual awareness. Additionally, extending the evaluation to a full-length dataset would offer a deeper understanding of the performance of these models in real medical research situations. By addressing

such recommendations, medical text annotation systems can be made more robust and flexible, thus improve the application and synthesis of clinical research findings.

7 Appendix

	precision	recall	f1-score	support
0	1.00	1.00	1.00	99
I-Control	0.00	0.00	0.00	36
I-Drug	0.66	0.63	0.65	65
I-Educational	0.20	0.08	0.12	12
I-Other	0.00	0.00	0.00	15
I-Physical	0.14	0.05	0.08	19
I-Psychological	0.04	1.00	0.08	4
I-Surgical	0.00	0.00	0.00	16
O-Adverse effects	0.00	0.00	0.00	29
O-Mental	0.11	0.06	0.07	18
O-Mortality	0.00	0.00	0.00	11
O-Other	0.33	0.02	0.04	53
O-Pain	0.00	0.00	0.00	5
O-Physical	0.82	0.18	0.29	79
P-Age	0.00	0.00	0.00	30
P-Condition	0.00	0.00	0.00	91
P-Sample size	0.00	0.00	0.00	84
P-Sex	0.00	0.00	0.00	20
micro avg	0.54	0.24	0.33	686
macro avg	0.18	0.17	0.13	686
weighted avg	0.34	0.24	0.25	686
samples avg	0.54	0.24	0.33	686

Table 8. Abstract-Level Performance Metrics for the LSTM Model, Evaluated on Abstracts

	precision	recall	f1-score	support
O-Mortality	0.00	0.00	0.00	90
I-Control	0.00	0.00	0.00	1748
P-Condition	0.04	0.01	0.02	676
0	0.80	0.75	0.77	22846
O-Physical	0.00	0.00	0.00	25
P-Age	0.00	0.00	0.00	100
I-Surgical	0.00	0.00	0.00	201
I-Educational	0.00	0.00	0.00	295
O-Mental	0.02	0.01	0.01	341
O-Pain	0.00	0.00	0.00	120
P-Sex	0.00	0.00	0.00	49
O-Other	0.00	0.00	0.00	555
I-Physical	0.00	0.02	0.01	253
O-Adverse effects	0.00	0.00	0.00	144
P-Sample size	0.00	0.00	0.00	161
I-Other	0.00	0.00	0.00	120
I-Drug	0.03	0.11	0.04	876
I-Psychological	0.93	1.00	0.96	20900
accuracy		0.77		49500
macro avg	0.10	0.11	0.10	49500
weighted avg	0.76	0.77	0.76	49500

Table 9. Token-Level Performance Metrics for the LSTM Model, Evaluated on Abstracts

	precision	recall	f1-score	support
0	1.00	1.00	1.00	99
I-Control	0.58	0.97	0.73	36
I-Drug	0.89	0.91	0.90	65
I-Educational	0.64	0.75	0.69	12
I-Other	0.17	0.67	0.27	15
I-Physical	0.61	0.58	0.59	19
I-Psychological	0.17	0.25	0.20	4
I-Surgical	0.50	0.69	0.58	16
O-Adverse effects	0.68	0.90	0.78	29
O-Mental	0.65	0.72	0.68	18
O-Mortality	0.69	1.00	0.81	11
O-Other	0.56	0.91	0.69	53
O-Pain	0.56	1.00	0.71	5
O-Physical	0.89	0.65	0.75	79
P-Age	0.76	0.87	0.81	30
P-Condition	0.92	1.00	0.96	91
P-Sample size	0.86	1.00	0.92	84
P-Sex	0.59	1.00	0.74	20
micro avg	0.73	0.89	0.80	686
macro avg	0.65	0.82	0.71	686
weighted avg	0.79	0.89	0.82	686
samples avg	0.74	0.89	0.80	686

Table 10. Abstract-Level Performance Metrics for the GPT-4o Model, Evaluated on Abstracts

	precision	recall	f1-score	support
0	0.80	0.84	0.82	22917
I-Control	0.02	0.03	0.02	120
I-Drug	0.04	0.03	0.03	892
I-Educational	0.01	0.00	0.01	295
I-Other	0.00	0.00	0.00	120
I-Physical	0.03	0.01	0.01	253
I-Psychological	0.00	0.00	0.00	33
I-Surgical	0.02	0.01	0.02	201
O-Adverse effects	0.00	0.00	0.00	144
O-Mental	0.00	0.00	0.00	341
O-Mortality	0.01	0.02	0.02	90
O-Other	0.02	0.02	0.02	562
O-Pain	0.00	0.00	0.00	25
O-Physical	0.04	0.01	0.01	1756
P-Age	0.00	0.00	0.00	100
P-Condition	0.03	0.04	0.03	684
P-Sample size	0.01	0.02	0.01	161
P-Sex	0.00	0.00	0.00	49
micro avg	0.68	0.67	0.67	28743
macro avg	0.06	0.06	0.06	28743
weighted avg	0.64	0.67	0.66	28743
samples avg	0.67	0.67	0.67	28743

Table 11. Token-Level Performance Metrics for the GPT-4o Model, Evaluated on Abstracts

	precision	recall	f1-score	support
0	0.83	0.90	0.87	10869
I-Drug	0.02	0.03	0.03	30
I-Educational	0.00	0.00	0.00	86
I-Other	0.02	0.01	0.01	170
I-Physical	0.00	0.00	0.00	98
I-Psychological	0.05	0.01	0.01	119
I-Surgical	0.04	0.03	0.03	120
O-Adverse effects	0.08	0.07	0.08	83
O-Mental	0.09	0.11	0.10	149
O-Morality	0.00	0.00	0.00	5
O-Other	0.04	0.02	0.03	256
O-Physical	0.13	0.04	0.06	603
P-Age	0.04	0.03	0.03	40
P-Condition	0.08	0.07	0.08	417
P-Sample size	0.08	0.06	0.07	99
P-Sex	0.00	0.00	0.00	16
micro avg	0.76	0.75	0.76	13160
macro avg	0.09	0.09	0.09	13160
weighted avg	0.70	0.75	0.73	13160
samples avg	0.75	0.75	0.75	13160

Table 12. Token-Level Performance Metrics for the GPT-4o Model, Evaluated on Full-Length Papers

	precision	recall	f1-score	support
O-Other	0.00	0.00	0.00	0
I-Educational	0.00	0.00	0.00	13
I-Psychological	0.00	0.00	0.00	46
P-Sample size	0.00	0.00	0.00	0
O-Mental	0.00	0.00	0.00	0
0	0.92	0.72	0.81	1394
P-Condition	0.00	0.00	0.00	0
I-Physical	0.00	0.00	0.00	0
O-Mortality	0.00	0.00	0.00	15
I-Control	0.00	0.00	0.00	0
P-Age	0.00	0.00	0.00	1
I-Drug	0.00	0.00	0.00	0
P-Sex	0.00	0.00	0.00	0
O-Pain	0.00	0.00	0.00	15
O-Adverse effects	0.00	0.00	0.00	0
O-Physical	0.00	0.00	0.00	16
I-Surgical	0.00	0.00	0.00	0
I-Other	0.00	0.00	0.00	0
micro avg	0.67	0.67	0.67	1500
macro avg	0.05	0.04	0.04	1500
weighted avg	0.86	0.67	0.75	1500

Table 13. Token-Level Performance Metrics for the LSTM Model, Evaluated on Full-Length Papers

References

1. Cochrane linked data: Pico ontology (2024), https://linkeddata.cochrane.org/pico-ontology
2. bepnye: Evidence based medicine - natural language processing, https://github.com/bepnye/EBM-NLP/tree/master
3. Brockmeier, A.J., Ju, M., Przybyła, P., Ananiadou, S.: Improving reference prioritisation with pico recognition. BMC medical informatics and decision making **19**, 1–14 (2019)
4. Demir, G.B., Süküt, Y., Duran, G.S., Topsakal, K.G., Görgülü, S.: Enhancing systematic reviews in orthodontics: a comparative examination of gpt-3.5 and gpt-4 for generating pico-based queries with tailored prompts and configurations. European Journal of Orthodontics **46**(2), cjae011 (2024)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Eriksen, M.B., Frandsen, T.F.: The impact of patient, intervention, comparison, outcome (pico) as a search strategy tool on literature search quality: a systematic review. Journal of the Medical Library Association: JMLA **106**(4), 420 (2018)
7. Higgins, J.P., Green, S.: Cochrane handbook for systematic reviews of interventions. The Cochrane database of systematic reviews (2008)
8. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
9. Li, L., Zhou, J., Gao, Z., Hua, W., Fan, L., Yu, H., Hagen, L., Zhang, Y., Assimes, T.L., Hemphill, L.: A scoping review of using large language models (llms) to investigate electronic health records (ehrs). arXiv preprint arXiv:2405.03066 (2024)
10. Muftić, F., Kadunić, M., Mušinbegović, A., Abd Almisreb, A.: Exploring medical breakthroughs: a systematic review of chatgpt applications in healthcare. Southeast Europe Journal of Soft Computing **12**(1), 13–41 (2023)
11. Nowak, A.J.: Artificial intelligence in evidence-based medicine. Artificial Intelligence in Medicine (2021)
12. Nye, B., Jessy Li, J., Patel, R., Yang, Y., Marshall, I.J., Nenkova, A., Wallace, B.C.: A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Proc Conf Assoc Comput Linguist Meet **2018**, 197–207 (Jul 2018)
13. OpenAI: Guides: Prompt engineering, https://platform.openai.com/docs/guides/prompt-engineering
14. OpenAI: Hello gpt-4o, https://openai.com/index/hello-gpt-4o/
15. OpenAI: Introduction to gpt-4o (2024), https://cookbook.openai.com/examples/gpt4o/introduction_to_gpt4o
16. Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Scientific Reports **14**(1), 6086 (2024)
17. Schmidt, L., Mutlu, A.N.F., Elmore, R., Olorisade, B.K., Thomas, J., Higgins, J.P.: Data extraction methods for systematic review (semi) automation: Update of a living systematic review. F1000Research **10** (2021)
18. Wang, Q., Liao, J., Lapata, M., Macleod, M.: Pico entity extraction for preclinical animal literature. Systematic Reviews **11**(1), 209 (2022). https://doi.org/10.1186/s13643-022-02074-4, https://doi.org/10.1186/s13643-022-02074-4

19. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)
20. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable ai: A brief survey on history, research areas, approaches and challenges. Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8 pp. 563–574 (2019)
21. Zhang, T., Yu, Y., Mei, J., Tang, Z., Zhang, X., Li, S.: Unlocking the power of deep pico extraction: Step-wise medical ner identification. arXiv preprint arXiv:2005.06601 (2020)