npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01673-4

Enhancing biomarker based oncology trial matching using large language models

Nour Alkhoury, Maqsood Shaik, Ricardo Wurmus & Altuna Akalin ✉

Clinical trials are an essential component of drug development for new cancer treatments, yet the information required to determine a patient's eligibility for enrollment is scattered in large amounts of unstructured text. Genomic biomarkers are especially important in precision medicine and targeted therapies, making them essential for matching patients to appropriate trials. Large language models (LLMs) offer a promising solution for extracting this information from clinical trial study descriptions (e.g., brief summary, eligibility criteria), aiding in identifying suitable patient matches in downstream applications. In this study, we explore various strategies for extracting genetic biomarkers from oncology trials. Therefore, our focus is on structuring unstructured clinical trial data, not processing individual patient records. Our results show that open-source language models, when applied out-of-the-box, effectively capture complex logical expressions and structure genomic biomarkers, outperforming closed-source models such as GPT-4. Furthermore, fine-tuning these open-source models with additional data significantly enhances their performance.

Cancer is one of the leading causes of death in the world[1]. According to the Global Cancer Observatory (GLOBOCAN) statistics produced by the International Agency for Research on Cancer (IARC), the estimation for the year 2022 is almost 20 million new cases of cancer and around 9.7 million cancer deaths, globally[2]. Surgery, chemotherapy, and radiotherapy, individually or in combination, represent the standard treatments for cancer patients[3–5]. Nevertheless, although these traditional methods are successful, they still have limitations. For instance, surgical removal of cancer is the most risk-averse option, but it is only possible in its early stages[5]. Moreover, chemotherapy and radiotherapy are not effective for all cancer types[3], as they are not devoid of side effects since they attack cancerous cells as well as healthy cells[4,6]. A promising direction in cancer treatment is precision medicine, which involves studying the patients' genetics, lifestyle, and environmental information to guide treatment decisions[7]. This approach improves care quality while reducing the need for unnecessary diagnostic tests and therapies by guiding healthcare decisions toward the most effective treatments for a given patient[8].

Therapies targeting specific genes or associated with certain response biomarkers have demonstrated considerable promise in the treatment of patients with different cancer types[1,9,10]. According to a study conducted on clinical trial records from 2000 to 2015, clinical trials that use biomarkers for patient stratification have higher success rates, particularly for oncology clinical trials[11]. As of 2022, the FDA acknowledges 45 genes and MSI-H and TMB-H as biomarkers that can predict a patient's response to a drug and five tumor-agnostic genomic biomarkers[12]. Despite the benefits of increasing patient participation in clinical trials, such as accelerating the development of treatments and improving trial generalizability[13], many clinical trials fail to be completed due to poor recruitment, amongst other reasons[14–16]. Even though the final decision to participate in a trial is in the hands of the patient, various barriers often prevent patients from being offered the opportunity to enroll in the trial enrollment process[17]. A systematic review defined four barrier domains affecting a patient's participation in a trial post-diagnosis, leading to only 8.1% of 8883 patients being enrolled[17]. Physicians may not discuss possible trials due to barriers such as treatment preferences[17], lack of awareness[18], or perceiving the process as time-consuming[19–21]. Improving enrollment relies on effective methods to match patients to clinical trials. Although clinical trial information is available through platforms like clinicaltrials.gov[17], the lack of structured data and non-standard naming conventions, coupled with the increasing number of trials lead to a time-consuming and overwhelming search process.

Historically, the process of finding matching patients for clinical studies was manual and labor intensive, often resulting in under-enrollment of trials and delays in the medical advancements. To overcome these challenges and expedite the search process, various computational approaches have been developed over the years. Early computational efforts utilized rule-based algorithms to encode eligibility criteria into structured queries[22]. These systems required significant manual effort to translate free-text eligibility criteria into machine-readable formats. While these methods improved the efficiency of patient-trial matching compared to purely manual methods, they were limited by their rigidity and inability to handle

Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine, Berlin, Germany. ✉e-mail: altuna.akalin@mdc-berlin.de

the complexity and variability of natural language used in medical records and trials descriptions.

In oncology, matching a potential trial to a patient could be even more complicated due to the increasing use of biomarkers and non-standard naming conventions. To enhance interoperability and standardization, oncology-based approaches were introduced. These methods employed medical ontologies like SNOMED CT and UMLS to standardize terminology used in eligibility criteria and patient data[23]. While these systems improved the consistency of data interpretations, they still struggled with the nuances of natural language and the contextual understanding required for accurate matching.

Advancements in NLP enabled the parsing and interpretation of free-text eligibility criteria and electronic health records which improved the patient-trial matching process[24]. An example of these technologies is the Watson for Clinical Trial Matching System. The system employs NLP to process structured and unstructured data such as clinical notes and pathology reports, alongside structured electronic health records (EHRs)[25]. Despite its high accuracy, the system still struggles with the complex clinical reasoning emphasizing the need for continued advancements in contextual understanding. Recent advancements in LLMs, such as GPT-3 and GPT-4, offer significant improvements over traditional methods. These models demonstrate superior language understanding as they are trained on vast amounts of data, enabling them to comprehend complex medical jargon and subtle linguistic nuances in eligibility criteria[26]. Their enhanced natural language processing (NLP) capabilities allow them to excel in tasks such as entity recognition, relation extraction, and semantic understanding, which are crucial for accurately matching patients to trials[27]. Additionally, LLMs reduce the need for manual encoding by processing unstructured data without extensive preprocessing, thereby lightening the workload for healthcare professionals. Furthermore, their adaptability and scalability enable them to be fine-tuned for specific domains or updated as new trials emerge, making them indispensable tools in a rapidly evolving field.

Amidst the rise of LLMs and the large number of ongoing clinical trials, many approaches have emerged to automate patient-to-trial matching, speeding up the process and increasing the chances of finding a suitable trial. Existing approaches apply either an end-to-end matching[28–30] or a structure-then-match strategy[24,31]. In an end-to-end strategy, the LLM is used to compare a patient's record with an unstructured clinical trial document and return a final decision on the trial's eligibility or ranking. In the structure-then-match strategy, the LLM is used to extract and structure entities from the clinical trial text, followed by the matching process. This process is straightforward, assuming the structuring part is done well.

In this study, we took a structure-then-match approach to enhance biomarker-based patient-to-trial matching by improving the biomarker extraction process from clinical trials descriptions, with a focus on the brief summary and eligibility criteria. Our work focuses on structuring these study descriptions for better patient-trial matching, without processing individual patient records. The primary challenge in trial matching is not comparing biomarkers—it's interpreting unstructured trial criteria. Our work structures ambiguous, free-text biomarkers (e.g., 'HER2/ERRB2 mut') into standardized formats with inclusion and exclusion criteria. This step is foundational: without it, even advanced algorithms fail due to nomenclature noise. We evaluate the model's ability to simultaneously extract genomic biomarkers from a clinical trial, pre-process the extracted data, and structure the logical connections (AND/OR) between biomarkers in the disjunctive normal form (DNF)[24]. The DNF is a representation of a logical formula as a disjunction (OR) of conjunctions (ANDs).

To achieve this, we investigated multiple large language models using various prompting techniques and fine-tuned an open-source model as depicted in Fig. 1. With our fine-tuned model, we achieved superior performance in extracting and structuring biomarkers in the DNF form.

Results

Data curation and trial data characteristics

As biomarkers are an important part of cancer drug development and diagnostics, we focused on curating a representative data set. The CIViC database (https://civicdb.org) is an open-source knowledgebase that includes extensive cancer-related biomarker datasets. We identified 500 biomarkers related to cancer diagnostics and therapy. Based on the large-scale AACR-genie cancer patient cohort, which included data from 171,957 patients we estimate that 23.56% of the patients had at least one mutation corresponding to a CIViC biomarker, indicating potential clinical relevance for targeted therapies, or enrollment in biomarker-driven clinical trials. Our findings also suggest that patients with colorectal cancer, breast cancer, and glioma had the highest frequencies of mutations associated with CIViC biomarkers (Fig. 2).

We have obtained the ongoing oncology clinical trials from clinicaltrials.gov. A clinical trial typically includes a summary of the trial plan, followed by a detailed description, and the inclusion and exclusion eligibility criteria required for enrollment. Then we query the trials with the list of biomarkers from the CIViC database. This allowed us to select 296 unique trials with the potential presence of a biomarker in the eligibility criteria. From these trials, we manually annotated 166 of the trials, detailing the inclusion and exclusion biomarkers for each trial in JSON format. We removed one outlier sample that had a significantly larger token count (Supplementary Fig. 1). We then split the data into a 70:30 ratio resulting in 116 training samples and 50 testing samples.

We prepared the training dataset for fine-tuning with Direct Preference Optimization (DPO) and split it into an 80:20 ratio which resulted in the first fine-tuning dataset DPO-92 with 92 samples as training set and 23 samples as validation set. To create a second fine-tuning dataset, the original dataset, DPO-92 was augmented with 80 synthetically generated samples using GPT-4 (See "Generation of Synthetic Dataset" for details). These additional samples underwent the same process of preparation for fine-tuning with DPO. The combined dataset was split following an 80:20 ratio resulting in the dataset DPO-156 with 156 samples for training and 39 samples for validation. Refer to "DPO Dataset Preparation" for more details.

Zero-shot, few-shot and prompt chaining performance

We employ the models using three prompting techniques: zero-shot prompting where the model is given instructions to be followed, prompt chaining where the task is performed using a chain of requests to the model, and finally, few-shot prompting where the model is provided with examples to demonstrate the task. Refer to the methods section for a detailed description of the prompting techniques.

In Table 1 we report the models' performance with the prompting techniques using a preliminary evaluation approach where we only measure the model's ability to extract the inclusion and exclusion biomarkers found in the clinical trial. With zero-shot prompting, GPT-3.5-Turbo had a moderate performance in the extraction of the inclusion biomarkers with an F2 score 0.45. However, the model struggled with the extraction of exclusion biomarkers resulting in one of the lowest F2 scores (0.06). This challenge for the GPT-3.5-Turbo model to extract the exclusion criteria correctly may be associated with the length of the clinical trial input and the position of the exclusion criteria at the end of the input, which might have caused the increase in hallucination for the model. Swapping GPT-3.5-Turbo with GPT-4 led to better performance with an F2 score of 0.56 for the inclusion biomarkers and 0.42 for the exclusion biomarkers. NousResearch/Hermes-2-Pro-Mistral-7B (https://huggingface.co/NousResearch/Hermes-2-Pro-Mistral-7B) is a large language model that excels at generating structured JSON outputs. Moreover, the model it was fine-tuned from, Mistral-7B, exhibits superior performance over LLAMA-7B across all benchmarks and outperformed LLAMA-34B in mathematical and coding tasks[32].

Hermes-2-Pro-Mistral-7B, demonstrated remarkable capabilities, outperforming the closed-source models at extracting both inclusion and exclusion biomarkers, achieving F2 scores of 0.98 and 0.66, respectively.

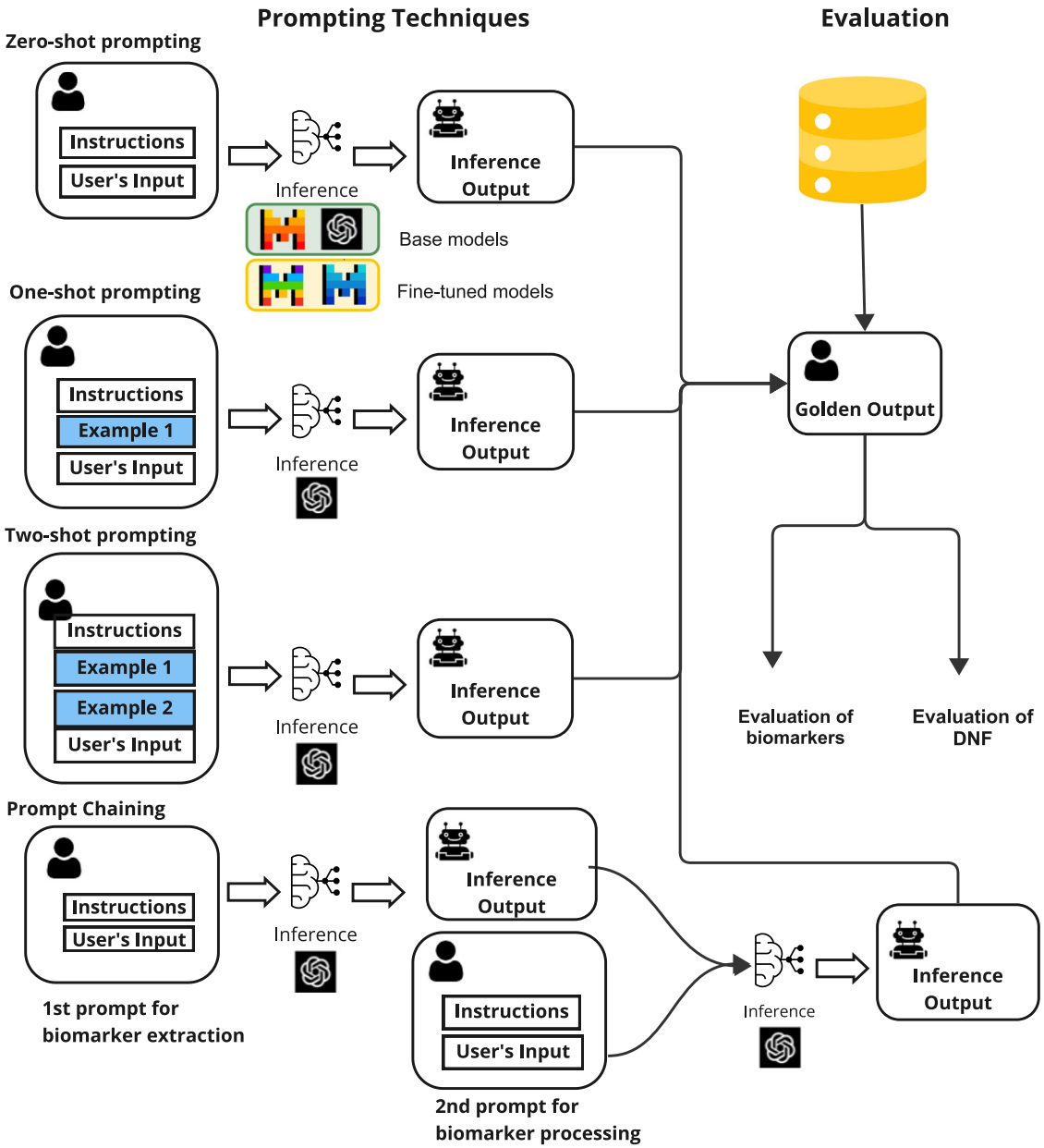

Fig. 1 | Overall framework for biomarker extraction and evaluation. This figure illustrates the prompting techniques explored for extracting genetic biomarkers from clinical trials. It underlines the evaluation of the extraction process both with and without assessing the model's capability to organize the biomarkers in Disjunctive Normal Form (DNF). The framework includes several prompting approaches, including zero-shot, one-shot, two-shot prompting, and prompt chaining, which involves sequential prompts for biomarker extraction and processing. The framework also highlights the usage of fine-tuned models.

Splitting the task into two subtasks with the prompt chaining technique where the first prompt's output is the input to the second had opposite effects on GPT-3.5-Turbo and GPT-4. For GPT-3.5-Turbo, prompt chaining resulted in a noticeable decline in performance compared to zero-shot prompting. The F2 score for the inclusion biomarkers decreased by approximately 42%, while the F2 score for the exclusion biomarkers decreased by around 16%. In contrast, GPT-4 's performance increased with prompt chaining to reach an F2 score of 0.76 for the inclusion biomarkers and 0.70 for the exclusion biomarkers. Few-shot prompting (1S and 2S) showed improvement over zero-shot prompting and prompt chaining for GPT-3.5-Turbo. However, providing two-shot prompting did not improve over one-shot prompting; overall there was a slight decrease in performance. This finding is surprising, as we generally observe an increase in performance with an increase in the number of examples provided to the model[33], suggesting that the selected examples might have introduced some conflict or caused the model to overfit.

In Table 2 we show the results with a more mature evaluation approach where we assess the models' ability to not only extract the biomarkers correctly but to also structure the biomarkers in DNF in the JSON output as seen in Fig. 3. GPT-3.5-Turbo with zero-shot prompting performed the worst with an F2 score of nearly zero for the inclusion and exclusion biomarkers. GPT-4 had an overall moderate performance with an F2 score of 0.29 for the inclusion and an F2 score of 0.43 for the exclusion biomarkers, hence outperforming GPT-3.5-Turbo. The open-source model Hermes-2-Pro-Mistral-7B with zero-shot prompting outperformed GPT-3.5-Turbo and GPT-4. At the extraction of the inclusion biomarker, Hermes-2-Pro-Mistral-7B was approximately 3.24 times better than GPT-4 with an F2 score of 0.94, and around 1.5 times better for the exclusion biomarkers with an F2 score of 0.65.

Employing GPT-3.5-Turbo with prompt chaining (PC) increased its performance slightly to reach an F2 of 0.14 for the inclusion biomarkers and an F2 score of 0.06 for the exclusion biomarkers. However, for GPT-4, applying chain prompting led to a decrease in performance, specifically, the model's F2 score decreased by 7% compared to zero-shot prompting in both inclusion and exclusion biomarker extractions.

Providing an example to demonstrate the task to GPT-3.5-Turbo (1S) caused a surprising improvement in its performance for the inclusion biomarkers to have an F2 score of 0.53. However, for the exclusion biomarkers even though there was an improvement compared to the zero-shot prompting (0S) and prompt chaining, the model still performed poorly with an F2 score of 0.12. When comparing two-shot prompting (2S) to one-shot prompting (1S) for GPT-3.5-Turbo, the F2 score was nearly zero. In contrast, there was no noticeable change in the performance of the inclusion biomarkers.

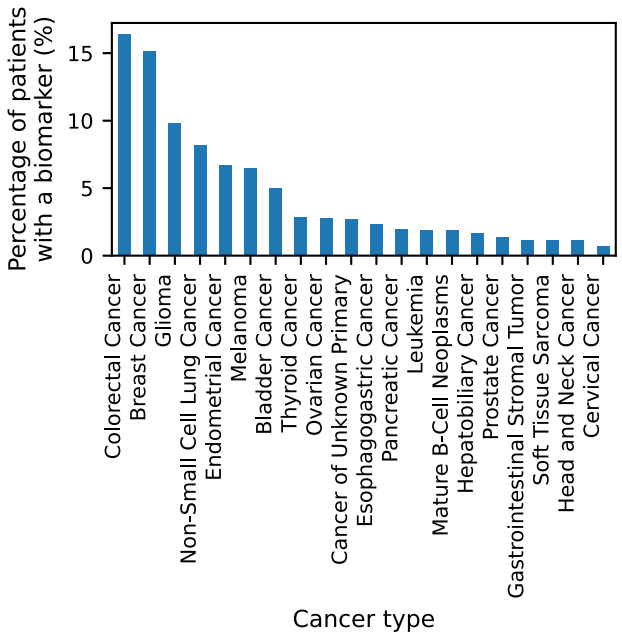

Fig. 2 | Distribution of Top 20 Cancer Types. Bar chart showing the distribution of the top 20 cancer types among patients who matched at least one of the 500 biomarkers selected from CIViC. The patients' records are collected from the American Association for Cancer Research (AACR) Project GENIE.

Fine-tuned model performance

As mentioned in the previous section, applying Hermes-2-Pro-Mistral-7B with zero-shot prompting already showed impressive results in performing our task, compared to the closed-source models. However, we notice a discrepancy between inclusion and exclusion in the model's abilities to extract the biomarkers and organize them in the DNF form as seen in Table 1 and Table 2. To potentially improve the extraction of exclusion biomarkers, we fine-tuned the model using Direct Preference Optimization (DPO). The purpose of fine-tuning is to adapt the LLM to a more specialized task by adjusting its parameters. The DPO algorithm adjusts the model's parameters to increase the likelihood of generating the desired output by dragging the probability distribution towards the preferred answer and away from the undesired one. Here, we fine-tuned two models, both initialized from Hermes-2-Pro-Mistral-7B, with each model fine-tuned on a dataset of different size. The first model, Hermes-FT was trained on the labeled dataset DPO-92, while the second model Hermes-FT-synth was trained on the larger datasets DPO-156. For more details on the datasets, please refer to the Methods section "DPO Dataset Preparation". Table 3 demonstrates how fine-tuning the model with DPO-92 which includes 92 manually annotated clinical trials samples decreased the overall performance. We observe a 10% decrease in the F2 score for the extraction of inclusion biomarkers and approximately a 7.5% decrease for the extraction of exclusion biomarkers. Nevertheless, Hermes-FT-synth, fine-tuned with a larger dataset (DPO-156) demonstrated superb capabilities, outperforming all models with F2 scores of 0.90 and 0.93 for inclusion and exclusion biomarkers, respectively. Despite the overall superiority in performance, the model demonstrated a slight decrease in recall for inclusion biomarkers extraction was observed compared to the base model, Hermes-2-Pro-Mistral-7B. This suggests that the fine-tuned model overlooked some biomarkers during extraction. This behavior could be attributed to the model learning to optimize the extraction of exclusion biomarkers during fine-tuning, thereby reducing hallucination and false positives. It is also possible that during fine-tuning, the model experienced changes in how it handles inclusion and exclusion criteria, causing it to provide brief responses to minimize hallucination in inclusion biomarkers, which led to a decrease in recall.

In Table 4, using Hermes-FT to extract and structure the biomarkers in DNF yielded unexpected results: a decrease in inclusion biomarkers' performance with an F2 score of 0.85, alongside a marginal increase in extraction of exclusion biomarkers with an F2 score of 0.67. Hermes-FT-synth resulted in the overall best performance with an F2 score of 0.86 for inclusion biomarkers and 0.94 for exclusion biomarkers.

Discussion

In this study, we explore the capabilities of both closed-source and open-source LLMs to enhance patient matching to biomarker-driven oncology

Table 1 | Performance of the open-source and closed-source base models using the different prompting techniques in extracting the inclusion and exclusion biomarkers from the clinical trials free-text documents

	Inclusion Biomarkers			Exclusion Biomarkers		
Technique	***Precision*** ↑	***Recall*** ↑	***F2*** ↑	***Precision*** ↑	***Recall*** ↑	***F2*** ↑
GPT-3.5-Turbo (0S)	0.61	0.42	0.45	0.02	0.18	0.06
GPT-3.5-Turbo (PC)	0.21	0.28	0.26	0.02	0.13	0.05
GPT-3.5-Turbo (1S)	0.46	0.60	0.56	0.06	0.21	0.14
GPT-3.5-Turbo (2S)	0.40	0.59	0.54	0.05	0.13	0.10
GPT-4 (0S)	0.55	0.56	0.56	0.47	0.41	0.42
GPT-4 (PC)	0.77	0.76	0.76	**0.75**	0.68	**0.70**
Hermes-2-Pro-Mistral-7B (0S)	**1**	**0.97**	**0.98**	0.42	**0.77**	0.66

The prompting techniques applied include zero-shot (0S) prompting, where the prompt describes the input, output, and task; one-shot (1S) prompting, where an additional example is provided to demonstrate the task; two-shot (2S) prompting, where two examples are given to illustrate the task; and prompt chaining (PC), which divides the task into subtasks with the output of one prompt serving as input for the next. Here, we apply a chain of two prompts where the first prompt handles extraction and its output is the input to the second prompt that handles the pre-processing and structuring the biomarkers in the JSON output. We ran the Hermes-2-Pro-Mistral-7B model three times, and the results in the table represent the consistent outcomes from these runs.
The bold font represents the best result across different techniques.

trials. It is worth noting that our study does not involve individual patient records but rather focuses on extracting biomarkers from clinical trial descriptions, as this is the primary challenge in trial matching using biomarkers. Once inclusion and exclusion biomarkers are extracted in a structured way, it will be more efficient to match patients based on biomarkers. Our results demonstrated a significant discrepancy between the closed-source models and the open-source model. This observed difference suggests that an increase in the volume of the model does not necessarily mean better performance[34] and highlights the importance of the training process. The closed-source models GPT-3.5-Turbo and GPT-4 underwent Reinforcement Learning with Human Feedback (RLHF) which helped the models excel at various standard natural language tasks. However, RLHF suffers from a phenomenon known as the "alignment tax"[35], which states that the model's ability to generate more human-like responses is at the expense of the model's deeper understanding of tasks[36,37]. In our work, where the model is required to handle a multi-step task, GPT-3.5-Turbo and GPT-4 models struggle to fully comprehend this complex task, even when applied with prompt chaining, one-shot, and two-shot prompting. While the 7 billion parameter model, Hermes-2-Pro-Mistral-7B, that underwent supervised fine-tuning from Mistral-7B, with zero-shot prompting had overall better reasoning capabilities outperforming the closed-source models.

Table 2 | Performance of the open-source and closed-source base models using the different prompting techniques (zero-shot (0S), one-shot (1S), two-shot (2S) and prompt chaining (PC)) in extracting the inclusion and exclusion biomarkers from clinical trials documents while structuring them in the JSON output while adhering to the disjunctive normal form (DNF)

	Inclusion Biomarkers			Exclusion Biomarkers		
Technique	*Precision* ↑	*Recall* ↑	*F2* ↑	*Precision* ↑	*Recall* ↑	*F2* ↑
GPT-3.5-Turbo (0S)	0.18	0.04	0.05	0	0	0
GPT-3.5-Turbo (PC)	0.15	0.13	0.13	0.02	0.12	0.06
GPT-3.5-Turbo (1S)	0.47	0.54	0.53	0.05	0.18	0.12
GPT-3.5-Turbo (2S)	0.44	0.55	0.52	0.01	0.02	0.01
GPT-4 (0S)	0.40	0.27	0.29	0.61	0.40	0.43
GPT-4 (PC)	0.47	0.25	0.27	**0.62**	0.37	0.40
Hermes-2-Pro-Mistral-7B (0S)	**0.99**	**0.93**	**0.94**	0.41	**0.75**	**0.65**

We ran the Hermes-2-Pro-Mistral-7B model three times to confirm its consistency.
The bold font represents the best result across different techniques.

To validate the robustness of our results, we ran the base model (Hermes-2-Pro-Mistral-7B) and the two fine-tuned models three times with a temperature setting of 0. Across all runs, the models consistently produced identical predictions, leading to the exact same inclusion and exclusion (with and without DNF) precision, recall, and F2 scores. This confirms the stability and reproducibility of the models in extracting and structuring biomarkers from clinical trial descriptions.

In our work, the open-source model, out-of-the-box, demonstrated superior performance. However, the model presented a disparity in accuracy between the inclusion and exclusion biomarkers. Fine-tuning the model with Direct Preference Optimization using a training dataset consisting of manually extracted samples and synthetically generated samples by GPT-4 not only helped close this gap but also achieved the best overall performance (Hermes-FT-synth).

Our study also reveals several key findings. For instance, GPT-3.5-Turbo with few-shot prompting performed competitively with GPT-4 in zero-shot prompting. However, the results highlight the importance of the examples selected, as we notice a significant increase in performance when extracting the inclusion biomarkers while the model still suffered at extracting the exclusion biomarkers. Furthermore, with two-shot prompting the model's performance decreased suggesting a bias in the first example towards our test set since adding a second example introduced conflict for the model causing it to hallucinate more. We discovered that prompt chaining was less reliable than zero-shot prompting. This might be that in prompt chaining the prompts are interdependent leading to a decrease in coherence, and leaving the model with not enough context to efficiently extract biomarkers[38]. Another key finding is that fine-tuning with DPO presents challenges related to dataset size. Fine-tuning with a relatively small dataset decreased the model's performance (Hermes-FT) due to overfitting

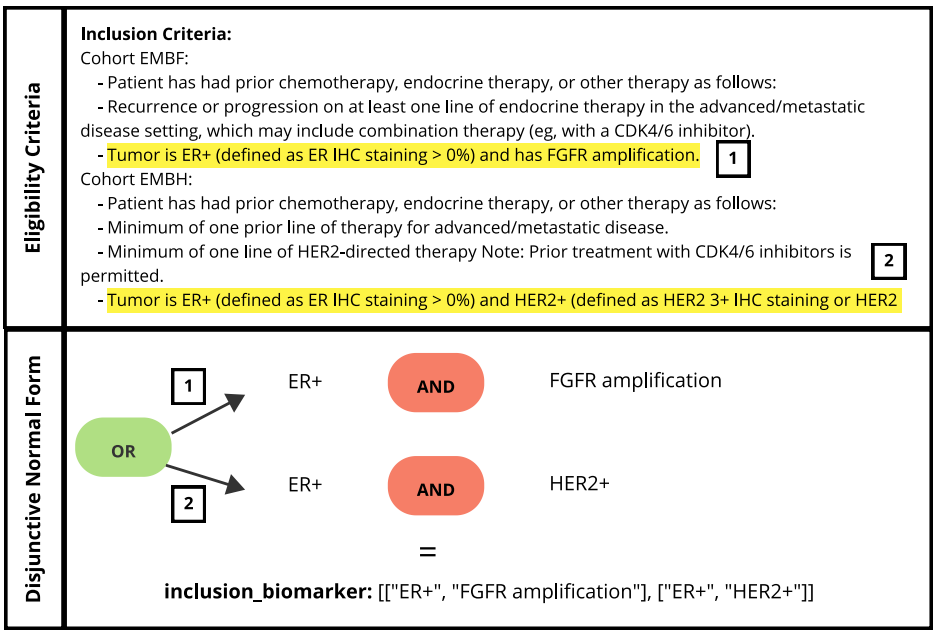

Fig. 3 | Representation of extracted biomarkers in the disjunctive normal form. The figure represents a snippet of the inclusion criteria in the clinical trial NCT04092673. The figure highlights the inclusion requirements for two cohorts EMBF and EMBH. Cohort one EMBF requires the tumor to be ER+ with FGFR amplification. Cohort two required the patient's tumor to be ER+ and HER2 + . These criteria are translated into a structured JSON output in the Disjunctive Normal Form (DNF), reflecting that patients are eligible if they meet either the condition of ER+ with FGFR amplification or ER+ with HER2 + .

Table 3 | Performance of the fine-tuned open-source models with Direct Preference Optimization (DPO) using zero-shot prompting in extracting the inclusion and exclusion biomarkers from clinical trial documents

	Inclusion Biomarkers			Exclusion Biomarkers		
Technique	***Precision*** ↑	***Recall*** ↑	***F2*** ↑	***Precision*** ↑	***Recall*** ↑	***F2*** ↑
Hermes-FT (0S)	0.65	**0.97**	0.88	0.25	0.95	0.61
Hermes-FT-synth (0S)	**0.99**	0.88	**0.90**	**0.82**	**0.96**	**0.93**

The models Hermes-FT and Hermes-FT-synth are fine-tuned from Hermes-2-Pro-Mistral-7B with a training dataset of 92 and 156 samples, respectively, to evaluate the impact of dataset size on model performance. We executed the evaluation of the models three times, confirming the consistency of the models across all runs.
The bold font represents the best result across different techniques.

Table 4 | Performance of the fine-tuned models Hermes-FT and Hermes-FT-synth using zero-shot prompting in extracting the inclusion and exclusion biomarkers from clinical trials documents while structuring them in the JSON output while adhering to the disjunctive normal form (DNF)

	Inclusion Biomarkers			Exclusion Biomarkers		
Technique	***Precision*** ↑	***Recall*** ↑	***F2*** ↑	***Precision*** ↑	***Recall*** ↑	***F2*** ↑
Hermes-FT (0S)	0.63	**0.93**	0.85	0.31	0.95	0.67
Hermes-FT-synth (0S)	**0.99**	0.84	**0.86**	**0.86**	**0.96**	**0.94**

The comparison highlights the influence of the fine-tuning dataset size on the performance of the model for our task. We ran each of the models three times and confirmed their consistency.
The bold font represents the best result across different techniques.

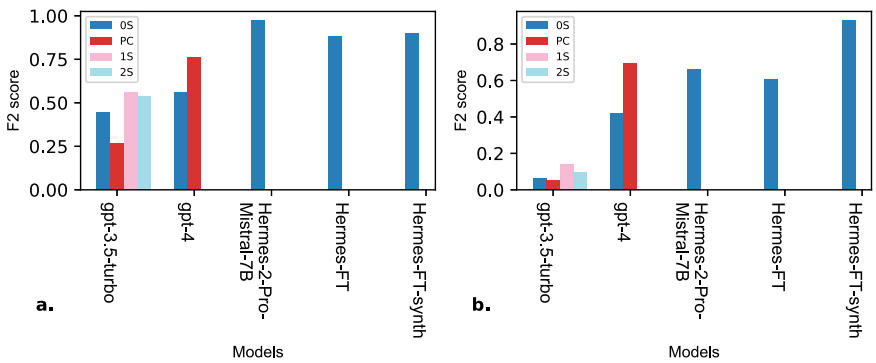

Fig. 4 | Comparison of biomarker extraction across models. This figure presents a comparative analysis of F2 scores for the base models: gpt-3.5-turbo, gpt-4, Hermes-2-Pro-Mistral-7B, and the fine-tuned models: Hermes-FT and Hermes-FT-synth using multiple prompting techniques: zero-shot (0S), prompt chaining (PC), one-shot (1S), and two-shot (2S). The figure highlights the performance in extracting the inclusion biomarkers (**a**) and the exclusion biomarkers (**b**) without considering the model's abilities in formatting the biomarkers in the Disjunctive Normal Form (DNF).

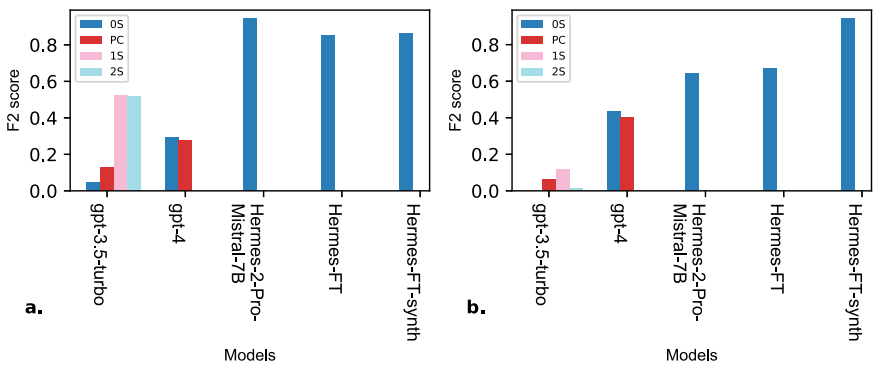

Fig. 5 | Comparison of biomarker extraction and representation in the Disjunctive Normal Form (DNF) across models. The figure presents a comparative analysis of F2 scores for the base models: gpt-3.5-turbo, gpt-4, Hermes-2-Pro-Mistral-7B, and the fine-tuned models: Hermes-FT and Hermes-FT-synth using multiple prompting techniques: zero-shot (0S), prompt chaining (PC), one-shot (1S), and two-shot (2S). The figure demonstrates the performance in extracting the inclusion biomarkers (**a**) and the exclusion biomarkers (**b**) while considering the model's abilities in organizing the biomarkers in the Disjunctive Normal Form (DNF).

and poor generalization[39]. Augmenting the training set with synthetically generated samples with GPT-4 improved the model's (Hermes-FT-synth) overall surpassing other models without having negative effects (Fig. 4).

Our results demonstrated that leveraging LLMs to extract a structured output while adhering to the DNF is possible (Fig. 5). This added layer of complexity provided insights into the models' reasoning abilities and capacity to handle real-world clinical data effectively. An adequate amount of training data is crucial since it allows for a successful supervised fine-tuning. Augmenting the manually annotated clinical trials training set with synthetically generated data reduces the time needed for the generation of a fully human-annotated dataset, opening the door for improved fine-tuning.

While this work focuses on clinical trial data, we recognize that the broader landscape of real-world clinical data often involves unstructured sources such as physician notes and tumor genotypes embedded in PDFs.

These unstructured formats present unique challenges, including variability in language, lack of standardization, and the difficulty of extracting data from non-machine-readable documents.

Recent advances in combining Optical Character Recognition (OCR) and NLP offer promising solutions to these challenges. For instance, deep learning-based post-correction methods have been shown to improve OCR accuracy for scanned medical documents[40]. Similarly, NLP approaches like zero-shot learning tools[41] and domain-specific models, such as CancerBERT[42], have successfully abstracted structured information from free-text oncological pathology reports and unstructured healthcare data. These advancements illustrate how methodologies similar to ours could potentially be extended or adapted for applications involving unstructured clinical data.

Incorporating such techniques into our approach represents an exciting avenue for future research. These integrations could enable the methodologies presented here to bridge the gap between different types of clinical data, addressing critical real-world challenges and broadening the impact of this work.

This study is not without its limitations. The dataset we used for testing our models' performance is rather small. This small dataset allows us to compare the models' performance and demonstrate the potential of LLM in performing complex tasks requiring reasoning abilities. However, the test set does not fully encapsulate the complexity of clinical trials criteria. Moreover, in few-shot learning, the selected examples showed a clear bias to the samples in our test set. To avoid such overfitting, more advanced prompting techniques, such as Retrieval-Augmented Generation[43], could be applied. While our study incorporated a wide range of prompting techniques and fine-tuning with DPO, two suggestions could help improve performance: hyperparameters tuning during supervised fine-tuning, and testing newly released models such as LLAMA3 (https://github.com/meta-llama/llama3) that may provide better results. However, one might need to consider other factors than improving the performance such as the inference latency and the token usage when optimizing the prompting techniques. In the real-world applications, scalability and cost constraints are key factors. In addition, we are only focusing on biomarkers in this study, other criteria such as disease stage, disease type, age, gender, and previous therapies are also important for eligibility for an oncology trial. However, with the increasing importance of biomarkers in oncology therapies the goal of this work is to improve this aspect of trial-matching performance.

Methods

Dataset curation and annotation

We retrieved and processed clinical trials from the clinicaltrials.gov database and performed keyword filtering to focus our study on oncology-related clinical trials, which resulted in around 15,856 trials stored in the vector database, ChromaDB. A clinical trial is a text document with a brief description mentioning the purpose of the trial, a detailed description of the trial and the participation criteria that lists the inclusion criteria required for a patient to be considered for enrollment and the exclusion criteria that if the patient possesses would disqualify them from participating in the trial. Our database, ChromaDB, includes only the "brief description" and "eligibility criteria" sections from the clinical trials document. No patient records data were used in this study.

To select trials that contained genomic biomarkers from our database, we used a list of 500 genomic biomarkers from the CIViC database to perform a semantic search against the cancer clinical trials.

Next, we manually annotated 166 clinical trials by assigning each trial the corresponding JSON output containing the inclusion and exclusion biomarkers present in that trial. After annotation, we assessed the token distribution to identify outliers that exhibit a significantly larger token count, which we removed to ensure uniformity in the data for downstream analysis. We used 70% of the trials for supervised fine-tuning as a training dataset (train set) and 30% as a test dataset.

Manual annotation. The expected output is JSON with two keys, *"inclusion_biomarker"* and *"exclusion_biomarker"*. The *"inclusion_biomarker"* contains the genomic biomarkers required for a patient's consideration in the enrollment process. In contrast, *"exclusion_biomarker"* contains genomic biomarkers that result in the patient being excluded from the trial enrollment process. To accurately capture the logical connections between biomarkers in the trial (AND/OR logic), we represent them in the DNF.

The DNF which can be described as the disjunction (OR) of conjunctions (ANDs), provides a standardized approach to expressing complex logical connections. In the context of our study, we implement a list of lists data structure to represent the biomarkers in the DNF formula. Each inner list represents a conjunction clause (AND), where all biomarkers conditions in that list must be satisfied. The outer list connects these conjunction clauses through disjunction (OR), where only one of the inner conjunction clauses must be met. This DNF structure allows for a better representation of the complex relationships between genomic biomarkers in clinical trials, enabling more precise patient-trial matching.

Generation of synthetic dataset

To augment the training dataset used in the supervised fine-tuning process, we generated synthetic samples with GPT-4 employed with the prompt provided in Supplementary Fig. 2. We included four examples selected from our training dataset in the prompt, with the placeholders in the template indicating where we inserted the examples. Additionally, we appended at the end of the prompt "Trial:" to encourage the language model to generate a new clinical trial that matches the style and format of the given examples. The generated samples comprise a clinical trial input and the corresponding JSON output containing the inclusion and exclusion biomarkers.

Additionally, we predicted the JSON output for our four examples using GPT-4 with zero-shot prompting (Supplementary Fig. 3) to ensure the output was coming from the same distribution to increase the consistency and reliability of the generated samples. Then, we manually reviewed the generated trials to ensure their quality and that they do not contain any unusual or unexpected content that could negatively impact the performance. Finally, we added the synthetically generated samples to the manually annotated training set.

DPO dataset preparation

In this section, we describe the process we followed to prepare our dataset for fine-tuning with Direct Preference Optimization (DPO)[44]. The DPO algorithm is a technique used to optimize the language model for a downstream task. This algorithm optimized the model by increasing the likelihood of preferred responses (winning completions) and decreasing the likelihood of less preferred responses (losing completions). Refer to "Direct Preference Optimization (DPO) Fine-tuning" section for more details.

We generated two datasets for fine-tuning with DPO, the first deriving from the manually annotated dataset (115 samples) and the second from the mixed dataset (195 samples). A dataset suitable for fine-tuning with DPO should include the input prompts, winning completions (y_W) and losing completions (y_L). The winning completions were either obtained through manual annotation or synthetically generated by GPT-4, with the latter being reviewed and validated by human annotators. To generate the losing completion, we sample from the supervised fine-tuned model Hermes-2-Pro-Mistral-7B that we used later for fine-tuning, which we will further describe later.

We used the zero-shot prompt template from Supplementary Fig. 4 to generate the prompt for each sample. Once the prompts are prepared, we tokenize them with left padding enabled to ensure that all inputs have the same length.

Using the tokenized prompts, we perform inference with the Hermes-2-Pro-Mistral-7B model, setting the temperature to zero. The generated completion is the losing completion. This inference process was repeated for every sample in each of the datasets.

We further split the samples into 80% as a training set to fine-tune the model and 20% as a validation set to avoid overfitting during fine-tuning.

Language models

We investigate the performance of the closed-source models from OpenAI GPT-3.5-turbo and GPT-4, as well as the open-source model NousResearch/Hermes-2-Pro-Mistral-7B. To ensure the robustness and consistency of our results, we ran the base model (Hermes-2-Pro-Mistral-7B) and the two fine-tuned models (Hermes-FT and Hermes-FT-synth) three times.

Even though for previous GPT models details about model architecture and training process were transparent, these details have not been disclosed for GPT-3.5-turbo and GPT4 models. From previous models, we know that GPT models are based on decoder-only transformers, a variant of the original transformer architecture[36]. It was reported that GPT-3.5-Turbo is the most capable model from the GPT-3.5 series[37]. In this paper, we use the GPT-3.5-Turbo-0125 and GPT-4 (GPT-4-0613) models. Additionally, the technical report released by OpenAI[36] mentions that GPT-4 outperforms their previous 175B parameter GPT-3 model.

Hermes-2-Pro-Mistral-7B is fine-tuned using teknium/OpenHermes-2.5 (https://huggingface.co/datasets/teknium/OpenHermes-2.5) from Mistral-7B[32], a pre-trained generative text model with seven billion (7B) parameters and a context length of 8192.

Prompting Techniques

Language models are versatile multitask learners[45], during training on large and diverse datasets, the pre-trained models learn to perform different tasks. The models' ability to learn a variety of tasks led to the success of natural language prompting, a method used to condition the model to perform a specific task without updating any of the pre-trained weights[26]. A prompt is used to provide the model with the necessary information to generate the desired output. Multiple techniques of prompting exist, in this study we employ zero-shot prompting, few-shot prompting and prompt chaining.

- Zero-shot prompting: The model is given, at inference time, a prompt that includes the description of the task that the LLM should perform. It usually includes details about the task, the user's input and the format of the output to be returned by the language model[46].
- Few-shot learning: Often referred to as in-context learning. Few-shot learning is similar to zero-shot prompting, but in addition, at inference, the model is provided with a set of k examples that demonstrate the task[26]. Each example typically includes an input and the corresponding expected output. Typically, in few-shot learning the model is given k examples and then followed by the user's input.
- Prompt chaining: Recent work[47] has shown that for complex tasks, decomposing the task into subtasks and chaining together prompts of each subtask can increase output quality. The idea behind prompt chaining is that the output from one prompt is the input of the following prompt, with each sub-task building on the last. This approach should in principle allow the model to tackle complex tasks that may be difficult to accomplish in a single prompt incrementally.

In the zero-shot prompts (Supplementary Figs. 3, 4) we start by giving a summary of our task. In our instructions, we define in detail the biomarkers as any changes at the level of chromosomes, genes, or proteins and the exact JSON output expected including a detailed definition of the DNF logic and how to represent it in the output. One step further, to decrease false positives, we added a description of some common eligibility criteria such as cancer type, pregnancy history, age, gender, etc. and instructed the language model to ignore this information since they do not classify as genetic biomarkers. We also reiterate and mention the important information more than once such as to focus on genomic biomarkers and ignore other details. At the end we describe the processing steps to be performed on the extracted biomarkers. It is noteworthy that while GPT and Hermes-2-Pro-Mistral-7B prompts differ in structure and wording, the fundamental details remain the same.

In our study, we applied two forms of few-shot learning: one-shot prompting, where the model is given only one example ($k = 1$), and two-shot prompting, where the model is given two examples ($k = 2$). The one-shot (Supplementary Fig. 5) and two-shot (Supplementary Fig. 6) prompts used with GPT-3.5-Turbo are similar to the zero-shot prompt with the exception of the additional examples consisting of a clinical trial input document and the corresponding JSON output. Examples of these prompts are illustrated in Supplementary Figs. 7 and 8, respectively.

For the chain of prompts used with GPT models, the aim was to reduce the number of tokens in the prompt and simplify the task for the language model by splitting it into two subtasks where the first prompt (Supplementary Fig. 9) focuses on the task of extracting relevant biomarkers from the clinical trial input and returning a list with AND/OR indicators. While the second prompt's (Supplementary Fig. 10) focus is processing the biomarkers and structuring the final JSON output with DNF.

Direct Preference Optimization (DPO) fine-tuning

Fine-tuning is a technique used to adapt the LLM to a downstream task and return responses that meet specific requirements. Compared to prompting techniques, during fine-tuning the model's pre-trained parameters are adjusted. The Direct Preference Optimization algorithm (DPO) optimizes the language model (also referred to as policy) by using preference data directly and minimizing the loss function, as shown in Eq. (1):

$$L_{DPO}(\pi_\theta;\pi_{ref}) = -E_{(x,y_w,y_L)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_W|x)}{\pi_{ref}(y_W|x)} - \beta\log\frac{\pi_\theta(y_L|x)}{\pi_{ref}(y_L|x)}\right)\right] \quad (1)$$

Where x is the prompt and y is the completion. The annotation y_W refers to the winning completion (desired output), the completion humans prefer over the losing y_L completion (rejected output). This loss function directly optimizes the policy model with parameters θ to increase the probability of the preferred completions $\pi_\theta(y_W|x)$ compared to the dispreferred completions $\pi_\theta(y_L|x)$. The reference model probabilities in the denominator constrain the update to not deviate too far from the original.

Fine-tuning large language models can be very expensive, it requires a lot of GPU memory. Parameter-efficient fine-tuning methods, such as Quantized Low Rank Adaption (QLoRA)[48], allow the adaptation of models with a large number of parameters without sacrificing performance by significantly reducing the number of trainable parameters. QLoRA combines the concept of quantization and Low-Rank Adaptation (LoRA)[49]. The LoRA technique assumes that a small number of weight parameters require adaptation when fine-tuning a pre-trained model with full-rank weight matrix $W_0 \epsilon R^{d\times k}$ for a downstream task. LoRA takes advantage of this by decomposing W_0 into two low-rank matrices $B \epsilon R^{d\times r}$ and $A \epsilon R^{r\times k}$, and rank $r \ll \min(d,k)$, such that the relationship in Eq. (2) holds.

$$W_0 + \Delta W = W_0 + BA \quad (2)$$

During fine-tuning, W_0 is fixed, and matrices A and B contain the trainable parameters.

Initially, ΔW is zero since a random Gaussian initialization is used for A and zero initialization for B. LoRA also rescales $\Delta W\ x$ by α/r, where α is a constant and r is the rank. This low-rank approximation allows efficient fine-tuning of a small subset of adapted weights B and A. Additionally, QLoRA involves freezing the pre-trained parameters in a quantized representation while fine-tuning the extra set of Low-Rank Adapters.

We employed DPO with QLoRA to fine-tune the Hermes-2-Pro-Mistral-7B model. During the training process, we used the paged AdamW 32 bit optimizer with a learning rate of 5e-5, specifying the step size taken by the optimizer during backpropagation to adjust the trainable parameters and minimize the loss function. Moreover, we allow the model to train for

200 steps with a batch size of 8. For the beta factor in the DPO loss function we used the default value 0.1.

For the LORA configuration, the rank is set to 2 and the scaling factor to 4. The trainable parameters are limited to the linear projection layers of the self-attention and multilayer perceptron modules. Furthermore, a dropout probability for the LORA layers of 0.05 is applied. A summary of the hyperparameters used during the fine-tuning process is provided in Supplementary Table 1.

The Hermes-2-Pro-Mistral-7B model weights are quantized to a 4-bit representation to reduce memory footprint. During the forward and backward propagation, the model weights are represented in Float16, making it suitable for neural network computations while maintaining sufficient precision.

Data availability

We have included all the data used in the github repository publicly and freely available.

Code availability

The source code is available at https://github.com/BIMSBbioinfo/oncotrialLLM. Additionally, the fine-tuning and fine-tuned models can be found at https://huggingface.co/nalkhou. Live demo of certain techniques for biomarker-based trial matching can be tested at https://onconaut.ai.

Received: 11 September 2024; Accepted: 24 April 2025;
Published online: 06 May 2025

References

1. Padma, V. V. An overview of targeted cancer therapy. *Biomedicine* **5**, 1–6 (2015).
2. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
3. Xing, K. & Shen, L. Molecular targeted therapy of cancer: the progress and future prospect. *Front. Lab. Med.* **1**, 69–75 (2017).
4. Zhong, L. et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduc. Target. Ther.* **6**, 201 (2021).
5. Kaur, R., Bhardwaj, A. & Gupta, S. Cancer treatment therapies: traditional to modern approaches to combat cancers. *Mol. Biol. Rep.* **50**, 9663–9676 (2023).
6. Pirovano, G. et al. TOPK modulates tumour-specific radiosensitivity and correlates with recurrence after prostate radiotherapy. *Br. J. Cancer* **117**, 503–512 (2017).
7. Manzari, M. T. et al. Targeted drug delivery strategies for precision medicines. *Nat. Rev. Mater.* **6**, 351–370 (2021).
8. Ginsburg, G. S. & Phillips, K. A. Precision medicine: from science to value. *Health Aff.* **37**, 694–701 (2018).
9. Choi, H. Y. & Chang, J.-E. Targeted therapy for cancers: from ongoing clinical trials to FDA-Approved drugs. *Int. J. Mol. Sci.* **24**, 13618 (2023).
10. Slamon, D. J. et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792 (2001).
11. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2018).
12. Suehnholz, S. P. et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. *Cancer Discov.* **14**, 49–65 (2023).
13. Unger, J. M., Cook, E., Tai, E. & Bleyer, A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am. Soc. Clin. Oncol. Educ. Book* 185–198 https://doi.org/10.1200/edbk_156686 (2016).
14. Stensland, K. D. et al. Adult cancer clinical trials that fail to complete: an Epidemic? *JNCI J. Natl Cancer Inst.* **106**, dju229 (2014).
15. Unger, J. M. et al. The scientific impact of positive and negative phase 3 cancer clinical trials. *JAMA Oncol.* **2**, 875 (2016).
16. Cheng, S. K., Dietrich, M. S. & Dilts, D. M. A sense of urgency: evaluating the link between clinical trial development time and the accrual performance of cancer therapy evaluation program (NCI-CTEP) sponsored studies. *Clin. Cancer Res.* **16**, 5557–5563 (2010).
17. Unger, J. M., Vaidya, R., Hershman, D. L., Minasian, L. M. & Fleury, M. E. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *J. Natl Cancer Inst.* **111**, 245–255 (2019).
18. Somkin, C. P. et al. Effect of medical oncologists' attitudes on accrual to clinical trials in a community setting. *J. Oncol. Pract.* **9**, e275–e283 (2013).
19. Organizational barriers to physician participation in cancer clinical trials. *PubMed* https://pubmed.ncbi.nlm.nih.gov/16044978/ (2005).
20. Benson, A. B. et al. Oncologists' reluctance to accrue patients onto clinical trials: an Illinois Cancer Center study. *J. Clin. Oncol.* **9**, 2067–2075 (1991).
21. Siminoff, L. A., Zhang, A., Colabianchi, N., Sturm, C. M. S. & Shen, Q. Factors that predict the referral of breast cancer patients onto clinical trials by their surgeons and medical oncologists. *J. Clin. Oncol.* **18**, 1203–1211 (2000).
22. Karystianis, G., Florez-Vargas, O., Butler, T. & Nenadic, G. A rule-based approach to identify patient eligibility criteria for clinical trials from narrative longitudinal records. *JAMIA Open* **2**, 521–527 (2019).
23. Tu, S. W. et al. A practical method for transforming free-text eligibility criteria into computable criteria. *J. Biomed. Inform.* **44**, 239–250 (2010).
24. Wong, C. et al. Scaling clinical trial matching using large language models: a case study in oncology. *arXiv.org* https://arxiv.org/abs/2308.02180 (2023).
25. Haddad, T. et al. Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: Retrospective pilot study. *JMIR Med. Inform.* **9**, e27767 (2021).
26. Brown, T. B. et al. Language Models are Few-Shot Learners. *arXiv.org* https://arxiv.org/abs/2005.14165 (2020).
27. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2019).
28. Hamer, D. M. D., Schoor, P., Polak, T. B. & Kapitan, D. Improving patient pre-screening for clinical trials: assisting physicians with large language models. *arXiv.org* https://arxiv.org/abs/2304.07396 (2023).
29. Jin, Q. et al. Matching patients to clinical trials with large language models. *Nat. Commun.* **15**, 9074 (2024).
30. Nievas, M., Basu, A., Wang, Y. & Singh, H. Distilling large language models for matching patients to clinical trials. *J. American Med. Inform. Assoc.* **31**, 1953–1963 (2024).
31. Datta, S. et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J. Am. Med. Inform. Assoc.* **31**, 375–385 (2023).
32. Jiang, A. Q. et al. Mistral 7B. *arXiv.org* https://arxiv.org/abs/2310.06825 (2023).
33. Minaee, S. et al. Large Language Models: a survey. *arXiv.org* https://arxiv.org/abs/2402.06196 (2024).
34. Touvron, H. et al. LLAMA: Open and efficient foundation language models. *arXiv.org* https://arxiv.org/abs/2302.13971 (2023).
35. Ouyang, L. et al. Training language models to follow instructions with human feedback. *arXiv.org* https://arxiv.org/abs/2203.02155 (2022).
36. OpenAI et al. GPT-4 Technical Report. *arXiv.org* https://arxiv.org/abs/2303.08774 (2023).
37. Ye, J. et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv.org* https://arxiv.org/abs/2303.10420 (2023).
38. Wu, T. et al. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. *arXiv.org* https://arxiv.org/abs/2203.06566 (2022).

39. Kang, K., Wallace, E., Tomlin, C., Kumar, A. & Levine, S. Unfamiliar finetuning examples control how language models hallucinate. *arXiv.org* https://arxiv.org/abs/2403.05612 (2024).
40. An OCR Post-Correction approach using deep learning for processing medical reports. *IEEE Journals & Magazine | IEEE Xplore* https://ieeexplore.ieee.org/document/9448197 (2022).
41. Kaufmann, B. et al. Validation of a Zero-Shot learning natural language processing tool for data abstraction from unstructured healthcare data. *arXiv.org* https://arxiv.org/abs/2308.00107 (2023).
42. Zhou, S., Wang, N., Wang, L., Liu, H. & Zhang, R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J. Am. Med. Inform. Assoc.* **29**, 1208–1216 (2022).
43. Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *arXiv.org* https://arxiv.org/abs/2005.11401 (2020).
44. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. *arXiv.org* https://arxiv.org/abs/2305.18290 (2023).
45. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 8–9 (2019).
46. Liu, P. et al. Pre-train, Prompt, and Predict: A systematic survey of prompting methods in natural language processing. *arXiv.org* https://arxiv.org/abs/2107.13586 (2021).
47. Wu, T., Terry, M. & Cai, C. J. AI chains: Transparent and controllable Human-AI interaction by chaining large language model prompts. *arXiv.org* https://arxiv.org/abs/2110.01691 (2021).
48. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLORA: efficient finetuning of quantized LLMS. *arXiv.org* https://arxiv.org/abs/2305.14314 (2023).
49. Hu, E. J. et al. LORA: Low-Rank adaptation of Large Language Models. *arXiv.org* https://arxiv.org/abs/2106.09685 (2021).

Acknowledgements

A.A. and M.S. received funding from Helmholtz to carry out this study.

Author contributions

A.A. conceptualized, supervised and designed the study. N.A. did the formal analysis and built the models. M.S. helped design prompts and LLM workflows. M.S. and R.W. provided software expertise on LLMs and user interface. M.S. and R.W. provided the cloud infrastructure to carry out part of the analysis. N.A. prepared the initial draft with input from A.A. All authors edited the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01673-4.

Correspondence and requests for materials should be addressed to Altuna Akalin.