Assessing the Difficulty of Inference Types in Natural Language Inference for Clinical Trials

Mathilde Aguiar, Pierre Zweigenbaum, Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
91400, Orsay, France
{mathilde.aguiar, pierre.zweigenbaum, nona.naderi}@lisn.fr

Abstract

Large Language Models (LLMs) achieve competitive results on Natural Language Inference when applied to clinical trials; however, it is not yet clear which type of inference LLMs perform well or poorly on. We address this by proposing new supplementary annotations for the existing NLI4CT dataset on the types of inferences observed in clinical trials. Our dataset supplements NLI4CT with a total of 1,949 new annotations using our carefully crafted guidelines for 17 types of inferences. To investigate how inference types affect the performance of LLMs, we prompt Flan-T5, Llama, Mistral, and Qwen and evaluate their performance using our newly annotated dataset. We found that logical inferences negatively affect the overall performance of Qwen3-4B, Qwen2.5-7B, and Qwen2.5-14B, whereas numerical inferences negatively affect the performance of Flan-T5-XL and Mixtral. Further analysis shows that MMed-Llama-3 struggles to understand the structure of clinical trial reports. Other parameters, such as the number of inference types involved or the section type in the premise, also influence the performance of the models. Our code and dataset are publicly available.

Keywords: Natural Language Inference, Clinical trials, Large Language Models.

1. Introduction

Large Language Models (LLMs) often obtain high accuracy or F1-score when evaluated on Natural Language Understanding tasks (Fourrier et al., 2024). They tend to outperform encoder-only architectures, such as BERT-like (Devlin et al., 2019) models, traditionally used for these discriminative tasks. Despite achieving competitive results, it is often not clear on what elements the model is relying on to make its prediction and what type of knowledge it uses. Natural Language Inference (NLI) consists of determining whether a statement can be inferred from a given premise. The possible outcomes are either *entailment*, *contradiction*, or *neutral*. This task can be quite challenging as the model needs to identify pieces of evidence in the premise and compare them to the statement to determine the inference relation. Often, the entailment relation has to be a multi-hop process, meaning that the model must perform several sub-inferences to deduce the final relation. Models need to understand links between concepts and use previously acquired knowledge, and can not only rely on words' surface forms to deduce the inference relation (Mahowald et al., 2024). We hypothesize that these sub-inferences involve different kinds of knowledge, some of which might be challenging to deal with for LLMs, especially in a domain-specific setting. We test this hypothesis in the clinical trials domain, using the NLI4CT dataset (Jullien et al., 2023a). This dataset applies NLI to clinical trials, covering cases where entailment is tested from their results, adverse events, interventions, or eligibility criteria sections.

We define 17 different inference types observed in NLI4CT, clustered in 6 categories. To do so, we manually annotate NLI4CT's test set and prompt 9 LLMs, pretrained on the medical or the general domain, and using 2 types of architectures: decoder-only models and Sequence-to-Sequence models. Our results show that Flan-T5-xl, Mixtral, Qwen3-4B, Qwen2.5-7B, and 14B are sensitive to the presence of certain inference types, like numerical and logical inferences, affecting the performance negatively when this type of inference is present.

The contributions of the paper are the following: (i) we propose an annotation scheme for inference types for clinical trials; (ii) we annotate the NLI4CT dataset according to this scheme; (iii) we analyze the performance of 9 LLMs on NLI4CT according to these inference types. (iv) finally, we perform an error analysis. Our code and dataset are publicly available on GitHub.[1]

2. Related Work

Dagan et al. (2006) define the task of NLI to be "*the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text*". This task has been studied and applied in various domains. However, there is no universal definition of the existing different inference types, since defining inference types depends on the studied domain and research purpose (Yan et al., 2020). In the following section, we propose

[1] Github repository

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5290–5300
11–16 May 2026.

a review of existing schemes for inference types definition.

2.1. Annotating Inference Types

Previous studies have defined different inference schemes specially tailored for their research focus and application domains.

Yan et al. (2020) create a schema to focus on linguistic aspects of inference types in the domain of opinion mining. The authors focused on providing annotations on whether an inference is present and, if so, which type between the following: *Logical*, *Pragmatic*, *Lexical*, *Enunciative*, and *Discursive*. Their aim was not to provide an analysis of models' performance on the NLI task but to analyze the different types of inference present in their corpus.

Instead of defining inference type schemes to predict which type of inference is present in a corpus, other studies have defined schemes to analyze the performance of language models on different inference types. These methods allow for evaluating models' reasoning steps instead of solely focusing on global models' accuracy or similar metrics, and to know which kind of knowledge the model is missing. To address this issue, a few annotation schemes (Nie et al., 2020; Joshi et al., 2020; Williams et al., 2022) have been proposed to assess a more fine-grained evaluation of models' performance on inference types (also called *reasoning types*). Nie et al. (2020) release the ANLI dataset to propose a novel challenging NLI dataset to The authors define the following reasoning types: *Numerical & Quant., Reference & Names, Standard, Lexical, Tricky, Reasoning & Facts, Quality*. Later on, Williams et al. (2022) reuse ANLI to perform a more in-depth analysis by defining 40 different fine-grained types, clustered in *Numeral, Basic, Reference, Tricky, Reasoning*, and *Imperfections* categories. TaxiNLI (Joshi et al., 2020) define a similar taxonomy with the *Linguistic, Logical, Knowledge* top-level categories. In the clinical trials domain, NLI4CT provides supplementary annotations by separating the *Numerical* inferences from the rest of the instances. In a more recent study, Jullien et al. (2025), define more types of reasoning for NLI4CT by introducing 6 types of reasoning including *expectation-driven evidence reasoning*, *clinical trial reasoning*, *lexical equivalence*, *world knowledge inference*, *domain-grounded quantitative derivation*, and *quantitative comparison*.

These schemes were mainly designed for general-domain applications, only one study (NLI4CT) addressed the clinical trials domain. In addition, these studies investigated Masked-Language Models, whereas our study focuses on Large Sequence-to-Sequence and decoder-only models. Sec. 3.2 gives a detailed comparison of these annotation schemes with the one we propose.

2.2. NLI4CT Dataset Description

The NLI4CT corpus consists of a collection of English breast cancer Clinical Trial Reports (CTRs) taken from clinicaltrials.gov. The NLI4CT task applies NLI to clinical trials with several use cases, such as checking whether a patient complies with the trial's eligibility criteria or whether a claim can be deduced from the trial's results. A premise is composed of a whole section of a CTR that can either be the *Eligibility Criteria*, the *Intervention*, the *Results*, or the *Adverse Events*. NLI4CT comprises two kinds of instances: *single*, where only one CTR is involved to perform the inference, and *comparison*, where two CTRs are needed to be compared. A statement is one or two sentences long and has been created artificially. By analyzing the statement and the premise together, a model should predict whether the statement entails or contradicts the premise. The original dataset is balanced between *Entailment* and *Contradiction* labels. The task depends on several kinds of inference, including biomedical reasoning, commonsense reasoning, and numerical reasoning.

3. Methodology

In this study, we first systematically investigate the types of inference and knowledge involved in the inference process by defining an inference types schema, annotating the NLI4CT dataset, and then examining which inference types remain challenging for LLMs. We first present our annotation process (Sec. 3.1), then provide a definition for each inference type (Sec. 3.2), and evaluate the models' performance on them (Sec. 3.3).

3.1. Annotation Process

We define and identify the different inference types in the NLI4CT dataset needed to solve the inference relation. The goal is to define labels that cover all observed inference types, are fine-grained enough, and have minimal overlap to prevent annotator disagreements. We follow an incremental process to define the different inference type labels, drawing on relevant ones from the literature to create our own labels and adapt them to our use case. We sample a first subset of 10% from NLI4CT's test set, ensuring that these 50 instances are representative of the full test set in terms of *Entailment/Contradiction* and *Single/Comparison* ratios. We use this first subset as a basis for defining our labels. We start by drawing super-classes of inference types (see Sec 3.2): *Task related expression*, *Logical*, *Generic knowledge*, *Biomedical*,

Numerical, and lastly *Typo/Error*. Then, we define more fine-grained categories (see Fig. 3) using the examples contained in the subset we sampled. These fine-grained labels are the ones used to annotate the test set. Each instance is labeled with one or more inference types whose combination is deemed necessary by the annotator to establish the inference relation. We asked three annotators, all NLP researchers and authors of this paper, to produce the annotations. After the first round of annotations, the annotators discussed the labels and revised the annotation guidelines. We conducted a second round of annotations, during which the annotators worked independently on a subset of 50 additional instances using the final annotation guidelines.[2] After the second round of independent annotations, the annotation conflicts were discussed and resolved during two additional rounds. Tab. 1 presents the final inter-annotator agreement using Krippendorff's alpha for this subset.

Pair	α
A1 vs A2	0.94
A1 vs A3	0.95
A2 vs A3	0.90
Average	**0.93**

(a) Pairwise inter-annotator agreement on inference type labels using Krippendorff's α.

Inf. type	α
TRE	1.00
Num	0.97
Bio	0.91
GK	1.00
Log	0.99
TE	1.00
Average	**0.98**

(b) Average Krippendorff's α per inference type between the 3 annotators.

Table 1: Inter-annotator agreement measures.

We achieve an average score of $\alpha = 0.93$, indicating a satisfactory agreement. Labels such as *Task-related expression* and *Typo/Error* were quite straightforward to annotate. Another difficulty is the lengthy premises and statements—averaging 125 words for premises (up to 1,388 words) and 22 words for statements—which makes them challenging to analyze and capture all subtleties. The remaining test set was annotated by annotator A1, resulting in a total of 1,949 annotations for 500 statement-premise pairs. Fig. 1 displays the corpus statistics. Fig. 2 presents the correlation matrix between the different inference types. Annotators were also asked to re-annotate the dataset in terms of *Entailment* and *Contradiction*. Annotators identified a few mistakes and achieved an α of 0.97 with the original NLI4CT's annotations. While the original test set was balanced between *Entailment* and *Contradiction*, we obtain slightly more *Contradiction* instances (253) than *Entailment* ones (247). We also provide these annotations.

[2] Annotation guide.

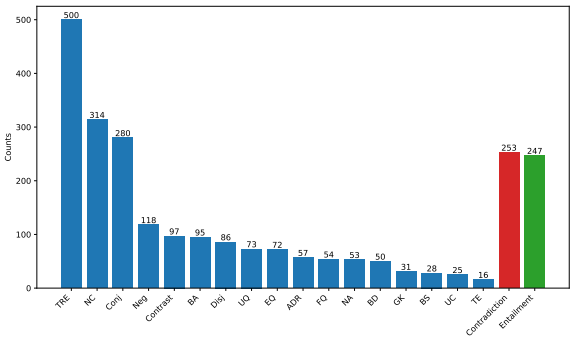


Figure 1: Count of occurrences of inference types labels and NLI labels in the test set.

3.2. Inference Types: Definitions

We define super-classes of inference types that comprise more fine-grained types. Fig. 3 illustrates the complete schema.

Task-related expression (TRE): any expression that refers to clinical trials or the structure of a clinical trial report. For instance, the different *cohorts*, *arms*, or *groups*.

E.g., **Statement (S)**: *Patients diagnosed with breast cancer may be eligible for all study groups in the primary trial and the secondary trial.*

Premise (P): *Inclusion Criteria: Healthy Participants (Part 1 only) . . . HER2-Positive Females (Parts 1 and 2) . . . Exclusion Criteria: Healthy Participants (Part 1 only) . . . HER2-Positive Females (Parts 1 and 2).* $\Rightarrow$ *part 1* and *part 2* are the *study groups*, while *inclusion* and *exclusion criteria* are determining the patient's eligibility.

Logical (super-class): connectives and quantifiers, as specified below.

Conjunction (Conj): terms in the statement expressing a logical AND.

E.g., **S**: *The primary trial intervention section requires surgical and imaging procedures.*

Disjunction (Disj): terms in the statement expressing a logical OR.

E.g., **S**: *Aes were not recorded for the primary trial or the secondary trial.*

Negation (Neg): a term in the statement expressing that some assertion is not true.

E.g., **S**: *Aes were not recorded for the primary trial or the secondary trial.*

Contrast: some terms in the statement expressing a contrast between two clauses, using words like *but*, *whereas*, *however*, etc.

E.g., **S**: *Children are not eligible for the primary trial however they are not explicitly excluded from the secondary trial.*

Universal quantifier (UQ): A term in the statement asserts that a condition applies to all elements

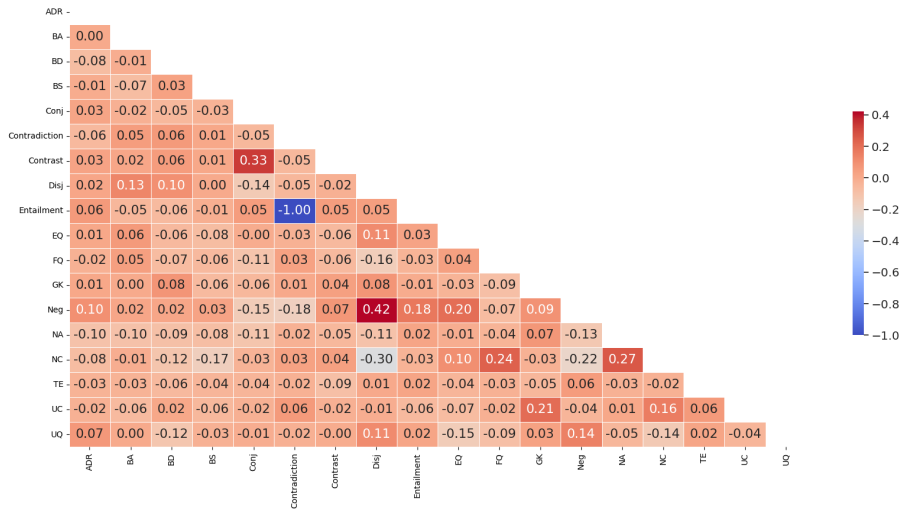


Figure 2: Correlation matrix of the inference type labels.

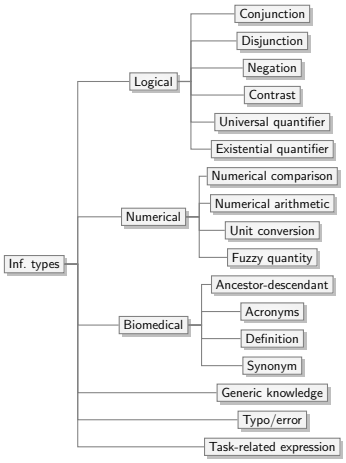


Figure 3: Our categorization of inference types in NLI4CT.

in a specified set (∀ symbol).

E.g., **S**: *Patients diagnosed with breast cancer may be eligible for all study groups in the primary trial and the secondary trial.*

Existential quantifier (EQ): a term in the statement asserts that a condition applies to at least one element in a specified set (∃ symbol).

E.g., **S**: *In total there are less cases of anemia in the primary trial than in the secondary trial.*

Numerical (super-class): fine-grained numerical inference types that include arithmetic operations, conversions, comparisons, and manipulation of numbers and quantities.

Numerical comparison (NC): when two numbers or quantities need to be compared and checked whether the number mentioned in the statement matches, is greater than, etc., the one in the premise.

E.g., **S**: *over 15% of patients in the primary trial and the secondary trial suffered from infections during the study period.*

P: Primary trial: *Infection * 1/32 (3.13%)*. Secondary trial: *Infections and infestations - Other, unspecified 1/15 (6.67%) . . . Infections and infestations - Other, unspecified 0/5 (0.00%)*.

Numerical arithmetic (NA): when an arithmetic operation (addition, subtraction, multiplication, division) is needed to solve the inference relation.

E.g., **S**: *All Patients receiving the placebo intervention in the primary trial experienced emesis.*

P: *Outcome Measurement: Number of Emesis Free Participants During the Study Period. . . . Overall Number of Participants Analyzed 20. Measure Type: Number. Unit of Measure: Participants 5.* $\Rightarrow \frac{5}{20} = 25\% \neq 100\%$, which means that not all participants experienced emesis.

Unit conversion (UC): a conversion from one unit to another is required to solve the inference relation.

E.g., **S**: *A Patient that has a primary tumour with a diameter of 33mm measured by clinical examination and echography, would be eligible for both the primary trial and the secondary trial.*

P: *Primary tumour greater than 2 cm diameter. . .* ⇒ convert the millimeters into centimeters to be able to compare the 2 measures.

Fuzzy quantity (FQ): a quantity subject to several interpretations is present in the statement, and its interpretation is needed to solve the inference relation, using words like *almost*, *several*, etc.

E.g., **S**: *Several recorded Aes in the primary trial occurred to cohort 1 patients.*

Biomedical (super-class): biomedical terms and knowledge that are needed to solve the inference relation.

Ancestor-descendant relation (ADR): terms that are either children or parents of the considered term in the premise.

E.g., **S**: *There was at least 1 recorded gastro-intestinal adverse event in the primary trial.*

P: *Vomiting 0/149 (0.00%).* ⇒ *vomiting* is a child of *gastro-intestinal adverse event*.

Biomedical acronym (BA): acronyms that shorten a biomedical term or expression. This acronym should be listed in one of the terminologies of the UMLS.

E.g., **S**: *Women suffering from both claustrophobia and IBS or not eligible for either the primary trial or the secondary trial.* ⇒ IBS = Irritable Bowel Syndrome.

Biomedical notion definition (BD): instances that require knowing the definition of a medical term and what properties it implies, and then to compare these properties with the information in the premise.

E.g., **S**: *Morbidly obese patients are eligible for the primary trial.*

P: *Have a BMI of 25 kg/m2 or greater.* ⇒ Morbidly obese patients have a BMI greater than 40 kg/m^2.

Biomedical synonym (BS): two biomedical terms, one in the statement and one in the premise, that are equivalent.

E.g., **S**: *Both interventions in the primary trial include the same dose of Paraplatin.*

P: *INTERVENTION 2: Arm 2 Taxotere/carboplatin/herceptin.* ⇒ *Paraplatin* is the brand name of *Carboplatin*.

Generic knowledge (GK): as opposed to domain-specific knowledge, notions that do not fall into the previous categories. In our case, these are often temporal words (*months*, *hours*) or concepts such as age, gender, ethnicity, nationality, etc.

E.g., **S**: *There are no conditions on mental health, bodyweight, age, Karnofsky/ECOG score or previous treatments that need to be met in order to be eligible for the primary trial.*

P: *Inclusion Criteria: Women 18 years...*

Typo/error (TE): we annotate a typo or error in the statement or the premise that would lead to a major misunderstanding of them and have an impact on the inference relation resolution of statement and premise(s).

E.g., **S**: *The the primary trial intervention section dose not describe the method of administration, dosage or cycle.* ⟹ *The* appears twice and *does* is replaced by *dose*.

Comparison to previous annotation schemes The original annotation provided with NLI4CT only separates the examples involving a numerical inference and does not provide more precise categories.[3] We computed the overlap of instances labeled as *Numerical* in NLI4CT with numerical instances using our definition, which shows that we share 83% of instances labeled as *Numerical*. This suggests that a similar definition was used in both studies. Our *Numerical* super-class aligns with those of ANLI (Nie et al., 2020) and Williams et al. (2022), and shares the *Mathematical* aspect of TaxiNLI's *Logical*. Our *Logical* largely encompasses TaxiNLI's Logical, which corresponds to ANLI's *Standard*, and Williams et al. (2022)'s *Basic* for conjunction and negation. *Generic knowledge* corresponds to some aspects of ANLI's and Williams et al. (2022)'s *Reasoning*, and TaxiNLI's *Knowledge*. *Typo/error* maps to Williams et al. (2022)'s *Imperfections*. *Biomedical* does not really map to any categories in other schemes and can be considered original since all the previous schemes are dealing with general-domain data.

3.3. Prompting Large Language Models

We select open-source LLMs, highest ranked in SemEval 2023 (Jullien et al., 2023b) and 2024 (Jullien et al., 2024). We evaluate the following models: Flan-T5-xl and xxl (Chung et al., 2024), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Llama-3.2-8B-Instruct (Dubey et al., 2024), Qwen3-4B-Instruct (Yang et al., 2025), Qwen2.5-7B and 14B-Instruct (Yang et al., 2024), all pretrained for the general domain. In addition, we use MMed-Llama-3-8B-EnIns (Qiu et al., 2024), a Llama3 model finetuned on the medical domain.

We use the same template as Kanakarajan and Sankarasubbu (2023) and added the mention "*Answer only with:*" to better constrain models to output the desired labels:

{Premise} Question: Does this imply that {Statement}? Answer only with: {options}, with *options* being *Entailment* and *Contradiction*.

We performed in a zero-shot setting, and used a temperature of 0.7, a top_p of 1.0, and top_k of 0. We set the maximum number of generated tokens to 10 and parse the produced answers using regular expressions. Accuracy is used to report the model's performance.

4. Results and Discussion

4.1. Overall Performance

In Tab. 2, we report the mean global accuracy (in column *All types*) to predict *Entailment* or *Contradiction* for all the instances of the test set of NLI4CT,

[3] Even though the paper talks about biomedical, commonsense, and numerical reasoning, the annotations only include numerical labels.

using the template described in Sec. 3.3. All the experiments are run three times, each with a different random seed (42, 55, and 3354). Qwen2.5-14B achieves the best results, with an accuracy of 0.73, followed by Qwen3-4B, with 3 accuracy points of difference but also 3.5 times fewer parameters than Qwen2.5-14B. On the other hand, MMed-Llama performs the worst with an accuracy of 0.55, which is close to random. We also compute the "hypothesis-only baseline" (Gururangan et al., 2018) by replacing the original premise with a random one. We observe that the worst-performing models perform similarly in both cases, indicating that they rely on superficial cues in the statement to predict the NLI label.

4.2. Performance per Inference Type

For each inference type, we compute the mean accuracy on the 3 runs. We define 2 subsets: the i subset, where we compute the accuracy only on instances labeled with inference type i, and the $\bar{i}$ set, where we compute the accuracy on all the instances that are not labeled with i.

To test whether the difference of performance between i and $\bar{i}$ is statistically significant, we perform a Chi-square (χ^2) test (Agresti, 2013) with a p-value threshold of 0.05. We define χ^2-*type*, where, for each model, the χ^2 is computed on one i and $\bar{i}$ inference type subsets (see Eq. 1). The null hypothesis is: "When inference type i is present, the model performs as well as when the inference type i is not present."

$$\chi^2_{type} = \sum_{k \in \{i,\bar{i}\}} \frac{(O_k - E_k)^2}{E_k} \quad (1)$$

Where O_k is k's observed frequency and E_k, k's expected frequency. For each of the three runs of a system, we compute its mean χ^2-*type* and the associated p-value. In all cases, the p-value of the three runs is on the same side of the threshold.

Tab. 2 reports these results, where highlighted cells show the pairs of results with significant differences. Globally, not all models are sensitive to the different types of inference involved. The best-performing models (scoring at least 17 points above randomness) show sensitivity to the different inference types. Even across different versions of the same type of model—for example, Flan-xl versus Flan-xxl—we do not necessarily observe a sensitivity. *Logical* and *Numerical* remain particularly challenging classes.

For Flan-T5-xl, overall performance is affected by *Numerical* inferences, with the presence of *Numerical comparison* significantly reducing performance ($NC = 0.63$ and $\overline{NC} = 0.74$). Mixtral is negatively impacted by *Unit conversion* for *Numerical* with $UC = 0.44$ and $\overline{UC} = 0.69$. Qwen2.5-7B is negatively impacted by *Negation* with $Neg = 0.62$ and $\overline{Neg} = 0.72$. Qwen3-4B is negatively impacted by *Negation* with $Neg = 0.57$ and $\overline{Neg} = 0.74$ and *Disjunction* with $Disj = 0.59$ and $\overline{Disj} = 0.72$. Qwen2.5-14B is also impacted by *Negation* with $Neg = 0.58$ and $\overline{Neg} = 0.77$ and *Disjunction* with $Disj = 0.63$ and $\overline{Disj} = 0.75$. While Jullien et al. (2023a) also had similar observations, where models struggled more with numerical inference than other types of inference, we found that *Logical* also has a significant impact on models' performance. In addition, as displayed in Fig. 2, *Negation* and *Disjunction* are highly positively correlated, which suggests that the presence of the two together makes it even harder for the models to deal with. In contrast, the other challenging labels are not correlated with each other.

5. Error Analysis

In this section, we perform a more in-depth analysis of our results. For better readability, we focus on the two best-performing models, the worst-performing model, and Flan-xl, which has a Seq2Seq architecture unlike the others.

5.1. Are LLMs able to understand the CTR structure?

Understanding the structure of a CTR is both essential and challenging for solving the NLI task. Indeed, the model needs to understand the different cohorts, groups, or arms, along with their corresponding results, interventions, eligibility criteria, and adverse events. This is the first step toward understanding the statement and considering the appropriate evidence to establish the inference relation.

All test set instances were annotated with *Task-related expressions*, indicating that understanding the structure of CTR and the relationship between the statement and the premise is a necessary step in reasoning. Most often, other reasoning types coexist with this label. Twelve instances in the dataset were labeled solely as *Task-related expression* and require only an understanding of the CTR structure, with no other type of reasoning involved. Tab. 3 displays the performance of different models on this label.

The results show that both Qwen models seem to be able to deal with CTR structure quite well since their performance on *TRE-only* is greater than the performance on all types of inference. On the other hand, MMed-Llama exhibits a poor performance, which suggests that the model does not understand the CTR structure and can explain why it struggles to solve examples involving multiple inference types.

Model	Log	$\overline{Log}$	Num	$\overline{Num}$	Bio	$\overline{Bio}$	GK	$\overline{GK}$	TE	$\overline{TE}$	All types	$H.O$
MMed-Llama-3	0.56	0.55	0.52	0.55	0.55	0.55	0.49	0.56	0.56	0.55	0.55	0.50
Llama-3	0.59	0.55	0.55	0.56	0.54	0.55	0.57	0.56	0.71	0.55	0.56	0.52
Mistral	0.57	0.59	0.60	0.58	0.61	0.58	0.55	0.56	0.67	0.55	0.59	0.49
Flan-T5-xl	0.68	0.67	0.56	0.70	0.67	0.67	0.68	0.67	0.81	0.67	0.67	0.53
Flan-T5-xxl	0.68	0.66	0.63	0.68	0.67	0.67	0.61	0.67	0.75	0.67	0.67	0.52
Mixtral	0.65	0.68	0.63	0.68	0.65	0.68	0.58	0.68	0.58	0.68	0.67	0.51
Qwen2.5-7B	0.67	0.70	0.69	0.69	0.69	0.69	0.65	0.70	0.63	0.70	0.69	0.51
Qwen3-4B	0.66	0.71	0.69	0.70	0.71	0.70	0.66	0.70	0.56	0.71	0.70	0.51
Qwen2.5-14B	0.69	0.73	0.72	0.73	0.72	0.73	0.62	0.73	0.63	0.73	**0.73**	0.50

Table 2: Mean accuracy per super-classes of inference type on the i and $\bar{i}$ subsets, in decreasing order of All-types accuracy. GK = *Generic knowledge*, Num = *Numerical*, Bio = *Biomedical*, Log = *Logical*, TE = *Typo/Error*. *Task-related expression* is not reported since it appears in every example of the dataset, which means its performance is equivalent to *All types*. Highlighted cells show pairs of results with statistically significant differences between the presence and absence of the inference type. *H.O* is the "hypothesis-only baseline". Standard deviations are all less than or equal to 0.01, so we do not report them in the table.

Model	TRE-only
MMed-Llama-3	0.25
Flan-T5-xl	0.63
Qwen3-4B	0.88
Qwen2.5-14B	0.79

Table 3: Mean accuracies on three runs, on examples labeled *only* with *Task-related expression (TRE)* label.

5.2. Is the number of inference types involved having an impact on performance?

We hypothesize that having a higher number of different types of inference would lead to a decrease in performance, due to the multiple types of knowledge that the model needs to put in contrast.

Fig. 4 shows the mean accuracy as a function of the number of inference types in each example.

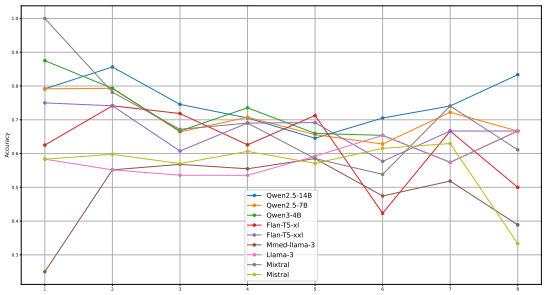
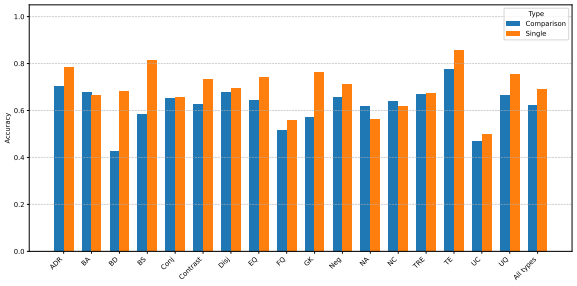


Figure 4: Evolution of the mean accuracy on the three runs as a function of the number of inference types involved in the considered example.

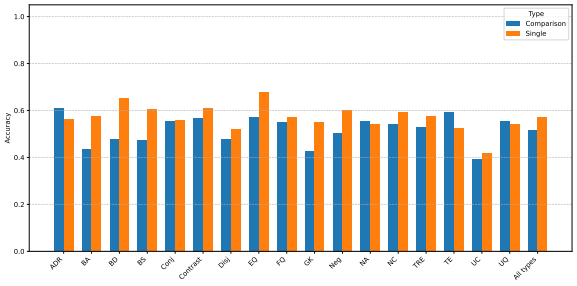
As mentioned in Sec. 5.1, when only one label is present, this is necessarily *Task-related expression*. All models except MMed-Llama achieve one of their highest performances in this case. When introducing two types of inference, model performance tends to increase slightly or remain stable, except for Qwen3-4B and Mixtral, whose performance decreases. Generally, adding more inference types simultaneously is more challenging. Surprisingly, for seven types of inference, Qwen3-4B, Qwen2.5-14B, and Mixtral achieve similar performance as when there are two types of inference present. However, models fall short when facing eight types of inference, except for Qwen2.5-14B. Note, though, that only 18 examples are labeled using 7 different types of inference while only 6 examples are labeled with 8 different inference types ; whereas 58 examples populate the other performance peak, where two inference types are involved. Qwen models were all struggling with *Negation*. For examples involving two inference types, *Negation* is only present in 2% of the examples, meanwhile for examples involving 7 types *Negation* is present for 67% of the examples. We hypothesize that in the case of 7 inferences, the presence of the other types of inference compensates for the difficulty of *Negation*. In contrast, when only two inference types are present, there are no additional types to provide compensation; however, since *Negation* accounts for only 2% of the instances, the model still achieves high performance.

5.3. Is *Comparison* harder than *Single*?

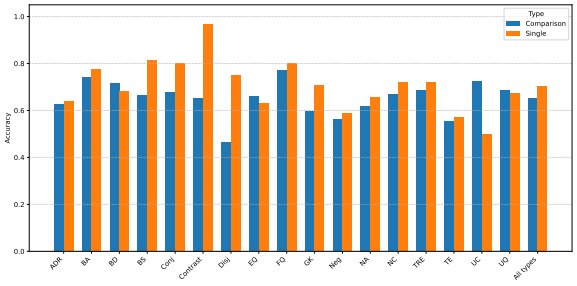
Comparison examples require confronting two CTRs for a given statement. One would expect them to be more challenging, as the model must handle two premises and identify more pieces of evidence compared with a *Single*-example case.



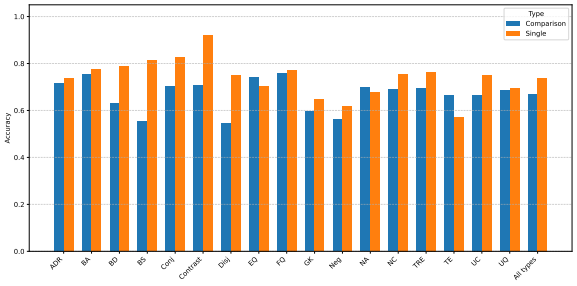
(a) Flan-xl



(b) MMed-Llama-3



(c) Qwen3-4B



(d) Qwen2.5-14B

Figure 5: Accuracy of the different models for *Comparison* and *Single* examples in function of the inference type involved.

Fig. 5 shows the accuracy of the models for *Comparison* and *Single* examples in function of the inference type involved.

For the four models, *Single* remains easier than *Comparison*. However, depending on the type of inference involved, this observation varies: Qwen-14B, Flan-xl, and MMed-llama achieve a higher performance on *Numerical arithmetic* for *Comparison* than *Single*. Qwen-14B also performs better on *Existential quantifier* and *Typo/error* for *Comparison*. Surprisingly, Qwen3-4B is much better on *Unit conversion*, *Biomedical notion definition*, *Existential quantifier*, and *Universal quantifier* when it is a *Comparison*. Some inference types in a *Comparison* setting are challenging for all models: *Biomedical synonym*, *Biomedical notion definition*, and *Generic knowledge*. These types of inferences particularly require comparing several pieces of text, e.g., when comparing synonyms in the first and second CTR premises with the statement.

5.4. *Adverse Events* vs *Results* vs *Intervention* vs *Eligibility*

In each example, the premise is composed of one whole section, either *Adverse Events*, *Results*, *Intervention*, or *Eligibility*. Each section has its own structure and involves different types of knowledge. Fig. 6 displays the mean accuracy of each model per CTR section.

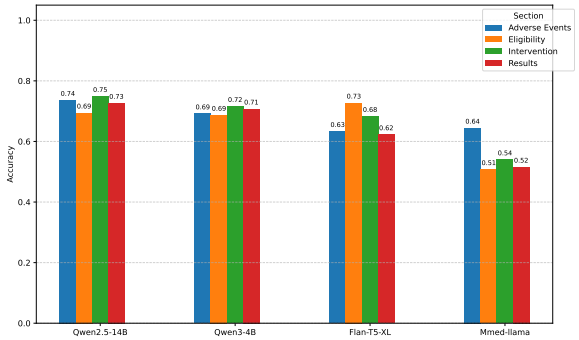


Figure 6: Mean accuracy in function of the section considered as the premise.

We also compute the top-3 most frequent inference type labels present for each section to know what kind of knowledge is most frequently needed to solve an inference involving a given section. While *Eligibility* and *Intervention* are mostly labeled with inference types belonging to the *Logical* superclass, *Adverse Event* is labeled equally with *Numerical* and *Logical* inference types, and *Results* is strongly dominated by *Numerical* inference types.

For Qwen models, the *Eligibility* section is the most challenging, with up to 6 points of difference with the other sections. Flan-T5-xl has a gap of up to 11 points between the *Eligibility* and *Results* sections. MMed-Llama-3 achieves almost random performance for all sections, except for *Adverse Events* with up to 13 more points. We note that the sections on which LLMs struggle are the ones with a higher presence of the inference type whose presence leads to a significant decrease in performance.

6. Conclusion

In this study, we proposed a definition of inference types and new annotations for the NLI4CT dataset for natural language inference on clinical trials. We defined 17 different types of inferences present in NLI4CT, clustered in six super-classes: *Task related expression*, *Logical*, *Biomedical*, *Numerical*, *Generic knowledge*, and *Typo/Error*. We believe that these definitions and this methodology could also be used for general-domain or other domain-specific applications.

We investigated the influence of each inference type on the performance of several open-source LLMs and found that not all models are sensitive to all inference types. Only the models with competitive performance exhibit sensitivity to the presence of certain inference types. In general, there is a significant drop in performance on logical and numerical inference types. We found that Flan-T5-xl is sensitive to *Numerical Comparison*, Mixtral to *Unit conversion*, and the three Qwen models are all sensitive to *Negation*, with Qwen3-4B and Qwen2.5-14B also sensitive to *Disjunction*. For future work, we plan to run the same experiments in a few-shot setting or using Chain-Of-Thought to see whether these approaches would improve results. We also plan to examine LLMs' weaknesses by analyzing natural-language explanations associated with predicted labels and determining whether they correlate with our observations.

7. Limitations

The annotation process remains complex, with sometimes an overlap between inference types such as *Ancestor-descendant relation* and *Biomedical synonym*, or *Generic Knowledge* and *Unit conversion* when it comes to dealing with months, hours, etc., which led to many discussions during the annotation process. As stated by Pavlick and Kwiatkowski (2019), these disagreements can reflect the full distribution of plausible human judgments. To provide a better understanding of the possible annotations produced during our process, we also release the 50 instances annotated by the three annotators and the corresponding justification for each instance.

8. Ethical Considerations

The NLI4CT task uses clinical data 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.

All annotators are NLP researchers, authors of this paper, and paid by their own institutions. They gave consent to annotate this dataset as part of their research activities.

9. Acknowledgements

This work received funding from the CNRS through grant 80PRIME 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.

10. Bibliographical References

Alan Agresti. 2013. *Categorical data analysis*. John Wiley & Sons.

Hyung Won Chung et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Albert Q. Jiang et al. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

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

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

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.

Mael Jullien, Marco Valentino, Leonardo Ranaldi, and Andre Freitas. 2025. Dissecting clinical reasoning in language models: A comparative study of prompts and model adaptation strategies.

Kamal Raj Kanakarajan and Malaikannan Sankarasubbu. 2023. Saama AI research at SemEval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *SemEval-2023*, pages 995–1003, Toronto, Canada.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language

inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Liyun Yan, Danni E, Mei Gan, Cyril Grouin, and Mathieu Valette. 2020. Inference annotation of a Chinese corpus for opinion mining. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4991–4999, Marseille, France. European Language Resources Association.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

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*.