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

Anonymous ACL submission

Abstract

LLMs achieve competitive results on Natural Language Inference (NLI) when applied to clinical trials; however, it is not yet clear on which type of inference LLMs perform well or not. We address this by proposing new supplementary annotations to the existing NLI4CT dataset on the types of inference observed in clinical trials. Our dataset supplements NLI4CT with a total of 1,145 new annotations using our 6 types of inferences. To enhance explainability, we also provide the justifications associated with the labels for a sample of 50 statements. To know on which type of inference LLMs perform worse or better, we prompt Flan-T5, Llama, Mistral, and Qwen and investigate their performance using our newly annotated dataset. We observe that for Flan-T5 and MMed-Llama, the presence of biomedical inference has a positive impact on the overall performance, while for Mistral and MMed-Llama, common knowledge has a negative impact, and for Flan-T5, numerical and linguistic inference have a negative impact. Our code is publicly available on GitHub[1] and the dataset on HuggingFace.[2]

1 Introduction

Large Language Models (LLMs) often obtain high performance in terms of accuracy or F1-score when evaluated on Natural Language Understanding tasks. They tend to outperform traditional encoder-only architectures, such as BERT-like (Devlin et al., 2019) models traditionally used for these discriminative tasks. Natural Language Inference (NLI) consists of determining if a statement can be inferred from a given premise. The possible outcomes are either *entailment*, *contradiction*, or *neutral*. This task can be quite challenging since the model needs to tackle pieces of evidence in both parts of the text and confront these pieces of evidence to determine the inference relation. Often, the entailment relation has to be a multi-hop process, meaning that the model needs to perform several sub-inferences to deduce the final relation. These sub-inferences also involve different kinds of knowledge. To obtain a more fine-grained evaluation of LLMs on these sub-inferences in the clinical trials domain, we provided 1,145 supplementary annotations to NLI4CT (Jullien et al., 2023a), covering 6 different observed inference types for both entailment and contradiction, with a Fleiss' kappa inter-annotator agreement of 0.36. We examine the performance of a set of LLMs on each inference type, with settings including and excluding one considered inference type. Our results show that LLMs such as Flan-T5, Mistral, and MMed-Llama are sensitive to the presence of certain inference types, either affecting the performance positively (biomedical knowledge) or negatively (common knowledge, numerical or linguistic inference).

The contributions of the paper are the following: we first propose an annotation scheme for inference types for clinical trials and apply it to the NLI4CT dataset. Second, we analyze the performance of various LLMs on NLI4CT on each inference type.

2 Related Work

2.1 Annotating Inference Types

In the case of NLI, most existing approaches focus on global accuracy or similar metrics and do not evaluate the reasoning steps or the model's performance on the different types of reasoning (Huang and Chang, 2023). Some previous works (Nie et al., 2020; Joshi et al., 2020; Williams et al., 2022) developed new annotation schemes to obtain a more fine-grained evaluation of the models' performance by proposing annotations for inference (or also called *reasoning*) types. These schemes are surveyed in Tab. 2. All the studies in Tab. 2 (except in NLI4CT) are designed for general-domain

[1] masked_for_anonymity
[2] masked_for_anonymity

applications, and none of them address the clinical trials domain. In addition, these studies investigated Masked-Language Models, while in our study, we focus on Large Sequence-to-Sequence and decoder-only models. Sec. 3.1 gives a detailed comparison of these annotation schemes with the one we propose.

2.2 NLI4CT Dataset Description

The NLI4CT corpus is freely available and consists of a collection of English breast cancer Clinical Trial Reports (CTR) taken from clinicaltrials.gov. This task uses NLI for clinical trials with several use cases, such as checking that a patient complies with the trial's eligibility criteria or checking that a claim can be deduced from the trial's results. NLI4CT comprises two kinds of instances: *single*, where only 1 CTR is involved to perform the inference, and *comparison*, where 2 CTRs are needed to be compared. A premise consists of a section of a CTR (or 2 CTRs if comparison) and a statement of 1 or 2 sentences. The model should predict whether the statement entails or contradicts the premise. The task involves several kinds of inference, both involving general domain and biomedical knowledge.

3 Methodology

In this study, we first systematically investigate the types of inference and knowledge involved in the inference process, and then examine on which type, the models perform the best or the worst. We first define the different inference types (Sec. 3.1), annotate the NLI4CT dataset using our types (Sec. 3.2), and evaluate the models' performance on the inference types (Sec. 3.3).

3.1 Inference Types: Definitions

We define and identify new categories of inference types in the NLI4CT dataset that are needed to solve the inference relation. The goal is to define labels that cover all the observed inference types with little overlap between them, while allowing multiple labels per sample. To define the different inference types, we started by adapting the existing relevant ones in the literature and further defining new categories by picking a few random examples from the dataset and annotating them while incrementally refining the definition of each type. As a result, we define the following inference types:

Logical examples where the inference can be formulated as a test where the output is a Boolean value. It usually involves operators such as negation, implication, conjunction, and disjunction. This label also includes comparison processes, using expressions such as *equal to*, *lower than*, or *greater than*.
E.g.: **Statement (S)**: *There were no cardiac or psychiatric Aes recorded during the primary trial and the secondary trial.* and, **Premise (P)**: ***primary_premise**: Adverse Events 1: Total: 0/344 (0.00%) Adverse Events 2: Total: 0/342 (0.00%). **secondary_premise**: Adverse Events 1: Total: 0/24 (0.00%) Adverse Events 2: Total: 0/23* $\Longrightarrow$ we can define the inference process as a logical test and express the statement using First-Order Logic:

$$\exists x \neg (C(x) \vee P(x)) \wedge (R(x, T_1) \wedge R(x, T_2))$$

$C(x)$: x is a cardiac adverse event.
$P(x)$: x is a psychiatric adverse event.
$R(x, t_i)$: x is an event recorded during trial t_i.
T_1: The primary trial.
T_2: The secondary trial.
With respect to the premise, the result expected would be *True*, so *Entailment*.

Numerical examples where the inference process involves numbers (ordinal, cardinal), converting units of measure or counting elements, as well as quantitative and qualitative descriptions of numerical expressions.
E.g., **S**:*"The primary trial only has a single adverse event recorded for its patient cohort."* and, **P**: *Adverse Events 1: Total: 1/29 (3.45%) Surgery: 1/29 (3.45%)* $\Longrightarrow$ counting the number of adverse events.

Biomedical knowledge examples involving any type of knowledge for which biomedical knowledge is needed. This can vary from medical acronyms, clinical hypernymy/hyponymy, and taxonomic relations among biomedical concepts.
E.g., **S**: *"Eating disorders were not common for the primary trial candidates."* and, **P**: *"Anorexia 1/50 (2.00%)"* $\Longrightarrow$ "anorexia" is an instance of "eating disorders" (taxonomic relation).

Common knowledge examples involving *basic* knowledge that each human possesses, what one could associate with *real-world knowledge*.
E.g.: **S**: *The primary trial uses a 3 week cycle for its intervention, the secondary trial, on the other hand does not have a cyclic treatment in place.* **P**:

Each treatment cycle was defined as 21 days. $\Longrightarrow$ using common knowledge, we know that a week is 7 days and that 3 weeks is indeed 21 days.

Linguistic knowledge examples involving linguistic expressions that are non-trivial, vague, and open to several interpretations.
E.g., **S**: *Eating disorders were not common for the primary trial candidates.* **P**: *Anorexia 1/50 (2.00%)* $\Longrightarrow$ "not common" is rather a vague concept. One could define "common" as a characteristic appearing in at least 50% of the population, but this notion may vary considering the context.

Typos/errors examples where the statement has one or several typos or grammatical errors.
E.g., *The the primary trial intervention section dose not describe the method of administration, dosage or cycle.* $\Longrightarrow$ *The* appears twice.

Comparison to previous annotation schemes The original annotation provided with NLI4CT only separates the numerical inference and does not provide more precise categories. We computed the overlap of instances labeled as *Numerical* in NLI4CT and by using our definition. We found out that we share 83% of instances labeled as *Numerical*, which suggests a similar definition. Our *Numerical* aligns with both ANLI's and Williams et al. (2022)'s definitions, and shares some 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*. *Common 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*. *Linguistic* and *Biomedical* do not really map to any categories in other schemes and can be considered original.

3.2 Annotation Process

To annotate the original NLI4CT dataset with our inference type labels, we ask 3 annotators, all NLP researchers and authors of this paper, to produce annotations for the test set. We sample 10% (50 instances) of NLI4CT's test set, keeping the sample representative of the full test set in terms of *Entailment/Contradiction* and *Single/Comparison* ratios. We provide the annotators with annotation guidelines (see Appx. B) and ask them to provide a short justification for each chosen label. These justifications are used to resolve conflicts between annotators. Each instance can be labeled with one or more inference types. We compute the inter-annotator agreement using the F1 score and Fleiss' Kappa κ (see Appx. D for detailed metrics), and obtain **0.36**, which suggests a fair agreement.[3] Considering our fair inter-annotator agreement, one annotator annotated the rest of the test set, resulting in a total of 1,145 annotations for 500 statements. Tab. 4 displays the detailed dataset statistics and Fig. 3 the correlation matrix between the different inference types.

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), Qwen2.5-7B and 14B-Instruct (Yang et al., 2024). We also evaluate 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) (see Appx. F) and added the mention "*Answer only with:*" to better constrain models to output the desired labels. 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. 1, we report the mean global accuracy to predict *Entailment* or *Contradiction* for all the instances of the test set of NLI4CT, using the template described in Sec. 3.3. All the experiments are run 3 times, each with a different random seed (42, 55, and 3354). Qwen-14B achieves the best results with 0.73 of accuracy; on the other hand, MMed-Llama performs the worst with an accuracy of 0.55.

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,

[3] For reference, the inter-annotator agreement in Joshi et al. (2020) was a Fleiss' Kappa of 0.226 on average for all inference types.

Model	Ck	$\overline{Ck}$	Num	$\overline{Num}$	Bio	$\overline{Bio}$	Log	$\overline{Log}$	T/E	$\overline{T/E}$	$Ling$	$\overline{Ling}$	All types
Flan-T5-xl	0.64	0.67	0.63	0.74	0.72	0.62	0.66	0.71	0.79	0.66	0.51	0.68	0.67
Flan-T5-xxl	0.60	0.68	0.63	0.72	0.72	0.62	0.67	0.67	0.69	0.67	0.49	0.68	0.67
Llama-3	0.47	0.57	0.55	0.58	0.58	0.54	0.55	0.63	0.53	0.56	0.52	0.56	0.56
MMed-Llama-3	0.45	0.57	0.56	0.54	0.55	0.55	0.54	0.63	0.55	0.55	0.58	0.55	0.55
Mistral	0.65	0.58	0.57	0.61	0.63	0.54	0.58	0.64	0.56	0.59	0.64	0.58	0.59
Mixtral	0.52	0.70	0.66	0.70	0.66	0.68	0.67	0.73	0.61	0.68	0.74	0.67	0.67
Qwen-7B	0.65	0.70	0.69	0.70	0.71	0.68	0.70	0.67	0.67	0.70	0.69	0.69	0.69
Qwen-14B	0.75	0.72	0.73	0.73	0.75	0.70	0.72	0.83	0.74	0.73	0.74	0.73	**0.73**

Table 1: Mean accuracy for all types and per inference type on the i and $\bar{i}$ subsets. CK = Common Knowledge, Num = Numerical, Bio = Biomedical, Log = Logical, T/E = Typo/Error, Ling = Linguistic. Highlighted cells indicate scores where the model statistically performs better (green) or worse (red), where the inference type is present. Standard deviations being less than or equal to 0.01, we do not report them in the table.

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. Tab. 1 reports these results.

We perform two kinds of Chi-square (χ^2) tests (Agresti, 2013) (see Appx. G for detailed formulas) with a p-value threshold of 0.05. First, χ^2-*all*, where, for each model, the χ^2 is computed on all *original* inference types subsets. We define our null hypothesis as: "There is no relation between the behavior of the system and the presence of any inference type". For each of the 3 runs of a system, we compute its χ^2 and the associated p-value. In all cases, the p-value of the 3 runs is on the same side of the threshold. We report the mean of the 3 runs. We observe that for most of the models, the inference types do not have an influence on the overall performance, except for Flan-xl and Flan-xxl, where the p-value was below the threshold, which suggests that depending on the inference types present, the model does not perform the same.

To know which inference type influences the performance, we define χ^2-*type*, where, for each model, the χ^2 is computed on one i and $\bar{i}$ inference type subsets. The null hypothesis is: "When inference type i is present, the model performs as well as when the inference type i is not present." For Flan-xl and Flan-xxl, the gap in performance is significant on the *Num* and *Ling* types, on which the 2 models are struggling more, whereas on the *Bio*, the model performs better. Mistral's χ^2-*type* also demonstrates a significantly better performance when the *Bio* inference is present. On the contrary, Mistral and MMed-Llama-3's χ^2-*type* show a significant loss of performance when Common Knowledge (*Ck*) is involved.

In addition, *Num* and *Ling* are positively correlated, while *Num* and *Bio* are strongly negatively correlated. As a consequence, since *Num* and *Ling* often appear together, if Flan-T5 struggles with one of these two inferences, this will also have an impact on the performance of the other inference. Jullien et al. (2023a) also had similar observations, where models struggled more with numerical inference than other types of inference.

5 Conclusion and Future Work

In this study, we proposed a definition of inference types along with new annotations on the NLI4CT dataset for natural language inference on clinical trials. 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 the types of inference involved. There is a significant drop in performance on linguistic, numerical, and common knowledge inference types. On the other hand, biomedical inference seems to be easier, resulting in better performance of models when this inference is present. We also believe that these definitions could be used for general-domain or other domain-specific applications. For future work, we plan to run the same experiments in a few-shot setting or by using Chain-Of-Thought to see whether it would improve the results. We also plan on looking into LLMs' weak points by analyzing the natural language explanations associated with the predicted labels and seeing if these explanations correlate with our observations.

6 Limitations

The disagreement between the annotators highlights the complexity of the annotation task. There is often an overlap between the different inference

labels (e.g., *logical* and *numerical* when it comes to number comparison), which leads 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 give a better understanding of the possible annotations produced during our process, we also release the 50 instances annotated by the 3 annotators and the corresponding justification for each instance.

7 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.

References

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

Mathilde Aguiar, Pierre Zweigenbaum, and Nona Naderi. 2025. Am I eligible? natural language inference for clinical trial patient recruitment: the patient's point of view. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 243–259, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. 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.

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.

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).

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.

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*.

A Other Annotation Schemes in the Literature

Tab. 2 displays different annotation schemes on similar tasks taken from the literature.

B Annotation Guidelines

All the annotators had access to the annotation guide with a description of the task and the different labels (see Fig. 1).

C Examples of Annotated Pairs

Fig. 2 shows an example of an annotated statement-premise pair. The highlighted spans are the pieces of text that correspond to each inference type label. Here, the label *Biomedical* was chosen because of the taxonomic relation between *eating disorder* and *Anorexia*. *Logical* and *Numerical* were chosen because of *Most,* but *Numerical* especially because of the highlighted numbers to process.

Study	Annotations
ANLI (Nie et al., 2020)	Numerical & Quant., Reference & Names, Standard, Lexical, Tricky, Reasoning & Facts, Quality
TaxiNLI (Joshi et al., 2020)	(top-level) Linguistic, Logical, Knowledge
Williams et al. (2022)	(top-level) Numeral, Basic, Reference, Tricky, Reasoning, Imperfections
NLI4CT (Jullien et al., 2023a)	NLI, Numerical
Our	Common Knowledge, Biomedical, Logical, Numerical, Typo/error, Linguistic

Table 2: Different annotation schemes in the literature.

D Detailed Inter-annotator Agreement

Tab. 3 reports the inter-annotator agreement using F1 and Fleiss' Kappa.

Pair	F1 Score
A1 vs A2	0.67
A1 vs A3	0.60
A2 vs A3	0.60
Average F1	**0.62**

(a) Pairwise F1 inter-annotator agreement.

Inf. type	Fleiss' κ
CK	-0.07
Num	0.51
Bio	0.57
Ling	0.44
Log	0.20
T/E	0.54
Average	**0.36**

(b) Fleiss' κ for inter-annotator agreement.

Table 3: Inter-annotator agreement measures.

Overall, we obtain a fair agreement, although labeling instances with *Common knowledge* seems to be challenging for the annotators.

E Annotations Distribution

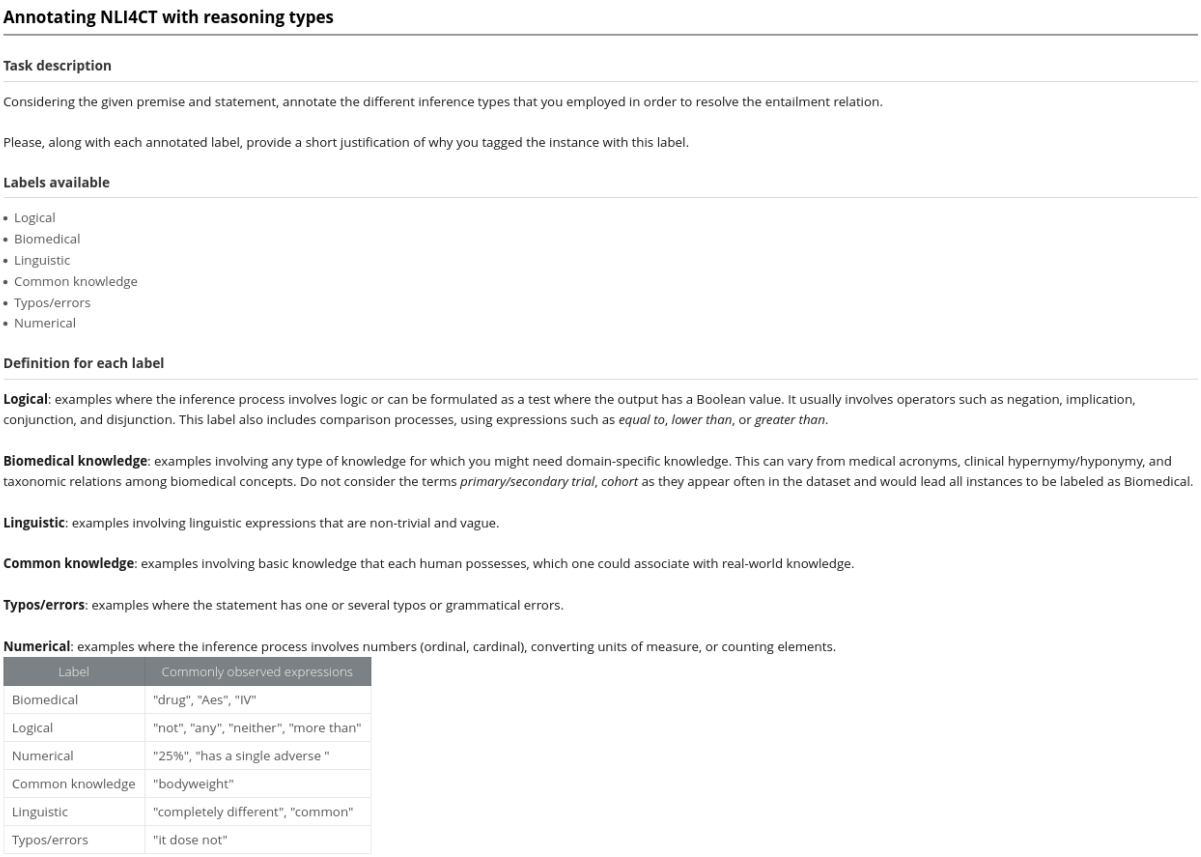
Tab. 4 displays statistics about the resulting dataset.

Fig. 3 shows the Pearson's correlation between the different inference types labels that occur in a single instance of the dataset.

The *Linguistic* inference type is positively correlated with *Numerical*, while *Numerical* being strongly negatively correlated with *Biomedical.*

F Prompts and Models

We selected Flan-T5-xl and xxl (Chung et al., 2024), respectively 3 and 11 billion parameters instruction-tuned sequence-to-sequence models



Annotating NLI4CT with reasoning types

Task description

Considering the given premise and statement, annotate the different inference types that you employed in order to resolve the entailment relation.

Please, along with each annotated label, provide a short justification of why you tagged the instance with this label.

Labels available

- Logical
- Biomedical
- Linguistic
- Common knowledge
- Typos/errors
- Numerical

Definition for each label

Logical: examples where the inference process involves logic or can be formulated as a test where the output has a Boolean value. It usually involves operators such as negation, implication, conjunction, and disjunction. This label also includes comparison processes, using expressions such as *equal to*, *lower than*, or *greater than*.

Biomedical knowledge: examples involving any type of knowledge for which you might need domain-specific knowledge. This can vary from medical acronyms, clinical hypernymy/hyponymy, and taxonomic relations among biomedical concepts. Do not consider the terms *primary/secondary trial*, *cohort* as they appear often in the dataset and would lead all instances to be labeled as Biomedical.

Linguistic: examples involving linguistic expressions that are non-trivial and vague.

Common knowledge: examples involving basic knowledge that each human possesses, which one could associate with real-world knowledge.

Typos/errors: examples where the statement has one or several typos or grammatical errors.

Numerical: examples where the inference process involves numbers (ordinal, cardinal), converting units of measure, or counting elements.

Label	Commonly observed expressions
Biomedical	"drug", "Aes", "IV"
Logical	"not", "any", "neither", "more than"
Numerical	"25%", "has a single adverse "
Common knowledge	"bodyweight"
Linguistic	"completely different", "common"
Typos/errors	"it dose not"

Figure 1: Annotation guide used by all the annotators

Count	Value
Contradiction	250
Entailment	250
Logical	455
Numerical	313
Biomedical	238
Common Knowledge	67
Linguistic	43
Typo/error	29
Comparison	271
Single	229

Table 4: Dataset statistics

ranking 2^{nd} in SemEval (SE) 2023 ; Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) is a 7 billion parameters decoder-only model, and Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024) is its equivalent with 45 billion parameters and using the mixture of experts approach, with both models ranking 1^{st} and 2^{nd} in SE 2024 ; Llama-3.2-8B-Instruct (Dubey et al., 2024) a decoder-only model with 8 billion parameters; and Qwen2.5-7B and 14B-Instruct (Yang et al., 2024) also achieving competitive results on NLI4PR (Aguiar et al., 2025), a task similar to NLI4CT.

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

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

G Chi-square Tests

We define two types of Chi-square tests: *chi-square-all* (Eq. 1) and *chi-square-type* (Eq. 2), where the first takes as input the performance on all the inference types $i_1, ..., i_6$ altogether, while the second takes as input one inference type i and the associated contrast subset $\bar{i}$.

Chi-square-all:

$$\chi^2_{all} = \sum_{k \in \{i_1,...,i_6\}} \frac{(O_k - E_k)^2}{E_k} \quad (1)$$

Chi-square-type:

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

Statement:

Most the primary trial candidates suffered from some kind of eating disorder during the study duration

Premise:

Adverse Events 1: Total: 17/50 (34.00%)
Fatigue 4/50 (8.00%)
Papulopustular rash 1/50 (2.00%)
Alanine aminotransferase increased 5/50 (10.00%)
Aspartate aminotransferase increased 4/50 (8.00%)
Alkalosis 1/50 (2.00%)
Anorexia 1/50 (2.00%)
Hyperglycemia 2/50 (4.00%)
Nervous system disorders - Other 1/50 (2.00%)
Dry skin 1/50 (2.00%)
Rash acneiform 1/50 (2.00%)

NLI label:

Contradiction

Inference types labels:

Numerical, Logical, Biomedical

Figure 2: Example of an annotated pair using *Biomedical*, *Logical* and *Numerical* inference types labels.

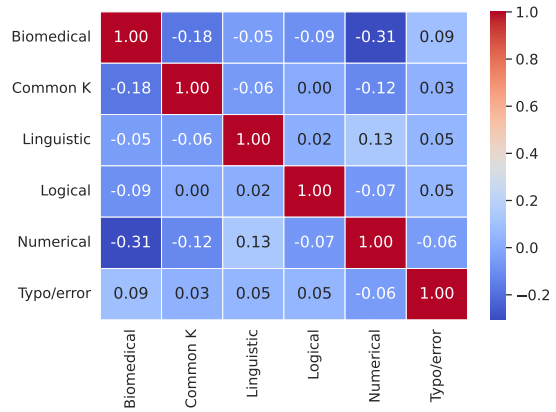


Figure 3: Correlation matrix of the different inference type labels.