Am I eligible? Natural Language Inference for Clinical Trial Patient Recruitment: the Patient's Point of View

Mathilde Aguiar, Pierre Zweigenbaum, Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
91405, Orsay, France
first.last@lisn.fr

Abstract

Recruiting patients to participate in clinical trials can be challenging and time-consuming. Usually, participation in a clinical trial is initiated by a healthcare professional and proposed to the patient. Promoting clinical trials directly to patients via online recruitment might help to reach them more efficiently. In this study, we address the case where a patient is initiating their own recruitment process and wants to determine whether they are eligible for a given clinical trial, using their own language to describe their medical profile. To study whether this creates difficulties in the patient-trial matching process, we design a new dataset and task, Natural Language Inference for Patient Recruitment (NLI4PR), in which patient-language profiles must be matched to clinical trials. We create it by adapting the TREC 2022 Clinical Trial Track dataset, which provides patients' medical profiles, and rephrasing them manually using patient language. We also use the associated clinical trial reports where the patients are either eligible or excluded. We prompt several open-source Large Language Models on our task and achieve from 56.5 to 71.8 of F1 score using patient language, against 64.7 to 73.1 for the same task using medical language. When using patient language, we observe only a small loss in performance for the best model, suggesting that having the patient as a starting point could be adopted to help recruit patients for clinical trials. The corpus and code bases are all freely available on our Github[1] and HuggingFace[2] repositories.

1 Introduction

Many efforts have been made to develop methods based on Natural Language Processing (NLP) to solve ongoing challenges in healthcare. These studies are targeting either medical professionals or patients. However, patients and medical professionals use different kinds of language. A system trained and designed on medical language might, therefore, fail when used with patient language.

Before releasing a new medicine on the market, clinical trials must be performed and recruit several cohorts of patients with profiles that comply with the inclusion and exclusion criteria of the trial. Recruiting patients can be challenging and costly, especially for studies focusing on certain diseases or targeting a specific population, e.g. a study targeting young children with a rare disease. This can cause major delays for the trial: in 2012, 80% of trials in the US were aborted because of the lack of fitting participants (Johnson, 2015). While enrollment into the trial is usually proposed by a medical practitioner to an already known patient, new online recruitment solutions[3] are promoting trials directly to patients who might not be familiar with clinical trials. These solutions could help speed up and reduce the cost of the patient recruitment process, allowing to recruit hard-to-reach populations, and target underrepresented populations (Brøgger-Mikkelsen et al., 2020).

In this study, we focus on patient recruitment for clinical trials by adopting the patient's point of view, thus using patient language (PL) to describe the patient's medical profile. To enable the research community to explore this setting, we design a novel task, Natural Language Inference for Patient Recruitment (NLI4PR). We create a dataset derived from patient profiles from the shared task TREC 2022 Clinical Trial Track (TREC-CT 2022) (Roberts et al., 2022) and clinical trials' eligibility criteria for which the patient would be eligible or excluded. We frame the recruitment task into a Natural Language Inference (NLI) task. Our aim is to evaluate models' ability to infer from a given premise (the trial's eligibility criteria) whether the

[1] https://github.com/CTInfer/NLI4PR

[2] https://huggingface.co/datasets/Mathilde/NLI4PR

[3] See for instance Klineo or DigitalECMT.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 243–259

statement (the patient's medical profile) is entailed or contradicts the given premise. If there is an entailment, the patient can be enrolled in the trial; otherwise, the patient does not match the trial's eligibility criteria. Since Large Language Models (LLMs) have demonstrated competitive results in similar shared tasks (Jullien et al., 2023b, 2024), we evaluate how they fare on the present new task. Our contributions are the following:

- Using patient language instead of medical doctor's language to describe the patient's medical profile and perform the patient-trial matching task.
- Creating a new dataset and task, NLI4PR, aiming at matching patients to clinical trials using patient language.
- Evaluating and comparing Large Language Models on the patient-matching task using medical and patient language.

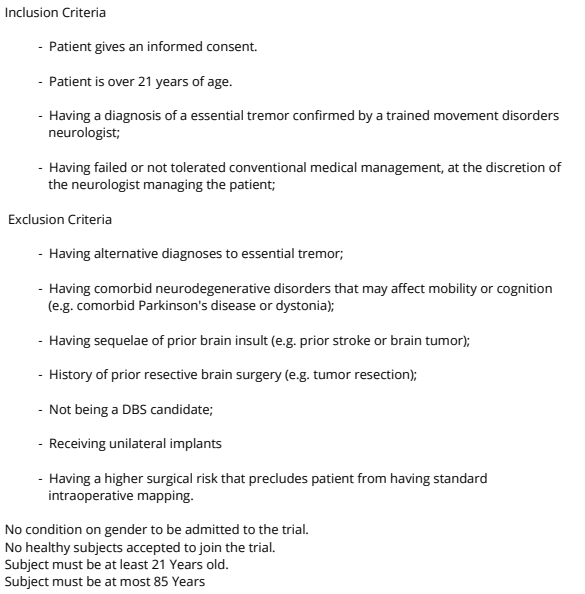
2 Related Work

2.1 Natural Language Processing for recruiting patients for clinical trials

Recruiting patients for clinical trials can be challenging and time-consuming. This is one of the main causes for trials to fail (Kantor and Morzy, 2024). Trials target a certain population, defined through the eligibility criteria designed at the beginning of the study (see Fig. 1). These criteria are expressed as free text in the Clinical Trial Reports (CTRs). The traditional way of promoting trials to patients was made directly by healthcare professionals to known patients that might fit the trial. However, this involves a long manual review of patient profiles, which can also lead to screening errors. Thanks to the digitization of patients' medical records, called electronic health records (EHRs), systems based on NLP (Ghosh et al., 2024; Murcia et al., 2024) aimed at providing support to solve the patient-trial matching task. These systems allow the automatic review of patients' profiles and trial eligibility criteria. They can either follow the trial-to-patients paradigm (for a given trial, the system suggests several patient profiles) or patient-to-trials (for a given patient, the system proposes several trials).

The TREC-CT 2021 (Soboroff, 2022), 2022 (Roberts et al., 2022), and 2023 (Soboroff, 2024) shared tasks promote the development of NLP-based systems that address the patient-trial matching problem. These tasks provide patient topics, which are a short description of a patient's medical profile, in free-text form in the 2021 and 2022 editions or as structured text (as questionnaires) in the 2023 edition. The goal is to provide for each patient a ranked list of CTRs for which the patient would be eligible, excluded, or not relevant. With the recent advent of Large Language Models, methods using these models (Jin et al., 2024; Nievas et al., 2024; Wornow et al., 2025) have been developed to perform the patient-trial matching. These methods have demonstrated competitive results compared to previous methods based on Masked Language Models.

Natural Language Inference is a task that aims to determine whether a statement can be inferred from a given premise. This task is quite challenging as it requires different kinds of knowledge, and involves finding evidence in the given pieces of text and confronting these pieces of evidence all together in order to conclude if there is an entailment or a contradiction. The NLI4CT task (Jullien et al., 2023a) uses NLI on clinical trials for various applications. Clinical trials are used as NLI premises, and statements have been manually generated. One of the targeted applications is patient recruitment, but the statements are using doctor's medical language. NLI4CT offers a benchmark to evaluate models on their common-sense, numeri-


Inclusion Criteria

- Patient gives an informed consent.
- Patient is over 21 years of age.
- Having a diagnosis of a essential tremor confirmed by a trained movement disorders neurologist;
- Having failed or not tolerated conventional medical management, at the discretion of the neurologist managing the patient;

Exclusion Criteria

- Having alternative diagnoses to essential tremor;
- Having comorbid neurodegenerative disorders that may affect mobility or cognition (e.g. comorbid Parkinson's disease or dystonia);
- Having sequelae of prior brain insult (e.g. prior stroke or brain tumor);
- History of prior resective brain surgery (e.g. tumor resection);
- Not being a DBS candidate;
- Receiving unilateral implants
- Having a higher surgical risk that precludes patient from having standard intraoperative mapping.

No condition on gender to be admitted to the trial.
No healthy subjects accepted to join the trial.
Subject must be at least 21 Years old.
Subject must be at most 85 Years

Figure 1: Example of a CTR's eligibility criteria. Taken from NCT04581941, available on clinicaltrials.gov

cal, and biomedical abilities applied to the clinical trial domain. Besides, these premises not only consist of the eligibility criteria section, but also, in some instances, they consist of result, intervention, or adverse events sections. Systems like that of Zhang et al. (2020) use NLI to model the patient recruitment task, using a fragment of the patient's EHR as the statement and the trial's eligibility as a premise. All of these approaches are based on the patient's EHR or other medical documents, and never on the patient's medical profile using patient language in a free-text form. Our task is the first to propose an approach using patient language to match patients to clinical trials.

2.2 Processing Patient Language

According to Seiffe et al. (2020), a medical, technical term is either used by a physician or comes from Latin or Greek; a lay term is a term that can easily be understood by patients or is based on everyday language. Here, we define patient language (PL) as the expressions, terms, and formulations expressed in natural language that patients use to talk about their health and any health-related topic, which is broader than the definition proposed by Seiffe et al. (2020). Processing such language poses different challenges from those in traditional medical texts. While medical language uses precise terms to describe a concept, patients will use less precise expressions due to a lower level of medical knowledge, which often causes the patient's text to be inaccurate and also longer compared to one written by a healthcare professional. The patient's medical language is also highly influenced by their health literacy, often depending on their social background, age, and education level. PL also often conveys a load of negative emotions, such as fear, worry, anger, or anxiety (Anderson et al., 2008). In written text, typos and misspellings can also occur. Lay terms (or plain English) bridge the gap between the jargon of a complex domain and "everyday life" language. In the medical domain, they allow patients to make informed decisions, as for instance in the README dataset (Yao et al., 2024), which aims to provide patients with definitions for technical terms found in their EHRs in lay terms. Medical to lay term glossaries have been created, such as that from the University of Michigan[4] or that of the European Medicines Agency.[5] MedlinePlus[6] (Miller et al., 2000) also provides a glossary of medical concepts explained using lay terms and other synonyms. The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a set of health, biomedical-related vocabularies and standards for the medical domain. In particular the Consumer Health Vocabulary (CHV) provides some medical term to lay language mappings.

Usually, the goal behind the use of lay language is to summarize (Giannouris et al., 2024) or simplify (Attal et al., 2023) the original technical text. Giannouris et al. (2024) summarized clinical trial reports with lay language to make them more easily accessible to non-experts but did not address the recruitment process. In this paper, we do not try to summarize or simplify the patient profile but we use lay terms to study whether patient language is processed as well as medical technical language in clinical trial matching, so that patients themselves could be the starting point of recruitment for clinical trials.

3 Corpus Creation

To the best of our knowledge, no dataset exists in which lay language descriptions of patient profiles are used to identify matching clinical trials. We therefore decided to create one. To do so, we employ a 3-step process: (i) we start from TREC-CT 2022's patient topics, which express patient profiles in free-text, medical language. We then rephrase these topics using patient language (see Sec. 3.1). (ii) We collect the CTRs labeled as *eligible* and *excluded* in TREC-CT. Finally, (iii) we convert the task into a 2-way NLI classification task (see Sec. 3.2). Figure 2 summarizes the process.

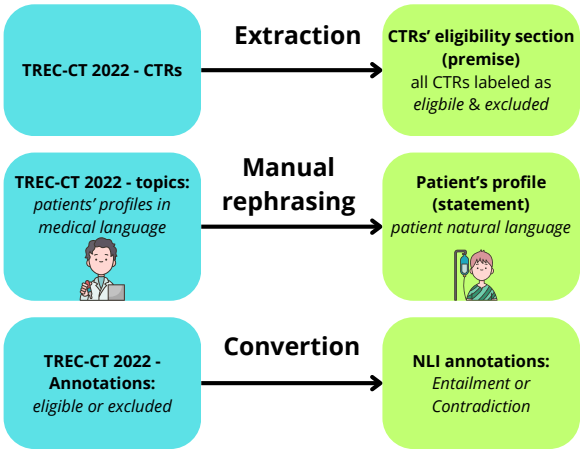


Figure 2: Corpus creation steps

[4] https://medicaldictionary.lib.umich.edu/

[5] https://www.ema.europa.eu/en/documents/other/ema-medical-terms-simplifier_en.pdf

[6] https://medlineplus.gov/

3.1 Rephrasing into Patient Language

We used the 50 TREC-CT 2022 patient topics that describe the patient's last medical visit (emergency room, clinic, or primary care physician). Topics are written using medical language. Following MIMIC-IV's (Johnson et al., 2023) descriptors, the patient topics contain the following information: chief complaint, history of present illness, patient demographics (age and gender), physical exams, and discharge diagnoses. Topics cover various diseases, such as genetic, endocrinal, or dermatological diseases, with patients presenting various profiles, from newborns to the elderly. To obtain PL topics, we tried two different approaches. The first consists of using Large Language models to rephrase the topics automatically. We tried with GPT-4o (OpenAI et al., 2024) and Llama3-8B-Instruct (Dubey et al., 2024) and applied a simple prompt, displayed in Appendix A. Both models seemed to grasp most of the information and adopt a patient perspective, using lay terms and the appropriate tone. However, they sometimes tended to remove quite important information (in the example displayed in Appendix A, in both cases, gender is missing). To avoid these issues, we discarded this approach and opted for the approach below.

To ensure consistency in the information contained in the topics, the first author manually rephrased the topics. This author is experienced in working on medical texts and performing annotation tasks on medical documents, but does not hold any medical degree. We estimate that this level of expertise is suitable for our task since we are trying to represent the health literacy of an average patient. To get a better grasp of different patients' writing styles, we first conducted a manual evaluation with 6 human annotators presenting various patient profiles, described in detail in Appendix B. We adopted a language similar to the one used by the participants. Apart from mapping the concepts from medical language to PL, we noticed that patients tend to use expressions representing their emotions, usually referring to fear, worry, or anxiety. We took this aspect into consideration in the rephrasing. Figure 3 gives an example of the rephrasing process:

1. Selecting the important concepts in the original patient topic (following the MIMIC-IV categories mentioned before).

2. Converting these concepts into patient language by using MedlinePlus for concepts unknown to the annotator or by using a lay-to-medical terms glossary. For each medical term, the annotator checks first MedlinePlus to understand the concept and look for lay language equivalents. They also check lay-to-medical glossaries to see other existing terms (although these glossaries often fall short for specific terms). If no equivalent was found in glossaries and MedlinePlus, the annotator paraphrases the term.

3. Styling the text using words that reflect the patient's emotions, by using adjectives that reflects fear, worry, etc. and by using exclamatory sentences. Additionally, we also tried to adjust language to the patient's age.

4. Proofreading to ensure consistency with the original topic.

To guide the rephrasing process, the annotator produced topics following this instruction (similar to the one given to the participants of the manual evaluation): "*Describe the purpose of your last doctor appointment, the tests undergone, the obtained results or diagnosis as well as your age, gender, and past medical history. All in no more than a dozen sentences.*".

Table 1 displays a small sample of reformulations of the initial medical terms. To analyze a few linguistic features of the NLI4PR dataset, we compute readability and similarity metrics. Using some of the scores of BioLaySumm 2024 (Goldsack et al., 2024), we computed BERTScore (Zhang* et al., 2020) for similarity between the patient and medical version of the topics, Flesch-Kincaid Grade Level (FKGL) (Flesch, 1948), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948) scores for readability. Tab. 2 reports the results of the different metrics.

The patient and medical topics still keep similar features with a high BERTScore of 89.5%. For the patient language topics, FKGL and DCRS scores both respectively indicate that a 11-17 years old student and a $11-12^{th}$ grade student could understand the topics written in patient language. Although, the CLI measure estimates the readability to be accessible for a 5-6th grader. However the topics produced are accessible to the majority of the population and correspond to what we would expect from an adult's average health literacy. For the

Medical term	PL example	Rephrasing strategy
ALS (amyotrophic lateral sclerosis)	sclerosis	MedlinePlus + name simplification
ear discharge	fluid in my ear	MedlinePlus' description
hearing loss	I could not hear as well as I used to	Paraphrase of the symptoms
His father died suddenly at age 35.	My dad died suddenly when he was 35, so I'm kind of scared.	Add emotion (fear)
dyslipidemia	cholesterol	MedlinePlus (Alternative names section)
Allopurinol	Zyloric	MedlinePlus (Brand names section)

Table 1: Examples of medical and patient language (PL) equivalents used in our task and the corresponding rephrasing strategy employed.

Metric	Patient	Medical
BERTScore	89.5%	
FKGL	6.24	8.83
DCRS	8.13	10.89
CLI	5.88	10.76

Table 2: Similarity (BERTScore) between patient and medical versions of the topics. Readability (FKGL, CLI, and DCRS) measures for patients vs medical versions of the topics.

medical version of the topics, the scores are higher (2.5 points more for FKGL and DCRS) and almost doubled for CLI, bringing the readability level to a 10-11th grade student. To see if the proportion of medical terms is more important in the medical version of the topics, we used QuickUMLS (Soldaini and Goharian, 2016) to extract medical concepts indexed in UMLS. For 92% of the topics, the medical language version contains more terms taken from the UMLS than its patient language equivalent. On average, patient language topics contain 21 terms taken from the UMLS versus 25 for the topic's medical version. Although the average length of a patient language topic is 116 words versus 98 for medical language. This suggests that patient indeed tend to use paraphrase to refer to medical terms.

3.2 Conversion into an NLI task

TREC-CT's original aim is to rank a large number of CTRs in terms of eligibility for a given patient topic. There are 3 ranking levels: *eligible* (the patient described in the topic can take part in the trial), *not relevant* (the trial's eligibility criteria do not seem relevant for the patient described in the topic and there is not enough information to qualify for the trial), and *excluded* (the patient described in the topic does not match the trial's eligibility criteria). Natural Language Inference aims to determine whether a statement entails a given premise, thus in our context, whether the patient topic (statement) entails the trial's eligibility criteria (premise). We map TREC-CT's annotations to NLI annotations: *eligible* is mapped to *entailment*, and *excluded* to *contradiction*. We did not map the instances labeled as *not relevant* to *neutral* as TREC-CT's goal was to rank trials by relevance and not to test patients' eligibility. We describe the internal inference process that should be employed in order to predict the right label. The patient topic Pat has a set of n features f (age, disease, gender, etc.): $Pat = \{f_1, ..., f_n\}$. The eligibility section is composed of m inclusion criteria Inc and k exclusion criteria Exc: $Inc = \{i_1, ..., i_m\}$ and $Exc = \{e_1, ..., e_k\}$. We define the inference relationship between the statement Pat and the premise Inc, Exc as:

$$\forall i \in Inc, \exists f \in Pat; entail(i, f) \quad (1)$$

$$\forall e \in Exc, \forall f \in Pat, contradict(e, f) \quad (2)$$

$$(1) \wedge (2) \Rightarrow Entailment \quad (3)$$

where *contradiction* holds if *entailment* does not. In other words, the model has to infer that for every feature f of a patient, it entails with every inclusion criteria and that it contradicts with every exclusion criteria, for the model to output *Entailment* as the final prediction.

For each topic, we extract all the CTRs labeled as *excluded* and *eligible* in TREC-CT, resulting in, for each patient topic, several *(patient topic, CTR)* pairs labeled either with *entailment* or *contradiction*. Our resulting task is a 2-way NLI classification task.

3.3 Resulting dataset

The resulting dataset consists of 7007 instances, split into training, development, and test sets (representing 70%, 10%, and 20% of the whole dataset, respectively). 3939 are labeled as *Entailment* and

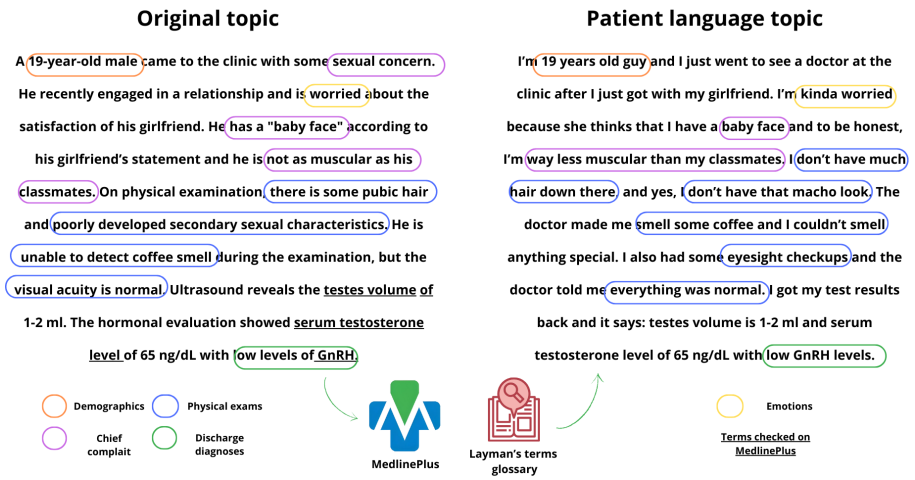


Figure 3: Rephrasing of a patient topic, following MIMIC-IV categories and using MedlinePlus.

3068 as *Contradiction*. Table 3 displays the number of instances per split and the label distribution. We provide two kinds of statements: *statement_medical*, which is the original TREC-CT's patient topic (in medical language), and *statement_pl*, which is the PL rephrased topic. The *premise* field is composed of the extracted eligibility section of the CTR. Additionally, we provide the study's title in the *NCT_title* field and its corresponding id in *NCT_id*. As in Jullien et al. (2023a), our dataset involves several challenges: biomedical reasoning, numerical reasoning, and commonsense reasoning. Appendix C displays more statistics. The dataset is freely available on HuggingFace.[7]

Split	# Entailment	# Contradiction
Train	2757	2147
Dev	295	230
Test	887	691

Table 3: Distribution of *Entailment* and *Contradiction* instances in the dataset splits.

4 Methods

Using this new dataset, we perform initial experiments to evaluate the ability of LLMs to solve the task with lay- vs. medical-language patient profiles.

We prompt four open-source Large Language Models using two prompting templates. The first template, *vanilla*, is made of a simple instruction described in Figure 4a; the second template, *persona*, aims at impersonating the model into a medical practitioner reviewing patient profiles and deciding whether they can participate into the trial or not (see Figure 4b).

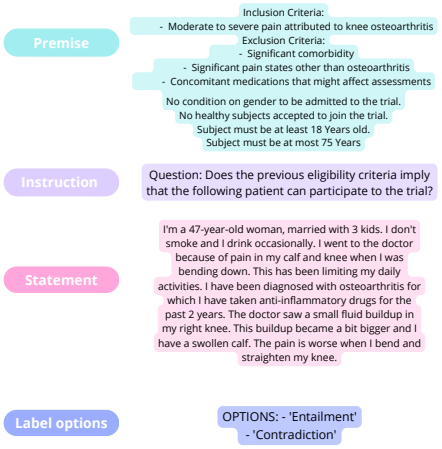
The templates are structured as follows: the *premise*, which is the eligibility criteria section of the clinical trial, the instruction, the *statement*, which is the patient profile, either expressed in PL or using medical language, and finally we provide the possible answers, *Entailment* or *Contradiction*. We perform all the experiments in a zero-shot setting, meaning that we do not show any previous demonstration to the model.

We use models that previously achieved competitive results on the similar SemEval task of NLI4CT:

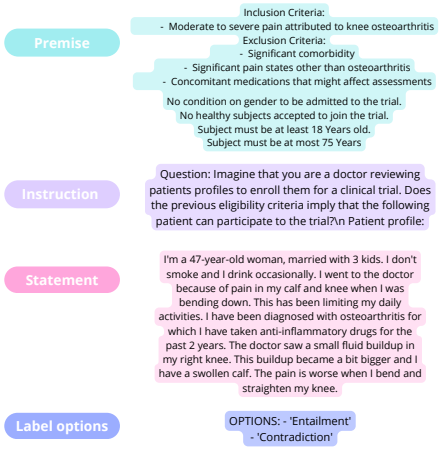
- Flan-T5-XXL (Chung et al., 2022), an 11 billion parameters instruction-tuned sequence-to-sequence model.
- Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al., 2024), instruction-tuned decoder-only models respectively with 7 and 14 billion parameters.
- Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), a 45 billion parameters decoder-only model pretrained using a mixture of experts approach.

These models are all pretrained on general domain data. As in Jullien et al. (2023a), we choose macro-F1 score as the evaluation metric. We perform the evaluation on the whole test set. We use a temperature of 0.7, a top_p of 1 and top_k of 0. For comparison, we compute the majority baseline corresponding to the case where all the predictions would be labeled as *Entailment*. Comparison is also done against a random classifier where the seed used is 42.

[7] https://huggingface.co/datasets/Mathilde/NLI4PR



(a) Example of a prompt using the *Vanilla* template



(b) Example of a prompt using the *Persona*-style template

Model	Lay-*V*	Lay-*P*	Med-*V*	Med-*P*
Majority		36.0		
Random		50.0		
Flan-T5-XXL	66.0	61.8	72.1	67.5
Qwen-7B	64.1	62.9	65.5	64.7
Qwen-14B	**71.8**	**69.8**	**73.1**	**73.7**
Mixtral-8x7B	60.7	56.5	70.8	71.2

Table 4: Macro F1 score (in %) for the different baselines, using our different prompting templates in a zero-shot setting, on the test set. *Lay* is *patient* language, *Med* is *medical doctor*'s language, *V* stands for *vanilla* template and *P* stands for *persona* template. The majority baseline is *Entailment*. Seed for the random baseline is 42.

5 Results

Table 4 displays the results obtained by the models on the two types of templates.

Qwen-14B achieves the best results for all kinds of templates, up to 37.7 points higher than the majority baseline and 23.7 for the random baseline. All models perform better on medical language than on PL. We believe this loss of performance may come in part from the lack of precision of layman terms used in PL, in comparison to medical terms that define a more precise concept. When trying to match eligibility criteria, the model might not be able to determine the patient's eligibility if in the PL statement, the concept is not precise enough. E.g., in the following example, the eligibility criteria states *"Subjects having a diagnosis of probable or definite **ALS** in accordance with the Revisited El-Escorial Criteria."*, the patient topic in medical language uses the acronym *ALS*, however in the patient topic in PL, the term used is simply *sclerosis* (see Table 1). With PL, the model cannot determine which type of sclerosis the patient is suffering from and thus might not match it to the trial.

Using a persona template did not necessarily lead to better results; Flan-T5 performed even worse when using PL. Despite being the larger model, Mixtral is the worst-performing when using PL, and in the worst case being only 6.5 points above the random baseline. In the case of Qwen, more parameters (increasing from 7B to 14B) improved performance, with a gain of up to 9 points for the Med-P template.

6 Error Analysis

Medical vs Patient Language We examined on which patient topic models tend to fail, either using PL or medical language: for this purpose, we compute the misclassification rate (MCR) for each patient topic t using the predictions of each model and the gold standard:

$$MCR(t) = \frac{misclassification\_topic\_t}{total\_count\_topic\_t}$$

We compute MCR for all topics with all models' predictions across all templates, where $misclassification\_topic\_t$ is the number of misclassifications for topic t and $total\_count\_topic\_t$ the number of instances using topic t as the statement in the dataset. We derive $MCR_{pl>med}$ where the models perform better with topics using medical language than PL, and conversely $MCR_{med>pl}$ where models were better using patient language, for each patient topic n:

$$MCR_{pl-med}(t) = MCR_{pl}(t) - MCR_{med}(t)$$

$$MCR_{pl>med} = \max_{t\in[1,50]} MCR_{pl-med}(t)$$

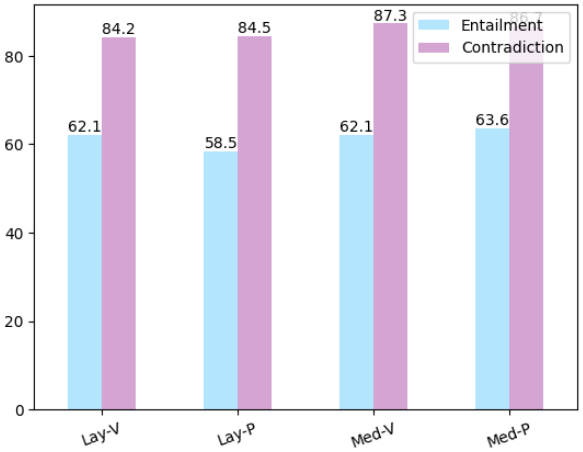
$$MCR_{med>pl} = -\min_{t\in[1,50]} MCR_{med-pl}(t)$$

Across all the models, the patient topic occurring the more often for $MCR_{pl>med}$ is patient #21, and the one for $MCR_{med>pl}$ is patient #30. Appendix D displays both patients' profiles. The descriptions of patient #21 in medical language and PL are similar in terms of demographics, chief complaint, and physical exams. However, for the discharge diagnosis, medical language mentions *ALS* while PL mentions *sclerosis* (see Tab. 1), thus not mentioning the specific kind of sclerosis diagnosed. For patient #30, the physical exam observations have been greatly simplified in the PL version. Otherwise, information remains consistent with the medical language version.

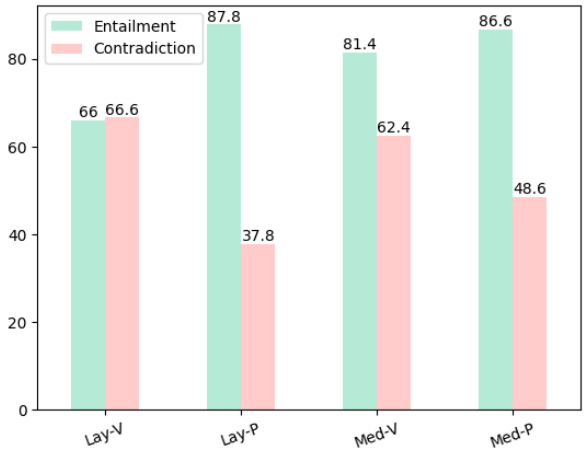
We quickly investigate if these differences can be the reason for misclassification. We allow Qwen-14B to output a longer sequence of tokens when prompted with a single example of patient #21 and #30 (see Appendix E). Qwen provides a brief explanation of the reason for its prediction. We compare the justifications given for the *Lay-V* and *Med-V* prompts.

For #21, Qwen predicted the right label (*Entailment*) for *Med-V* and the wrong label for *Lay-V*. The misclassification comes from the case depicted in Sec. 5. Qwen mentioned that *sclerosis* does not necessarily involve an ALS, which is technically true. PL lacks precision, which can lead to misclassification, whereas the model can predict the right label with medical language for the same case.

For #30, Qwen predicted the right label (*Entailment*) for *Lay-V* and the wrong label for *Med-V*. The patient topic describes a woman suffering from osteoarthritis. In order to solve the inference, the model has to perform numerical inference to determine if her age fits the age range of the inclusion criteria, check that the diagnosis of osteoarthritis fits with the inclusion and exclusion criteria and that the patient does not suffer from other disorders. For *Lay-V* the model reports that it compared the age range, the osteoarthritis diagnostic with the eligibility criteria. For *Med-V* the model got misled by one of the symptoms and inferred another potential disease, that could fall under one exclusion criterion. In this case, having more information that was not directly linked to the criteria confused the model and led to a wrong prediction.



(a) Entailment and Contradiction accuracy for Qwen-14B's predictions.



(b) Entailment and Contradiction accuracy for Flan-T5-XXL's predictions.

Figure 5: *Lay* is *patient* language, *Med* is *medical* doctor's language, *V* stands for *vanilla* prompt and *P* stands for *persona* prompt.

Which is harder, *Entailment* or *Contradiction*? We compute the accuracy per label for the two best-performing models, Qwen-14B and Flan-T5-XXL (see Figure 5). Qwen is performing up to 26 points better on *Contradiction* than on *Entailment*. This behavior is consistent with all the types of templates. Surprisingly, Flan-T5 obtains up to 50 points more in predicting *Entailment* than *Contradiction*, and this observation applies to all templates except *Lay-V*. Predicting *Contradiction* seems rather simple compared to predicting *Entailment*. Since a patient would not be eligible for a clinical trial if their characteristics do not comply with at least one of the exclusion criteria, this would

lead to a direct assertion of the *Contradiction* label. Meanwhile, for *Entailment*, the model has to go through all the patient's features and compare them to all the inclusion and exclusion criteria, which involves more knowledge and computations.

7 Future Work and potential applications

One direction for future work would be to fine-tune the models (using our training and development sets) to see if it would improve performance. Systematically evaluating models' explanations would also allow to determine if the model is predicting the right label for the right reason and, hence, detecting the right pieces of evidence in the text to make its prediction. This evaluation could be done using the LLM-as-a-judge paradigm (Zheng et al., 2023), where one or several LLMs could evaluate if the retrieved evidence and explanations are correct. Expanding the dataset with new patient profiles with various health literacy levels and diseases would also allow to evaluate the models on more diverse cases.

We hope this work can pave the way to the development of more NLP applications to promote clinical trials directly to patients, using their own language. We believe that proposing these kinds of interfaces would allow to reduce the recruitment workload and to promote trials to a wider population.

8 Conclusion

In this study, we present a novel task, Natural Language Inference for Patient Recruitment (NLI4PR), that aims to use patient language to match patients to clinical trials. The patient-to-trial matching is usually done using a description of the patient in doctor's medical language. Here, we adopt another approach where the patient describes their own profile using their own language. Patient language presents major differences compared to doctor's medical language due to the patient's limited health literacy. We evaluated the ability of several open Large Language Models to deal with patient language and compare it to the use of medical language. We frame the task as a Natural Language Inference task. For this, we create a new dataset derived from the patient profiles provided by TREC-CT 2022, and the clinical trials ranked as *eligible* and *excluded* in TREC-CT 2022.

We found that all models obtained an F1 score much higher than the majority baseline on our test set, using medical language but also using patient language. Models struggled more with patient language than with medical language, however the gap between the two settings was rather low. We found that this gap in performance is mainly coming from the loss of precision in the terms used by patients compared to medical professionals. We observe that most of the errors are coming from issues in enforcing the inclusion and exclusion criteria rather than the difference between the patient language and the doctors' medical language.

9 Limitations

The patient topics have been built to keep important patient information. We hypothesize that in a real-world scenario, a patient describing their medical history and condition might miss some elements, making the task even more challenging. In addition, despite the effort made to diversify the phrasings, the rephrasing has been done by a single annotator, which might limit diversity. An improved approach would be to ask a diverse pool of patients to produce their own statements to represent differences in pathology, social background, levels of health literacy, and phrasing.

Since we perform a 2-way classification and we did not consider the cases labeled as *not relevant* from TREC-CT, we cannot directly compare our results with those of TREC-CT participants.

10 Ethical Considerations

The patient profiles are taken from the original TREC-CT 2022 shared task. They do not contain any element or piece of information that could lead to identification of any individual. The rephrased version using layman's terms does not contain any personal information either. Clinical trials are extracted and processed from clinicaltrials.gov. This resource is freely available, provided by the National Library of Medicine, and is an official U.S. Department of Health and Human Services website.

Acknowledgments

This work received funding from the CNRS through grant 80|PRIME and the French "Agence Nationale pour la Recherche" under grant agreement ANR-22-CPJ1-0087-01. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015277 made by GENCI.

References

Wendy G. Anderson, Stewart C. Alexander, Keri L. Rodriguez, Amy S. Jeffreys, Maren K. Olsen, Kathryn I. Pollak, James A. Tulsky, and Robert M. Arnold. 2008. "What concerns me is..." Expression of emotion by advanced cancer patients during outpatient visits. *Supportive Care in Cancer*, 16:803–811.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1).

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Mette Brøgger-Mikkelsen, Zarqa Ali, John R Zibert, Anders Daniel Andersen, and Simon Francis Thomsen. 2020. Online patient recruitment in clinical trials: Systematic review and meta-analysis. *J Med Internet Res*, 22(11):e22179.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Meri Coleman and T. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Satanu Ghosh, Hassan Mohammed Abushukair, Arjun Ganesan, Chongle Pan, Abdul Rafeh Naqash, and Kun Lu. 2024. Harnessing explainable artificial intelligence for patient-to-clinical-trial matching: A proof-of-concept pilot study using phase I oncology trials. *PLOS ONE*, 19(10):1–14.

Polydoros Giannouris, Theodoros Myridis, Tatiana Passali, and Grigorios Tsoumakas. 2024. Plain language summarization of clinical trials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 60–67, Torino, Italia. ELRA and ICCL.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang et al. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10.

Otis Johnson. 2015. An evidence-based approach to conducting clinical trial feasibility assessments. *Clinical investigation*, 5:491–499.

Mael Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962, Mexico City, Mexico. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Klaudia Kantor and Mikołaj Morzy. 2024. Machine learning and natural language processing in clinical trial eligibility criteria parsing: a scoping review. *Drug Discovery Today*, 29(10):104139.

Naomi Miller, E.M. Lacroix, and Joyce Backus. 2000. MEDLINEplus: Building and maintaining the national library of medicine's consumer health web service. *Bulletin of the Medical Library Association*, 88:11–7.

Victor Murcia, Vinod Aggarwal, Nikhil Pesaladinne, Ram Thammineni, Nhan Do, Gil Alterovitz, and Rafael Fricks. 2024. Automating clinical trial matches via natural language processing of synthetic electronic health records and clinical trial eligibility criteria. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2024:125–134.

Mauro Nievas, Aditya Basu, Yanshan Wang, and Hrituraj Singh. 2024. Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*, 31(9):1953–1963.

OpenAI et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2022. Overview of the TREC 2022 clinical trials track. In *Text Retrieval Conference*.

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones - resources for medical laymen to technical language and vice versa. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6185–6192, Marseille, France. European Language Resources Association.

Ian Soboroff. 2022. Overview of TREC 2021. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD.

Ian Soboroff. 2024. Overview of TREC 2023. Special Publication (NIST SP), National Institute of Standards and Technology.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *Proceedings of the MedIR workshop, SIGIR*, Pisa, Italy.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle A. Jindal, Kenneth W. Mahaffey, and Nigam H. Shah. 2025. Zero-shot clinical trial patient matching with LLMs. *NEJM AI*, 2(1):AIcs2400360.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. README: Bridging medical jargon and lay understanding for patient education through data-centric NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Xingyao Zhang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2020. DeepEnroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1029–1037, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Rephrasing patient topics using Large Language Models

Table 5 displays an example of the different outputs in PL obtained. The prompt used with the Large Language Models is the following:

> Rephrase the following patient profile like it would be expressed by the patient themselves: *original_topic*

Source	Output text
Original topic	A 19-year-old male came to clinic with some sexual concern. He recently engaged in a relationship and is worried about the satisfaction of his girlfriend. He has a "baby face" according to his girlfriend's statement and he is not as muscular as his classmates. On physical examination, there is some pubic hair and poorly developed secondary sexual characteristics. He is unable to detect coffee smell during the examination, but the visual acuity is normal. Ultrasound reveals the testes volume of 1-2 ml. The hormonal evaluation showed serum testosterone level of 65 ng/dL with low levels of GnRH.
ChatGPT	I'm 19, and I recently started a relationship. I've been feeling a bit worried about whether I'm making my girlfriend happy. She mentioned that I have a 'baby face,' and I've noticed I'm not as muscular as most of my classmates. I also can't smell coffee, which I realized during this visit. I do have some pubic hair, but I think I might be behind in developing other features that guys my age have. The doctor checked me out, and they said my testosterone levels are low, and my testicles are smaller than normal. They also mentioned something about low levels of another hormone, GnRH.
Llama-3-8B-Instruct	I'm 19 years old and I've been dating someone for a little while now. To be honest, I'm really worried that I'm not good enough for her. She says I have a 'baby face' and I'm not nearly as muscular as my friends. I feel like I don't measure up. When we're together, she sometimes mentions how unsatisfied she is with our intimate moments, which makes me even more anxious. Physically, I notice that I don't have much body hair or muscle development compared to others my age. During our last exam, I couldn't even smell the coffee because of my poor sense of smell! But at least my eyesight seems okay. I also just had an ultrasound done on my testicles, and the doctor told me they're smaller than usual. And when they checked my hormone levels, it turned out that my testosterone levels are super low - way lower than normal. It's making me feel really self-conscious and unsure about who I am and what I can do to make things better between me and my girlfriend.
Human	I'm 19 years old guy and I just went to see a doctor at the clinic after I just got with my girlfriend. I'm kinda worried because she thinks that I have a baby face and to be honest, I'm way less muscular than my classmates. I don't have much hair down there, and yes, I don't have that macho look. The doctor made me smell some coffee and I couldn't smell anything special. I also had some eyesight checkups and the doctor told me everything was normal. I got my test results back and it says: testes volume is 1-2 ml and serum testosterone level of 65 ng/dL with low GnRH levels.

Table 5: Outputs obtained after rephrasing the original topic into PL. Using either LLMs or a human annotator.

B Rephrasing patient topics using patients

Goal To have examples of the way patients express themselves, we asked 6 participants to produce statements using patient language. These participants are volunteers who are not authors of this paper. They all have different profiles in terms of health literacy, cultural backgrounds, educational backgrounds, and age. All the materials (persona, generated medical reports and guidelines) are available here[8].

Participants We note A_i a participant.

A1: 25-year-old Cambodian man with no medical training. He holds a Master's degree in computer science. His mother tongue is Khmer, but has a B2 English level.

A2: 69 years old French man. He has a PhD in biochemistry but no particular training in medicine. His native language is French.

A3: 60-year-old French woman. She is a secondary teacher with no special training in medicine. Her native language is French.

A4: 29 years old Greek woman. She is a PhD student majoring in NLP. She has no particular medical training. Her native language is Greek, but she has a C2 English level.

A5: 26 years old Chinese man. He is a PhD student in biology working on oncology. His native language is Chinese, but he has a B2 level in French.

A6: 25 years old French woman. She is currently pursuing her 8th year of medical training, specializing in anesthesiology and intensive care. Her native language is French.

Settings We designed our experiment in order to obtain as spontaneous as possible answers from the participants. We gave all participants a persona directly adapted from one of the TRECT-CT's patient topics. To avoid influencing the participants, the persona was built as a set of fields (age, gender, tests undergone, chief complaint, medical history, current medication) with keywords only. We also used ChatGPT to generate mock-ups of test results[9] with the results mentioned in the original TREC-CT 2022 topic. We gave the following instruction to the participants:

> Following the provided persona and test report, describe the purpose of your last doctor appointment, the tests undergone, the obtained results or diagnostic as well as your age, gender and past medical history. All in no more than a dozen of sentences.

Results **A1:** "I'm a 25 year old man. My roommate always says that mostly when we are watching TV in the afternoon, he noticed that I usually fall asleep even while we are talking. it's a bit embarrassing. I think I am not a person like that. I feel something is wrong. So I decided to meet a doctor to check about this. So the doctor asked me to do the sleep test which is called Multiple Sleep Latency Test. This test just requires me to take a bunch of short naps during the daytime and the doctor monitored my sleep. It is so surprising that the test also confirms that I quickly fall asleep and enter REM sleep. The doctor said this could be narcolepsy, which is a sleep disorder that causes excessive daytime sleepiness. I'm waiting to do an overnight sleep study and see a sleep specialist to figure out what the next steps are."

A2 (translated): "My medical check-up is not very encouraging. The blood test results are alarming. Serum antibodies are abnormally high, indicating an inflammatory state. The muscle biopsy confirms the presence of inflammation. The interpretations provided by the medical analysis laboratory suggest systemic inflammatory muscle damage. The doctor suspects dermatomyositis. However, he suggests further tests. In my quest to understand dermatomyositis, I learn that it is associated with various cancers. I think my life is about to be turned upside down if the doctor's diagnosis proves correct. Maybe I'll have to undergo chemotherapy and so on. At my age, my professional and emotional life could be seriously devalued. I'll decide after the next medical examinations."

A3 (translated): "I am a 33-year-old woman. I consulted a doctor because I'm persistently tired, have less appetite than before and have lost 4 kilos in two months. At the hospital, tests showed generalized hyperpigmentation. A blood test revealed abnormally low levels of cortisol, a sign of "primary adrenal insufficiency", a dysfunction of the glands above the kidneys that produce hormones. Could my health problem be due to my past history? I suffer from Hashimoto's disease, an autoimmune disorder caused by a thyroid disorder. I am treated

[8] https://github.com/CTInfer/NLI4PR/tree/main/human_evaluation

[9] For instance, a generated karyotype report here.

for this disease by taking levothyroxine on a regular basis. But obviously, this medication no longer seems sufficiently effective. To resolve the adrenal insufficiency I'm suffering from, and thus stimulate the function of the adrenal glands, the doctors have prescribed a glucocorticoid-based treatment, on a long-term basis but under supervision. I will need to consult my GP regularly to observe the clinical symptoms - the extent of the brown spots on my skin - and to prescribe a blood test to monitor cortisol levels and, if necessary, adjust and rebalance the treatment. I was also recommended a parallel follow-up with an endocrinologist."

A4: "I am a 67 year old woman. I went to the ophthalmologist because I was not seeing well from both of my eyes especially during the night. The physical exam showed that my pupils are normal in diameter both in the light and darkness, however the acuity test results confirmed that I have a blurry vision of 50/100 that is probably linked to cataract. I will need to consult a second ophthalmologist to confirm the diagnosis and I might need to do further lab tests."

A5 (translated): "My name is Jean Martin and I'm 52 years old. On 22/01 I had two radiology examinations: a thoracic X-ray and an oeso-gastro-duodenal transit. The purpose of these examinations was to find an explanation for my symptoms of thoracic burning and acid reflux, which have been treated piecemeal with PPIs (proton pump inhibitors = anti-acids). I have no other antecedents than my obesity, I don't smoke or drink. Dr. Dupuis, a radiologist, interpreted these examinations and concluded that I had a hiatal hernia due to stomach sliding, with no signs of complications: no ulcerations, no digestive perforation and permeability of the lower esophageal sphincter, with no visualized esophageal reflux. Treatment with ipp is indicated, as is follow-up by a specialist in gastroenterology. If the symptoms become too incapacitating, I'm advised to undergo 2nd-line laparoscopic surgery to reconstruct the stomach, which is still a major operation. I prefer to try medical treatment in 1st intention as agreed. I have been informed of the serious signs of my illness, which require me to undergo urgent appointment."

A6 (translated): "Hello, I'm currently 26 years old. I went to the clinic today because I felt down at the gym. I exercise often but it's been the 4th time that this happens. From time to time, I experience vertigo while I'm resting and I don't understand why. I exercise everyday and I don't have any other diseases for now. At the emergency room, the doctor asked me to do an X-ray and he showed me that I have a heart malformation. He told me that the volume of my left and right side are not equivalent. What's wrong? Should I stop working out?"

Conclusion Most participants followed the instructions correctly or at least partially (A2 forgot to mention their age and gender). A1, A2, A3, and A5 expressed some kind of worry regarding their symptoms and diagnosis, especially for A1 and A2, where the participants inquired about the consequences of their disease. All participants use reported speech to talk about their test results or the doctor's diagnosis. A3 and A4 directly cite some results directly taken from their test results. We observe that A2 and A3 did some supplementary research regarding their diagnosis (probably by searching their diagnosis in a search engine).

C Corpus statistics

Metric	Value
# of CTRs (whole dataset)	6649
# of CTRs (train dataset)	4713
# of CTRs (dev dataset)	523
# of CTRs (test dataset)	1564

Table 6: Dataset metrics

D Patients #21 and #30

D.1 Patient #21

Medical language: A 47-year-old man comes to the clinic for the follow up of his neuromuscular disease. He experienced gradual, progressive weakness of the left upper extremity over the last year. Over the last few months, he has also noticed weakness in the right upper extremity. BP is 120/75, PR is 80 and temperature is 37 C. Reflexes are brisk in the upper extremities, and the plantar responses are extensor. Mild gait ataxia is present. The patient is under treatment of Riluzole 50 mg BID with the diagnosis of ALS.

Patient language: I've been suffering from a neuromuscular disease for a while now, and I went to my doctor's office. I'm now a 47-year-old man and over the past year I experienced a progressive and gradual weakness of my left upper extremity, and over the past month, I also noticed a weakness over my right upper extremity. My heart rate was 120/75, and my PR was 80 with 37°C for temperature. My reflexes are not good in my upper extremities, and I have trouble with my balance. I'm also under Exservan 50 mg for my sclerosis.

D.2 Patient #30

Medical language: A 47-year-old woman comes to the office complaining of pain in the calf and knee when she bends down. The pain limits her activity. Her medical history is significant for osteoarthritis, for which she uses nonsteroidal anti-inflammatory drugs (NSAIDs) for the past two years. She is living with her husband and has 3 children. She doesn't smoke but drinks alcohol occasionally. Her vital signs are normal. On physical examination, there is a small effusion in the right knee. The effusion grew a little larger and she developed a tender swelling in the popliteal fossa and calf. Both the pain and swelling worsened as she bent and straightened her knee.

Patient language: I'm a 47-year-old woman, married with 3 kids. I don't smoke and I drink occasionally. I went to the doctor because of pain in my calf and knee when I was bending down. This has been limiting my daily activities. I have been diagnosed with osteoarthritis for which I have taken anti-inflammatory drugs for the past 2 years. The doctor saw a small fluid buildup in my right knee. This buildup became a bit bigger and I have a swollen calf. The pain is worse when I bend and straighten my knee.

E Qwen-14B prompted for explanations

E.1 Patient #21

Premise (NCT03160898): See Fig 6.

Inclusion Criteria:

- Diagnosis of familial or sporadic ALS ≤ 24 months prior to screening
- Upright Slow Vital Capacity (SVC) ≥ 60% of predicted for age, height and sex at screening
- Able to swallow tablets
- A caregiver (if one is needed)
- Able to perform reproducible pulmonary function tests
- Pre-study clinical laboratory findings within the normal range or, if outside the normal range, deemed not clinically significant by the Investigator
- Male patients who have not had a vasectomy and confirmed zero sperm count must agree after receiving the first dose of study drug until 10 weeks after the last dose to either use acceptable methods of contraception or abstain from sex
- Female patients must be post-menopausal or sterilized or must not be breastfeeding, have a negative pregnancy test, have no intention to become pregnant during the study and use acceptable methods of contraception or abstain from heterosexual intercourse from Screening until 10 weeks after last dose of study drug
- Patients must be either on riluzole for at least 30 days prior to screening or have not taken riluzole for at least 30 days prior to screening and not planning to start riluzole during the course of the study.
- Patients on edaravone must have completed at least 2 cycles of dosing with edaravone at the time of screening or have not taken edaravone for at least 30 days prior to screening and not planning to start edaravone during the course of the study.

Exclusion Criteria:

- At the time of screening, any use of non-invasive ventilation (NIV), e.g. continuous positive airway pressure [CPAP], noninvasive bi-level positive airway pressure [NPPV] or noninvasive volume ventilation [NVV] for any portion of the day, or mechanical ventilation via tracheostomy, or on any form of oxygen supplementation
- Neurological impairment due to a condition other than ALS
- Presence at screening of any medically significant cardiac, pulmonary, GI, musculoskeletal, or psychiatric illness that might interfere with the patient's ability to comply with study procedures or that might confound the interpretation of clinical safety or efficacy data
- Has taken any investigational study drug within 30 days or five half-lives of the prior agent, whichever is longer, prior to dosing
- Known to have received CK-2127107 or tirasemtiv in any previous clinical trial
- Has received or is considering receiving during the course of the study any form of stem cell therapy for the treatment of ALS
- Has received or is considering receiving during the course of the study any form of gene therapy for the treatment of ALS
- Has received or is considering obtaining during the course of the study a diaphragmatic pacing system
- History of substance abuse within the past 2 years
- Use of certain medications

No condition on gender to be admitted to the trial.
No healthy subjects accepted to join the trial.
Subject must be at least 18 Years old.
Subject must be at most 80 Years

Figure 6: Eligibility criteria from trial NCT03160898, used as the premise.

Medical answer:

"Entailment

Explanation: Based on the information provided about the 47-year-old male patient, it seems he meets some key inclusion criteria such as being diagnosed with ALS (within 24 months before screening), currently taking Riluzole which aligns with the pre-screening requirement of having been on riluzole for at least 30 days prior to screening, and being within the specified age range (at least 18 but at most 80 years). However, there isn't enough information given to determine whether his SVC is $\geq 60\%$, if he has an able caregiver, can swallow tablets, etc., nor is there sufficient detail to ascertain if he fulfills all exclusion criteria (such as absence of neurological impairment not related to ALS, no history of substance abuse within the past two years, specific medical conditions that could interfere, or recent participation in another clinical trial). The presence of mild gait ataxia mentioned does raise concern potentially regarding the exclusion criterion of "Presence at screening of any medically significant...psychiatric illness" or possibly indicating another neurological disorder; however, without explicit confirmation that this is not due to ALS, we cannot definitively rule him out based solely on this symptom alone. Similarly, without more specifics on the rest of the criteria, particularly those relating to functional capacity and absence of contraindicating health conditions, it is prudent to default towards entailment barring concrete contradictions. Therefore, the answer leans toward entailment pending verification of additional details necessary for full eligibility assessment."

PL answer:

"Contradiction

The provided patient information does not specify whether they have been diagnosed with Amyotrophic Lateral Sclerosis (ALS). Furthermore, it mentions that the patient is taking medication for sclerosis rather than riluzole which is typically used for ALS. Additionally, there is no mention of their SVC percentage which is required to meet the inclusion criteria. Hence based solely on the provided information, we cannot conclude that this patient meets all the necessary conditions to be eligible for the trial as described."

E.2 Patient #30

Premise (NCT03563690): See Fig 7.

Medical answer:

"Contradiction

The patient's presentation includes a tender swelling in the popliteal fossa and calf along with an enlarged effusion due to bending and straightening the knee, which may indicate other causes of her knee pain beyond osteoarthritis, such as possible vascular issues (e.g., deep vein thrombosis). This could fall under the exclusion criteria point number 1 ("Knee osteoarthritis patients with ... other causes of knee pain..."). Therefore, based on this information, it contradicts the full eligibility to participate merely under the inclusion criteria for osteoarthritis without further clarification or exclusion of other potential diagnoses."

Inclusion Criteria:

1. Meet the Clinical Classification Criteria for Osteoarthritis of the Knee, diagnosed by specialist.
2. Aged 40-75 years old;
3. History of chronic knee pain for more than 3 months, and complain of knee pain on most days of the past month;
4. Kellgren and Lawrence Grades 1-3;
5. With related knee pain in one or both knees rated > 4 cm on a 10 cm Visual Analog Scale(VAS) (VAS:0 to 10)
6. Who volunteer to participate and sign the consent form .

Exclusion Criteria:

1. Knee osteoarthritis patients with gout, infection, tumor, autoimmune diseases, trauma or other causes of knee pain or knee deformities;
2. Patients who have local skin damage, poor skin conditions or coagulant dysfunction and patients who are not suitable for acupuncture;
3. Patients who are accompanied with serious medical problems or mental disorders, cognitive dysfunction，disorders of shallow sensation or who cannot cooperate with the treatment;
4. Patients who have received acupuncture or needle knife in the past month;
5. Other patients who are not suitable for acupuncture treatment.

No condition on gender to be admitted to the trial.
No healthy subjects accepted to join the trial.
Subject must be at least 40 Years old.
Subject must be at most 75 Years

Figure 7: Eligibility criteria from trial NCT03563690, used as the premise.

PL Answer:

"Entailment

Based on the information provided about your condition and comparing it against the inclusion and exclusion criteria outlined in the clinical trial requirements, there is no direct contradiction found that would exclude you from participating in the trial. However, final determination should involve a thorough review by the healthcare professionals involved in the study as they will need to confirm all specific details such as severity grade, pain level on VAS, and absence of any other listed exclusions like concurrent illnesses that could disqualify participation. Given the description of meeting age range, having chronic knee pain due to confirmed osteoarthritis diagnosis, experiencing significant pain which impacts daily life, and willingness to take part (assuming voluntary agreement and signed consent), you appear to meet several key

criteria which allows for potential entailment into the trial framework."