Human Experts vs. Large Language Models: Evaluating Annotation Scheme and Guidelines Development for Clinical Narratives

Ana Luísa Fernandes[1,2,*,†], Purificação Silvano[1,2,9,†], Nuno Guimarães[1,2,†], Rita Rb-Silva[7,8,†], Tahsir Ahmed Munna[1,2,†], Filipe Cunha[1,2,3,†], António Leal[2,4,9], Ricardo Campos[1,5,6,†] and Alípio Jorge[1,2,†]

[1]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
[2]University of Porto, Portugal
[3]University of Minho, Braga, Portugal
[4]University of Macau, China
[5]University of Beira Interior, Covilhã, Portugal
[6]Ci2 - Smart Cities Research Centre (IPTomar), Covilhã, Portugal
[7]Research Centre of the Portuguese Institute of Oncology of Porto (CI-IPOP), Porto, Portugal
[8]RISE-Health, Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Portugal
[9]CLUP - Centre of Linguistics of the University of Porto, Portugal

Abstract

Electronic Health Records (EHRs) contain vast amounts of unstructured narrative text, posing challenges for organization, curation, and automated information extraction in clinical and research settings. Developing effective annotation schemes is crucial for training extraction models, yet it remains complex for both human experts and Large Language Models (LLMs). This study compares human- and LLM-generated annotation schemes and guidelines through an experimental framework. In the first phase, both a human expert and an LLM created annotation schemes based on predefined criteria. In the second phase, experienced annotators applied these schemes following the guidelines. In both cases, the results were qualitatively evaluated using Likert scales. The findings indicate that the human-generated scheme is more comprehensive, coherent, and clear compared to those produced by the LLM. These results align with previous research suggesting that while LLMs show promising performance with respect to text annotation, the same does not apply to the development of annotation schemes, and human validation remains essential to ensure accuracy and reliability.

Keywords

clinical narratives, annotation schemes, LLM, Electronic Health Records, health data

1. Introduction

Electronic Health Records (EHRs) contain extensive volumes of unstructured narrative text, presenting considerable challenges for their organization, curation, management, and effective reuse for both clinical and research purposes [1]. Given that an estimated 70-80% of the clinical information within EHRs is text-based [2], Natural Language Processing (NLP) techniques play a pivotal role in automating the retrieval, processing, and extraction of relevant biomedical data [1]. However, manual information extraction remains a highly labor-intensive process that requires significant clinical expertise and

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

*Corresponding author.
ana.l.fernandes@inesctec.pt (A. L. Fernandes); msilvano@letras.up.pt (P. Silvano); nuno.r.guimaraes@inesctec.pt (N. Guimarães); rrsilva@med.up.pt (R. Rb-Silva); tahsir.a.munna@inesctec.pt (T. A. Munna); lfc@di.uminho.pt (F. Cunha); antonioleal@um.edu.mo (A. Leal); ricardo.campos@ubi.pt (R. Campos); alipio.jorge@inesctec.pt (A. Jorge)
https://github.com/analuisacardosofernandes/Human-Experts-vs.-Large-Language-Models (A. L. Fernandes)
0009-0009-0552-3904 (A. L. Fernandes); 0000-0001-8057-5338 (P. Silvano); 0000-0003-2854-2891 (N. Guimarães); 0000-0002-1422-0974 (R. Rb-Silva); 0000-0001-9269-502X (T. A. Munna); 0000-0003-1365-0080 (F. Cunha); 0000-0002-6198-2496 (A. Leal); 0000-0002-8767-8126 (R. Campos); 0000-0002-5475-1382 (A. Jorge)

extensive training to achieve a high level of Inter-Annotator Agreement (IAA). Moreover, manual extraction is often impractical for studies involving large datasets, such as clinical trials, underscoring the need for more efficient computational methods [2]. While the implementation of high-performance information extraction algorithms has become increasingly feasible due to advancements in NLP, the creation of high-quality annotated corpora for training and evaluating automatic models continues to pose a significant challenge [3]. To ensure the development of high-quality datasets, it is essential to establish a robust and comprehensive annotation scheme that accurately accounts for the unique characteristics of clinical text and represents them with precision and completeness.

Annotation schemes consist of descriptive and analytical labels that, during the process of annotation, are associated with linguistic data, guided by predefined guidelines that specify the labels, features, annotation units (e.g., token, phrase, clause, or document) and instructions on how to proceed. To ensure consistency, labels and units must have clear operational definitions, facilitating agreement among human annotators. In cases where annotation supports machine learning, schemes may highlight features correlated with annotation labels. Modern annotation workflows often employ specialized tools that enable span identification, label assignment, and relationship marking, alongside measures of IAA to assess consistency and inform the development of automatic annotation systems [4].

One of the primary challenges in developing annotation schemes for clinical narratives arises from the substantial heterogeneity in content and writing styles across different hospitals [3], as well as across various departments and services within the same institution. Clinical text is typically composed in a free-form, spontaneous manner, exhibiting a wide-ranging diversity of medical domain topics and concepts. Throughout a patient's hospital journey, a variety of medical reports are generated, including admission reports and discharge summaries following hospitalization. Furthermore, clinical text in EHRs differs significantly from non-clinical text due to the specialized nature of medical language and the frequent use of abbreviations, which significantly increase processing complexity [5]. For instance, the Unified Medical Language System (UMLS) encompasses over two million terms representing approximately 900,000 concepts across more than 60 biomedical terminologies, as well as 12 million relationships among these concepts [6]. Additionally, biomedical terminology is highly intricate, with some terms exhibiting context-dependent meanings [1].

To gain a better understanding of clinical narratives, especially concerning the chronological progression of the patient's hospital journey, it becomes crucial to analyze the sequence of medical reports generated during their care. This analysis involves considering the temporal semantics inter-document, which helps to create a structured timeline of hospital events that accurately reflects the patient's history. However, this necessity complicates the development of annotation schemes.

Bearing in mind all these challenges, designing an annotation scheme for clinical reports is a complex endeavor, even for experts in both linguistics and the medical domain. In this study, we aim to investigate the extent to which Large Language Models (LLMs) can address this challenge. LLMs have, since their surge, led to an increasing reliance on automated and generative methods for data annotation [7]. While LLMs can achieve competitive performance in annotation tasks compared to human annotators, existing research [8, 9] has shown that human expertise continues to overcome LLMs, particularly in more complex annotation tasks. However, to the best of our knowledge, no prior studies have specifically evaluated the effectiveness of LLMs in developing annotation schemes for representing temporal information in clinical narratives.

Accordingly, this study makes the following key contributions:

1. Performance Assessment of LLMs: We evaluate the capability of an LLM in generating an annotation scheme and corresponding guidelines, with a specific focus on temporal information in clinical narratives.
2. Comparative Analysis: We conduct a systematic comparison between human-generated and LLM-generated annotation schemes, assessing their efficiency, consistency, and applicability.
3. Multilingual Expansion: Unlike most prior studies focused on English, our research extends the evaluation of LLM performance to Portuguese, broadening the understanding of their capabilities across languages.

The structure of this paper is outlined below. Section 2 presents various annotation schemes for clinical narratives, emphasizing their scarcity and incompleteness. Section 3 outlines the methodology, beginning with the process of developing an annotation scheme for clinical narratives by a human expert (3.1.1) and an LLM (3.1.2), followed by the creation of an evaluation framework in Section 3.2 designed to assess both annotation schemes. This section provides a comprehensive description of the metrics and procedures employed for evaluating human annotation using the two schemes, as well as the qualitative assessments of the guidelines utilizing Likert scales. Finally, Section 4 presents and discusses the results. In Section 4.1, we present the problems encountered in the annotation performed according to each of the schemes, specifically the results and analysis of the curation process and the IAA values. Section 4.2 presents the results of the annotation scheme evaluation conducted by the annotators using Likert scales.

2. Annotation schemes for clinical narratives

Annotation schemes for clinical narratives remain scarce, with limited availability of annotated corpora. One of the earliest efforts, the *Clinical E-Science Framework* (CLEF) [10], annotated a corpus following a scheme which included entities, relations, modifiers, co-references, and temporal information. Patel et al. [11] annotated a large corpus of clinical documents following a scheme with 11 semantic groups mapped to UMLS semantic types [6]. The *Temporal Histories of Your Medical Events* (THYME) [12] corpus and the *i2b2* project [13] integrated event and temporal relation annotations extending ISO TimeML [14] to the annotation of clinical reports. Additionally, the *MiPACQ* corpus [15] applied syntactic and semantic annotations based on the UMLS semantic hierarchy.

Over the years, most clinical NLP research has focused on English, Chinese being the second most common language [3], and significantly fewer efforts have been dedicated to other languages. In Spanish, the *IxaMed-GS* corpus applied an *SNOMED-CT*-based scheme for disease and drug entity annotations [16], while the *MERLOT* corpus developed a comprehensive semantic annotation framework for clinical documents in French [17].

For Portuguese, annotation efforts remain particularly limited. In Brazilian Portuguese, Souza et al. [18] created a Named Entity Recognition (NER) system for various clinical narrative types, while Oliveira et al. [19] introduced *SemClinBr*, the first semantically annotated clinical corpus for this language, with a scheme incorporating UMLS semantic types alongside additional tags for negation and abbreviations. Rocha et al. [20] extended this research by manually annotating patient reports for automatic information extraction, while the *MedAlert Discharge Letters Representation Model* (MDLRM) focused on entity annotation in 90 Brazilian Portuguese hospital discharge summaries. In European Portuguese, Lopes et al. [21] describe a clinical text collection with entities manually annotated, and Nunes et al. [22] introduced the *MediAlbertina* model, a BERT-based encoder pre-trained on Portuguese electronic medical records, which is annotated with entities and their status (present or absent).

Despite these initiatives, most annotation schemes remain limited in scope, primarily focusing on NER. Moreover, the reliance on medical ontologies constrains the annotation of morphosyntactic and semantic features, underscoring the need for more comprehensive and multilingual annotation frameworks in clinical NLP. The inclusion of deeper morphosyntactic and semantic structures, such as relationships between entities and temporal information, is crucial to properly represent relevant clinical information. For instance, the relevant antecedents for understanding the clinical cases are organized in a specific temporal order, which may not coincide with the linear order of discourse. In that case, temporal relations are determined, for instance, by expressions that can be of different kinds (nouns, adjectives, adverbs), which have to be identified and labeled during annotation. Another source of temporal ordering is the aspectual nature of the situations themselves: states are unbounded situations, which tend to establish temporal inclusion with other situations, whereas transitions are telic situations, which trigger temporal precedence. Therefore, aspectual information is paramount to determining temporal organization and should also be included in annotation frameworks for clinical narratives.

3. Methodology

3.1. Developing a temporal annotation scheme and guidelines for clinical narratives

In this section we describe how two annotation schemes and respective guidelines were designed to capture temporal information in clinical narratives in European Portuguese. The first scheme is defined by a human expert, while the second is defined by an LLM. We will provide a detailed account of both processes, outlining the steps, criteria, and methodologies that guided each approach.

3.1.1. By Human

A specialist in linguistics and pharmaceutical sciences, with expertise in semantic annotation, developed the annotation scheme through a comparative analysis of various existing frameworks. This analysis encompassed the temporal layer of the *Text2Story* scheme [23, 24, 25, 26], which is a general annotation scheme based on ISO 24617 - Language Resource Management – Semantic Annotation Framework [14] and is designed for annotating morphosyntactic and semantic information in European Portuguese news texts. Additionally, the study examined the *i2b2* [13] and *MERLOT* [17] frameworks, both of which are specifically developed for the annotation of clinical texts.

For this comparative analysis, six pseudo-anonymized medical reports from patients diagnosed with Acute Myeloid Leukemia, followed by IPO-Porto (Portugal), were annotated according to the guidelines of the *Text2Story*, *i2b2*, and *MERLOT* annotation schemes. The results revealed that, although *Text2Story* captured morphosyntactic and semantic information, it lacked labels specific to the medical domain. In contrast, the *i2b2* and *MERLOT* frameworks, while including domain-specific labels, were overly broad in scope.

Based on these preliminary findings, we used 40 medical reports from patients diagnosed with Acute Myeloid Leukemia, followed by IPO-Porto, to identify additional tags needed to enhance the *Text2Story* annotation scheme in order to capture medical domain-specific information. This corpus included admission reports, discharge summaries, and general medical reports. This work was conducted in collaboration with a medical specialist from IPO-Porto, who validated the most relevant clinical elements for annotation. In selecting the semantic classes, the UMLS Metathesaurus ontology was considered, ensuring a systematic approach aligned with international standards. The definition of medical labels was further grounded in the work of Leite [27].

The guidelines for the temporal annotation framework and the proposed set of labels can be consulted in detail in the GitHub repository.

3.1.2. By LLM

For the development of an annotation framework and guidelines by an LLM capable of capturing temporal and clinical information, we utilised Gemini. We chose Gemini 1.5 Flash due to its performance and because we had access to a paid version. It is comparable to that of GPT-4 models across various tasks[1], and because it belongs to a family of models that currently offer context windows larger than those provided by OpenAI [2]. This latter characteristic is particularly significant, as it enables the application of the methodology presented in this study to more extensive annotation guidelines.

For the development of the prompts provided to the model, we employed an adaptation of ablation studies. We chose this method as ablation studies offer valuable insights into the contribution of each component of the prompts to the performance of the LLM [28]. Based on this, we began by providing the model with a simpler prompt and gradually enriched its structure. In total, three distinct prompts were developed plus two variants of two of them, reflecting an iterative process of refinement. In what follows, we provide a detailed description of the content of each prompt, along with the evaluation conducted for each. To conduct the prompts' assessment, we defined the parameters and questions presented in Table 1 establishing Likert scales [29].

[1]The reader is referred to the leaderboard presented at https://lmarena.ai/?leaderboard
[2]Cf. information at https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#context-window

Table 1
Evaluation parameters and corresponding questions.

Parameters	Questions
Markables (tag spans)	How clear and consistent are the guidelines for defining markables?
Events (grammatical domains)	How effective are the guidelines for identifying and classifying events at the morphosyntactic and grammatical levels?
Events (medical domain)	How well-defined and comprehensive are the tags for the medical domain?
Temporal Expressions	How effective are the guidelines for identifying and classifying temporal expressions?
Temporal Relations	How accurate are the guidelines for identifying and classifying temporal relations between events?
Definitions	How clear and detailed are the definitions of each tag?
Ambiguities	How effective is the approach to resolving ambiguities?
Coherence and Clarity	How coherent and clear are the provided guidelines?
Adaptation of Text2Story	How effective is the adaptation of the Text2Story scheme for medical reports?
Examples	How relevant and illustrative of the tags are the examples?

Table 2
Means and standard deviations of the results obtained from evaluating each Prompt's output using Likert scales (the Prompt content is a summary).

Prompt	Content	Evaluation	
		$\bar{x}$	σ
Prompt 1	Development of a temporal annotation scheme capturing morphosyntactic, grammatical, and medical domain-specific information.	2.1	0.74
Prompt 2A	Same as Prompt 1, + T2S guidelines + instructions to capture morphosyntactic and grammatical information.	2.7	1
Prompt 2B	Same as Prompt 2A + request to include tags specific to the medical domain.	3	1
Prompt 3A	Same as Prompt 2B + inclusion of synthetic reports as examples.	2.8	0.84
Prompt 3B	Same as Prompt 3A + an emphasis on deriving examples from synthetic reports + providing detailed specifications of markables.	3.2	0.63

The scores for each parameter ranged from 1 to 5. For example, for the first parameter, "Markables", the question "How clear and consistent are the guidelines for defining markables?" offered five response options: 1 – The guidelines are nonexistent or very vague; 2 – The guidelines are partially clear but frequently ambiguous; 3 – The guidelines are satisfactory but occasionally lack clarity; 4 – The guidelines are clear in most cases, with some areas for improvement; 5 – The guidelines are precise, unambiguous, and consistently applicable.

This information regarding the content of each prompt and its evaluation is presented in Table 2.

The first prompt (Prompt 1) included instructions for the development of a temporal annotation scheme capturing morphosyntactic, semantic, and medical domain-specific information.

As the output of Prompt 1, the LLM addressed all elements of the input, generating guidelines that included domain-specific medical tags ("diagnosis", "treatment", "symptom", "procedure", "test", and "state"), as well as temporal expressions ("date", "time", "duration", and "frequency") and temporal relations ("before", "after", "simultaneously", "includes", "is included", and "during"). The guidelines provided basic examples for some tags. For instance, for the tag "Treatment", examples included "start of chemotherapy", "bone marrow transplant", and "radiotherapy". However, the guidelines failed to clearly specify the markables, significantly undermining the annotation process. For example, in the case of "start of chemotherapy", the example implies that the annotation should use a single tag. Yet, this phrase involves two distinct events: "start" and "chemotherapy". The lack of a precise definition for the markables thus introduces ambiguity into the annotation process. The proposed scheme also

exhibited limitations in capturing more detailed morphosyntactic and grammatical information, as events were generically annotated as "events" and supplemented only by the medical domain tags. Prompt 1 received the lowest average evaluation score (2.1), as shown in Table 2.

Due to the limitation in including morphosyntactic and semantic tags, and to ensure that the LLM was provided with the same conditions as the human expert, we decided, for the second prompt (Prompt 2A), to supply the guidelines from the original *Text2Story* annotation scheme. Thus, Prompt 2A was structured similarly to Prompt 1, with the addition of the *Text2Story* guidelines and the following phrase appended to the end of the prompt: "For capturing morphosyntactic and grammatical information, refer to the *Text2Story* guidelines provided in the attached document".

As the output of Prompt 2A, the LLM adopted the same tags for events, temporal expressions, and temporal relations present in the *Text2Story* guidelines, while introducing a new tag, "domain", with attributes such as "diagnosis", "treatment", "prognosis", and "etc.". The guidelines included a clarification of markables, following the standard established by *Text2Story*. However, examples were provided only for temporal expressions. At the end of the guidelines, an annotated example sentence was included but exhibited significant limitations. The output of Prompt 2A received an average Likert scale score of 2.7.

Since the tags related to the medical domain were insufficient and qualitatively inferior to the output of Prompt 1, we decided to reinforce the final sentence of the second prompt with the following request: "Ensure to include tags specific to the medical domain" (Prompt 2B). As a result, the output was similar to Prompt 2A but included an expansion of the medical domain tags, such as "diagnosis", "treatment", "procedure", "symptom onset", "symptom resolution", "prognosis", "follow-up", "lab results", and "medication administration". However, significant limitations persisted in the output, including the absence of clear definitions for the tags and examples based on only one annotated sentence, which was marked by contradictions and inconsistencies. For instance, in establishing temporal relations, the directionality of the arrow was not specified, resulting in an example where the same events were linked by both "after" and "before". The average evaluation score for the output of Prompt 2B was 3.

Regarding Prompt 3A, with the aim of improving the quality of examples generated by the LLM and, once again, ensuring that the LLM had conditions comparable to those of a human expert, we opted to provide the model with synthetic reports created by a specialist physician. Thus, Prompt 3A was structured similarly to Prompt 2B, with the addition of a set of synthetic reports and the following instruction at the end: "Use the five attached medical reports as examples". As a result, the output was of lower quality than that generated by Prompt 2B, receiving an average Likert scale score of 2.8. The LLM produced only one example derived from the provided reports, which exhibited limitations and inconsistencies. Additionally, the guidelines presented were unclear in specifying the markables, merely referring to the *Text2Story* guidelines.

In light of these results, we decided to refine Prompt 3A further, developing Prompt 3B, with an emphasis on deriving examples from synthetic reports and providing detailed specifications of markables. The output generated by Prompt 3B stood out for providing the best guidelines among the outputs of the prompts tested achieving the highest average score (3.2). Nonetheless, these guidelines still exhibited several weaknesses, such as a lack of clarity in defining markables, the absence of well-defined tags for the medical domain, limitations in the approach to resolving ambiguities, and issues with the quality of the examples provided.

The annotation scheme and guidelines generated as output of Prompt 3B were used as a reference for comparison with the guidelines developed by the human expert, since they obtained the best result.

The content of the prompts, as well as the guidelines produced as outputs and their assessement, can be found in the GitHub repository.

Figure 1 depicts the sequential steps involved in developing the annotation scheme, outlining the processes followed by both the human expert and the LLM.

3.2. Evaluation framework

To evaluate the annotation schemes and guidelines produced by both a human and the LLM, two complementary approaches were employed. First, the two annotation schemes and their respective

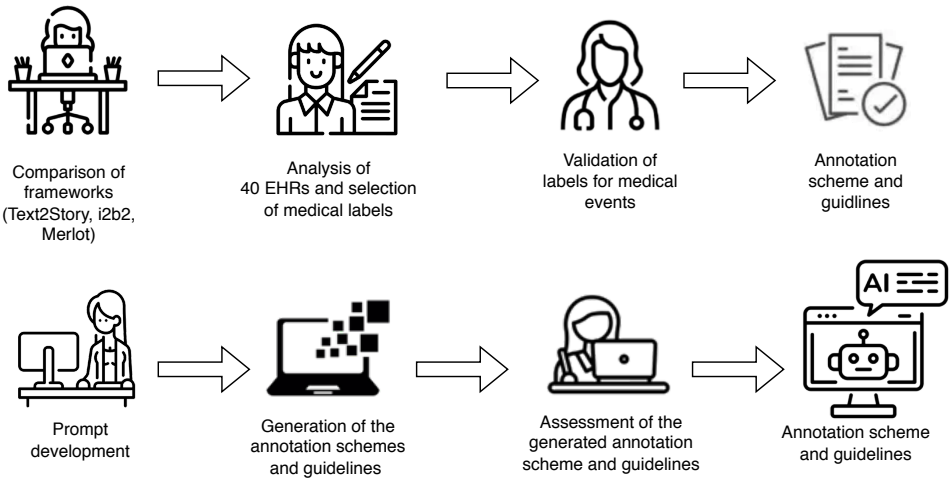

Figure 1: Overview of the annotation scheme development process followed by a human expert and by the LLM

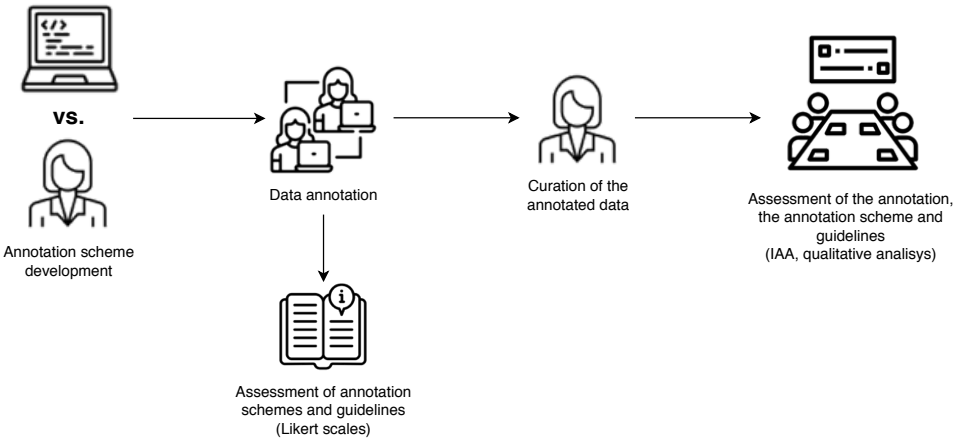

Figure 2: Overview of the two evaluation approaches used to assess the efficacy of the two annotation schemes.

guidelines were applied to the annotation of clinical reports by two annotators. This human annotation was evaluated by a specialist in linguistics and pharmaceutical sciences during the curation process and by IAA metrics. Second, both the human- and LLM-generated schemes and guidelines were assessed by the two annotators. Figure 2 provides a summary of these two approaches.

Regarding the first approach, the corpus used for annotation consisted of synthetic reports for two patients diagnosed with Acute Myeloid Leukemia, written by a specialist physician from IPO-Porto. For each patient, an admission report, two discharge reports, and a general report were provided.

The sets of reports were annotated according to the two annotation schemes (human- and LLM-generated). The reports for Patient 1 followed the guidelines developed by a human expert, while the reports for Patient 2 were annotated according to the guidelines formulated by the LLM as output from Prompt 3B. Annotation was conducted by two linguistics students with extensive experience in semantic annotation.

To minimize biases during the annotation process, such as the potential influence of one scheme on another, the following approach was adopted: Annotator 1 commenced the task using the scheme developed by a human expert, annotating the reports for Patient 1. Simultaneously, Annotator 2 began with the scheme created by the LLM, annotating the reports for Patient 2. Upon completion of this initial phase, it was the other way around, ensuring that each annotator contributed to all combinations of scheme and patient.

Curation was conducted by an expert to ensure the accuracy and consistency of the annotations, as

well as to verify whether the annotators had correctly interpreted and applied the guidelines to the reports. The annotations and curation were performed using the INCEpTION tool [30]. Additionally, to assess the quality and reliability of the guidelines, the IAA was calculated. This percentage allows for the identification of ambiguities or difficulties in the interpretation of the guidelines, with values closer to 100% indicating higher reliability [31].

Another approach used to assess the quality of the annotation schemes and guidelines created by both the human expert and the LLM involved the application of the same Likert scales [29] that were developed to evaluate the prompt outputs, presented in Table 1. In this instance, the evaluation was conducted by the annotators.

4. Results and discussion

4.1. Human annotation

As explained in the previous section, to assess the quality and efficacy of the annotation schemes and guidelines produced by the human expert and by the LLM, we devised an experiment in which both annotation frameworks were used to represent information from clinical reports. Once the annotation was finished, the curation was carried out by a specialist in linguistics and pharmaceutical sciences, enabling some general observations.

Regarding the annotation based on the guidelines developed by a human expert, the curator observed that the identification of events and temporal expressions was performed unanimously and in accordance with the annotation guidelines. Minor inconsistencies were reported, particularly in the classification of "medical domain" attributes, possibly due to the annotators' lack of expertise in the medical domain. As for the annotation based on the guidelines developed by the LLM, several inconsistencies were reported. The lack of clarity in the guidelines led one annotator to assign attributes with morphosyntactic and grammatical categories to all events, while the other did not apply these attributes to events labeled under the "medical domain".

For both schemes, the primary source of variance was the annotation of temporal relations, particularly in the selection of the relation type. Additionally, although the rules regarding arrow directionality were clarified in the guidelines created by the human expert, the annotators did not always follow them. The scheme developed by the LLM did not specify any rules on this matter, further contributing to variance among the annotators.

The Inter-Annotator Agreement (IAA) results were consistent with the curator's observations. For event annotation based on the human expert's scheme, the exact match at span-level was 72%, with agreement on the label in 81% of cases. Regarding the attributes of the specialized event class in the medical domain, agreement reached 85%.

In contrast, for the LLM-generated annotation scheme, the exact match percentage at span-level 64%, with agreement on the label in 78% of cases. Annotators agreed on the attributes of the specialized event class in only 61% of cases.

These findings are indicative that the human expert's annotation scheme provided clearer guidelines, resulting in more consistent annotations. The lower agreement observed with the LLM-generated scheme suggests reduced reliability, particularly in domain-specific event classification, likely due to ambiguous or inconsistently applied definitions.

4.2. Likert-scale performance assessment tool

As mentioned in Subsubsection 3.2, the two annotators were asked to evaluate each scheme and their respective annotation guidelines using Likert scales after the annotation process. The average evaluation results and the corresponding standard deviations are presented in Table 3.

The scheme developed by the human expert demonstrated superior performance across all evaluated parameters, consistently achieving the highest ratings, particularly in the identification and classification of medical domain events, temporal expressions, and temporal relations. Regarding ambiguity resolution,

Table 3
Means and standard deviations of the results obtained from the evaluation of the schemes/guidelines developed by LLM (Prompt 3B) and by Human.

	Evaluation			
	Human		LLM	
Annotator	$\bar{x}$	σ	$\bar{x}$	σ
Annotator 1	4.1	0.7	1.8	0.4
Annotator 2	4.5	0.5	2.5	0.8

Annotator 2 characterized the human-generated scheme as "a robust and effective approach to resolve all ambiguities encountered", whereas Annotator 1 acknowledged that "reasonable strategies are offered, but a comprehensive approach is lacking". In terms of the relevance and illustrative capacity of the examples, Annotator 2 considered the guidelines produced by the human expert to "provide well-developed examples, with broad coverage and clear application". Annotator 1, however, noted that the guidelines "provide sufficient examples, but they do not cover the full diversity of scenarios or contain inconsistencies". With respect to the definition of markables in the human-generated guidelines, both annotators agreed that "the guidelines are clear in most cases, with some areas for improvement". Similarly, in the evaluation of the coherence and clarity of the guidelines, both annotators stated that "the guidelines are coherent and clear in most cases". Overall, the scheme and guidelines developed by the human expert received an average rating of 4.1 from Annotator 1 and 4.5 from Annotator 2.

Conversely, the scheme and guidelines produced by the LLM received lower ratings across all evaluated parameters. Regarding the medical domain tags, Annotator 1 noted that "identification is limited, with clear inconsistencies in classification", while Annotator 2 stated that "the tags are absent or poorly defined for the medical domain". In terms of ambiguity resolution, Annotator 2 observed that in this LLM-generated scheme "reasonable strategies are offered, but a comprehensive approach is lacking", whereas Annotator 1 argued that "no strategies are offered to resolve ambiguities". Concerning the quality of the examples, both annotators agreed that the guidelines "include few examples, without addressing complex cases or containing incorrect annotation". Regarding the LLM definition of markables, Annotator 2 found that "the guidelines are satisfactory but occasionally lack clarity", while Annotator 1 remarked that "the guidelines are partially clear but frequently ambiguous". As for the coherence and clarity of the guidelines, Annotator 2 considered that "the guidelines are reasonable but have gaps and inconsistencies", whereas Annotator 1 asserted that "the guidelines are inconsistent and difficult to understand". As a result, the LLM scheme and guidelines received an average rating of 1.8 from Annotator 1 and 2.5 from Annotator 2.

5. Conclusions and Future Work

This work assessed the performance of an LLM in generating a temporal annotation scheme capable of capturing morphosyntactic, semantic, and medical domain information in clinical narratives in European Portuguese, comparing the automatically generated scheme with that produced by a human expert. The study contributes to expanding research on LLM performance to Portuguese, being, to our knowledge, the only one comparing the annotation scheme and guideline generation capacity of an LLM with that of a human expert.

The results show that, although LLMs are capable of creating annotation guidelines, the guidelines produced by a human expert were more comprehensible, coherent, and clear. The lack of specificity and clarity in the LLM's guidelines led to inconsistencies in the annotation, particularly in the definition of markables and the resolution of ambiguities. A hybrid approach, combining the performance of LLMs with human validation, could mitigate the observed inconsistencies.

In future research, we aim to conduct a comprehensive analysis of human annotation results, with a particular emphasis on instances of annotator disagreement. Our objective is to identify the underlying factors contributing to these discrepancies and refine the human-designed annotation scheme to reduce

ambiguities. Furthermore, we plan to integrate additional automated evaluation metrics to complement manual assessments, focusing on aspects such as readability, clarity, and structural coherence. To enhance scientific reproducibility, we also intend to incorporate a broader range of large language models (LLMs), including open-source alternatives. Additionally, we seek to extend this study to other annotation layers, particularly the referential layer, to systematically capture participants and their relationships.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments on earlier versions of this paper. We express our gratitude to Inês Cantante and Ana Filipa Pacheco for conducting the annotations. This research was partially funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC) and by the Centre of Linguistics of the University of Porto, within the project UIDB/00022/2020 (DOI 10.54499/UIDB/00022/2020).

References

[1] O. Irrera, S. Marchesin, G. Silvello, Metatron: Advancing biomedical annotation empowering relation annotation and collaboration, BMC Bioinformatics 25 (2024) 1–41. doi:10.1186/s12859-024-05730-9.

[2] C. Lindvall, C.-Y. Deng, E. Moseley, N. Agaronnik, A. El-Jawahri, M. K. Paasche-Orlow, J. R. Lakin, A. Volandes, J. A. Tulsky, Natural language processing to identify advance care planning documentation in a multisite pragmatic clinical trial, Journal of Pain and Symptom Management 63 (2022) e29–e36. doi:10.1016/j.jpainsymman.2021.06.025.

[3] E. Zhu, Q. Sheng, H. Yang, Y. Liu, T. Cai, J. Li, A unified framework of medical information annotation and extraction for chinese clinical text, Artificial Intelligence in Medicine 142 (2023) 1–12. doi:10.1016/j.artmed.2023.102573.

[4] N. Ide, Introduction: The handbook of linguistic annotation, in: N. Ide, J. Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer, Dordrecht, 2017, pp. 1–18.

[5] G. Moharasan, T.-B. Ho, Extraction of temporal information from clinical narratives, Journal of Healthcare Informatics Research 3 (2019) 220–244. doi:10.1007/s41666-019-00049-0.

[6] O. Bodenreider, The unified medical language system (umls): Integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[7] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, H. Liu, Large language models for data annotation and synthesis: A survey, 2024. URL: https://arxiv.org/abs/2402.13446. arXiv:2402.13446.

[8] A. H. Nasution, A. Onan, Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks, IEEE Access 12 (2024) 71876–71900. doi:10.1109/ACCESS.2024.3402809.

[9] N. Pangakis, S. Wolken, N. Fasching, Automated annotation with generative ai requires validation, 2023. URL: https://arxiv.org/abs/2306.00176. arXiv:2306.00176.

[10] A. Roberts, R. J. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, A. Setzer, Building a semantically annotated corpus of clinical texts, Journal of biomedical informatics 42 5 (2009) 950–66. URL: https://api.semanticscholar.org/CorpusID:17473913.

[11] P. Patel, D. Davey, V. Panchal, P. Pathak, Annotation of a large clinical entity corpus, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2033–2042. URL: https://aclanthology.org/D18-1228/. doi:10.18653/v1/D18-1228.

[12] W. F. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, J. Pustejovsky, Temporal annotation in the clinical domain, Transactions of

the Association for Computational Linguistics 2 (2014) 143–154. URL: https://aclanthology.org/Q14-1012/. doi:10.1162/tacl_a_00172.

[13] W. Sun, A. Rumshisky, O. Uzuner, Annotating temporal information in clinical narratives, Journal of Biomedical Informatics 46 (2013) S5–S12. URL: https://doi.org/10.1016/j.jbi.2013.07.004. doi:10.1016/j.jbi.2013.07.004.

[14] International Organization for Standardization, Iso 24617:2012 - language resource management – semantic annotation framework, 2012.

[15] D. Albright, A. Lanfranchi, A. Fredriksen, I. WilliamF.Styler, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. D. Nielsen, J. H. Martin, W. H. Ward, M. Palmer, G. K. Savova, Towards comprehensive syntactic and semantic annotations of the clinical narrative, Journal of the American Medical Informatics Association : JAMIA 20 (2013) 922 – 930. URL: https://api.semanticscholar.org/CorpusID:15409975.

[16] A. Miranda-Escalada, M. López-Arevalo, J. Armengol-Estapé, M. T. Martín-Valverde, F. Sanz, L. I. Furlong, M. Krallinger, A clinical gold standard corpus in spanish: Mining adverse drug reactions, Journal of Biomedical Informatics 56 (2015) 318–332. URL: https://doi.org/10.1016/j.jbi.2015.06.016. doi:10.1016/j.jbi.2015.06.016.

[17] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, A. Névéol, A french clinical corpus with comprehensive semantic annotations: Development of the medical entity and relation limsi annotated text corpus (merlot), Language Resources and Evaluation 52 (2018) 571–601. URL: https://doi.org/10.1007/s10579-017-9382-y. doi:10.1007/s10579-017-9382-y.

[18] J. V. A. d. Souza, Y. B. Gumiel, L. E. S. e. Oliveira, C. M. C. Moro, Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups, in: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), 2019, pp. 318–323. URL: https://doi.org/10.5753/sbcas.2019.6269. doi:10.5753/sbcas.2019.6269.

[19] L. E. S. e. Oliveira, A. C. Peters, A. M. P. da Silva, C. P. Gebeluca, Y. B. Gumiel, L. M. M. Cintho, D. R. Carvalho, S. Al Hasan, C. M. C. Moro, Semclinbr—a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks, Journal of Biomedical Semantics 13 (2022) 13. URL: https://doi.org/10.1186/s13326-022-00269-1. doi:10.1186/s13326-022-00269-1.

[20] N. C. Rocha, A. M. P. Barbosa, Y. O. Schnr, J. Machado-Rugolo, L. G. M. de Andrade, J. E. Corrente, L. V. de Arruda Silveira, Natural language processing to extract information from portuguese-language medical records, Data 8 (2023) Article 1. URL: https://doi.org/10.3390/data8010011. doi:10.3390/data8010011.

[21] F. Lopes, C. Teixeira, H. Gonçalo Oliveira, Contributions to clinical named entity recognition in portuguese, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, 2019, pp. 223–233. URL: https://doi.org/10.18653/v1/W19-5024. doi:10.18653/v1/W19-5024.

[22] M. Nunes, J. Boné, J. C. Ferreira, P. Chaves, L. B. Elvas, Medialbertina: An european portuguese medical language model, Computers in Biology and Medicine 182 (2024) 109233. URL: https://doi.org/10.1016/j.compbiomed.2024.109233. doi:10.1016/j.compbiomed.2024.109233.

[23] P. Silvano, A. Leal, F. Silva, I. Cantante, F. Oliveira, A. Jorge, Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus, in: H. Bunt (Ed.), Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Association for Computational Linguistics, Groningen, The Netherlands (online), 2021, pp. 1–13. URL: https://aclanthology.org/2021.isa-1.1/.

[24] A. Leal, P. Silvano, E. Amorim, I. Cantante, F. Silva, A. Jorge, R. Campos, The place of iso-space in text2story multilayer annotation scheme, in: H. Bunt (Ed.), Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022, European Language Resources Association, 2022, pp. 61–70. URL: https://aclanthology.org/2022.isa-1.8.

[25] P. Silvano, E. Amorim, A. Leal, I. Cantante, F. d. Silva, A. Jorge, R. Campos, S. S. Nunes, Annotation and visualisation of reporting events in textual narratives, in: Proceedings of Text2Story 2023: Sixth Workshop on Narrative Extraction From Texts, CEUR Workshop Proceedings, Dublin, Ireland, 2023, pp. 47–59.

[26] P. Silvano, E. Amorim, A. Leal, I. Cantante, A. Jorge, R. Campos, N. Yu, Untangling a web of temporal relations in news articles, in: Proceedings of Text2Story - Seventh Workshop on Narrative Extraction From Texts Held in Conjunction with the 46th European Conference on Information Retrieval (ECIR 2024), 2024, pp. 77–92. URL: https://repositorio-aberto.up.pt/handle/10216/158767.

[27] M. A. Leite, Ontology-Based Extraction and Structuring of Narrative Elements from Clinical Texts, Mater´s thesis, Universidade do Porto, 2024.

[28] S. Sheikholeslami, M. Meister, T. Wang, A. H. Payberah, V. Vlassov, J. Dowling, Autoablation: Automated parallel ablation studies for deep learning, in: Proceedings of the 1st Workshop on Machine Learning and Systems, 2021, pp. 55–61. URL: https://doi.org/10.1145/3437984.3458834. doi:10.1145/3437984.3458834.

[29] R. Likert, A technique for the measurement of attitudes, Archives of Psychology, Nova Iorque, 1932.

[30] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, I. Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, in: D. Zhao (Ed.), Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2018, pp. 5–9. URL: https://aclanthology.org/C18-2002.

[31] R. Artstein, Inter-annotator agreement, in: N. Ide, J. Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer Netherlands, 2017, pp. 297–313. URL: https://doi.org/10.1007/978-94-024-0881-2_11. doi:10.1007/978-94-024-0881-2_11.