AnnoHID: LLM-Assisted Annotation Framework for Low-Resource Medical Texts

Annisa Maulida Ningtyas[1], Guntur Budi Herwanto[1], Yunita Sari[1], Rifki Afina Putri[1]
Filip Kovacevic[2], Alaa El-Ebshihy[2], Varvara Arzt[2], Florina Piroi[2]
[1]Universitas Gadjah Mada, Indonesia [2]TU Wien, Austria
{annisamaulidaningtyas, gunturbudi, yunita.sari, rifki.putri}@ugm.ac.id

Abstract

This paper introduces AnnoHID, a semi-automated annotation framework designed for medical texts in low-resource languages. The system integrates large language models (LLMs) for pre-annotation and human validation to support efficient and consistent annotation. We demonstrate its application to Bahasa Indonesia medical social media texts from Alodokter, a medical Q&A platform, for Named Entity Recognition (NER) and Medical Concept Normalization (MCN). We conducted a user study with 21 participants to demonstrate the effectiveness of AnnoHID. The results show that LLM-assisted annotation yields higher inter-annotator agreement for both NER ($\kappa = 0.76$) and MCN ($\kappa = 0.63$) and that human review improves raw LLM NER output, raising the F1 score from 0.39 to 0.45. However, LLM assistance did not reduce annotation time and may introduce normalization bias in MCN. The framework is multilingual, human-in-the-loop, and interoperable with standard medical terminologies, such as SNOMED-CT. Future work focuses on mitigating pre-annotation bias, reducing annotation overhead, and scaling evaluations to larger datasets and additional low-resource languages.

1 Introduction

Social media platforms allow laypersons to search for, communicate with medical experts and discuss health information, helping them better understand medical conditions and improve their health literacy (Fage-Butler and Nisbeth Jensen, 2016; Ningtyas et al., 2024). However, there is a language gap between popularized terms used by laypersons on these platforms and specialized terms used by medical experts (e.g., "skin itch" vs. "paresthesia"), which leads to difficulties in interpreting and aggregating patient-related information. The task of Medical Concept Normalization (MCN) addresses this gap by detecting popularized medical phrases in user-generated text and linking them to standardized medical concepts in controlled vocabularies such as SNOMED-CT (Miftahutdinov and Tutubalina, 2019).

To train MCN models we need high-quality annotated data that links lay expressions to standardized medical concepts. Creating such datasets is costly, time-consuming, and limited by medical experts availability. For low-resource languages, the scarcity of training data becomes critical. Therefore, in this work, we propose an annotation system for efficient and high-quality MCN training dataset creation for low-resource settings, focusing on Indonesian language or Bahasa Indonesia.

Our proposed platform, AnnoHID[1], reduces annotation cost and time by combining LLM-assisted pre-annotation with human validation. This makes reliable annotated datasets more feasible in domains and languages where resources are scarce, supporting downstream medical NLP applications. Our contributions in this work are: 1) a semi-automatic NER and MCN annotation pipeline that uses LLMs for pre-annotation to reduce manual effort, 2) an adaptable framework designed for low-resource languages, with a primary focus on Bahasa Indonesia and the potential for extension to other languages (by substituting the target language corpus of the English system prompts), and 3) a human-in-the-loop design that ensures high-quality annotations through expert validation.

2 Related Work

Annotated datasets exist for English but are scarce for low-resource languages. As shown in Table 1, several tools support medical text annotation, including general-purpose platforms such as BRAT (Stenetorp et al., 2012) and INCEp-

[1]A demo of the system is available at https://medical.komunitech.id/. The installation is publicly available at https://github.com/gunturbudi/annohid. The annotation platform is released under the MIT License.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 683–691

TION (Klie et al., 2018) and domain-specific systems such as MedCAT (Kraljevic et al., 2021), CLAMP (Soysal et al., 2018), cTAKES (Savova et al., 2010), and SciSpaCy (Neumann et al., 2019). These tools provide features such as dictionary lookup, machine learning–based entity recognition, and integration with medical knowledge bases. However, these systems are developed and evaluated for English and rely on (training) data that are simply not available for low-resource languages. In addition, support for efficient semi-automatic workflows with human validation, and integrated NER and concept normalization remains limited. For these reasons, existing solutions are often difficult to adapt to new languages and domains. Our system addresses these limitations by providing an integrated, LLM-assisted, human-in-the-loop framework for low-resource MCN annotation.

We benchmark our system against existing annotation tools based on a set of dimensions. Table 1 summarizes five descriptive features manually extracted from the relevant publications: External KBs, NER model / approach, benchmark treebanks / datasets, supported entity types, and tool / framework dependencies. We also analyze four categorical features: source code availability, executable availability, graphical user interface (GUI), and API support. Each such feature is encoded using three availability levels: green (available), red (not available), and black (information not reported).

Our system provides additional capabilities not commonly reported in prior tools, including integrated LLM-assisted pre-annotation with human-in-the-loop validation, end-to-end support for both NER and medical concept normalization within a single workflow, and explicit design for low-resource languages.

3 System Overview

AnnoHID is intended to be used by medical NLP researchers and annotators who are working in low-resource languages. There are two roles in the system: administrator and annotator.

3.1 Admin Side

The administrator role provides access to several core features, including data and user management. Here, administrators can upload text documents to be annotated, create annotator accounts, and assign specific documents to individual annotators. An administrator can access the annotation performance report.

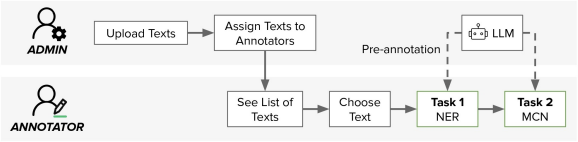

Figure 1: General annotation workflow.

3.2 Annotator Side

In this role, there are two main annotation tasks: Named Entity Recognition (NER) and Medical Concept Normalization (MCN). An admin will assign to each annotator a number of LLM-pre-annotated texts. The annotation process has the steps shown in Figure 1.

Step 0: Data Selection In this step, the annotator chooses one text from the list of his or her assigned texts. These assigned texts are pre-selected and distributed by the admin, ensuring that each annotator works on the specific documents allocated to them.

Step 1: NER annotation The annotator may choose to perform manual annotation or utilize the LLM-assisted annotation mode. When the latter is selected, an LLM pre-annotates the given text. In this step we used Gemini-2.5 Flash in combination with LangExtract[5]. The annotator then reviews these generated annotations which they may accept or reject as shown in Figure 2. If a suggestion is rejected, the annotator may revise or re-annotate the corresponding span of text.

Step 2: MCN annotation After an annotator completes the NER task, she proceeds to link the cleaned entities to medical concepts in SNOMED-CT. To find this link, we use the same LLM, Gemini-2.5 Flash, as for pre-annotations. Since SNOMED-CT is not yet available in Bahasa Indonesia[6], the LLM runs a three-step normalization pipeline: (1) translates the annotated entity span into English, which may still result in lay or colloquial terms; (2) normalizes the translated term into standard medical terminology in English; and (3) maps it to the corresponding SNOMED-CT concept ID. The SNOMED-CT API is then used to retrieve the appropriate concept IDs. The annotators may also accept or reject the suggested linking. If they reject it, they can manually correct the result

[5] https://github.com/google/langextract

[6] Currently, the Ministry of Health in Indonesia is in the process of translating SNOMED-CT into Bahasa Indonesia

Tool	External KBs	NER Model/Approach	Benchmark Treebanks/Dataset	Entity Types/Classes	Tool and framework dependency	Comment
Pain symptoms (Wang et al., 2022)	-	dict lookup, CRF	-	Pain	CLAMP	
Model for Stanza (Zhang et al., 2021)	-	LSTM	CRAFT, GENIA, MIMIC	Disease, Problem, Treatment, others[2]	CoreNLP	
MedCAT (Kraljevic et al., 2021)	UMLS, SNMD	dict lookup, Word2Vec, Bio ClinicalBERT	MedMentions, MIMIC III, ShARe (MIMIC II)[3]	Different UMLS semantic types	SciSpaCy	
SciSpaCy (Neumann et al., 2019)	-	transition-based chunking (Lample et al., 2016)	MedMentions, BC5CDR, CRAFT, JNLPBA, BioNLP13CG, AnatEM, BC2GM, BC4CHEMD, Linnaeus, NCBI-Disease	Chemicals, Disease, Cell type, Cell line, Protein, Gene, DNA, RNA, Cancer genetics, others[4]	spaCy	
SemEHR (Wu et al., 2018)	UMLS	not evident	MIMIC III	Different UMLS semantic types	Bio-YODIE	semantic search system
Bio-YODIE (Gorrell et al., 2018)	UMLS	-	MIMIC II	-	-	EL only
Adverse Drug Events (Li et al., 2018)	-	BiLSTM + CRF	MADE 1.0	Medication, Indication, Frequency, Severity, Dosage, Duration, Route, ADE, SSLI	-	no EL
CLAMP (Soysal et al., 2018)	UMLS	CRF, dict lookup, regex	i2b2, MTSampes, UTNotes	Problem, Treatment, Test	BRAT, OpenNLP, UIMA, CARD	
INCEpTION (Klie et al., 2018)	HPO, SNMD	dynamic label suggestions	-	no classification	Spring Boot	customizable General-purpose annotation platform
Sophia (Divita et al., 2014)	UMLS	dict lookup	Problem Reference Standard Corpus(?), i2b2	no classification	UIMA-AS, SPECIALIST Text Tools, LVG tools, v3NLP, conText	no POS tagging, exact matching
Body sites (Dligach et al., 2014)	UMLS	POS tagger & dict lookup	SHARP, ShARe	UMLS entity types	cTAKES	cTAKES-based dict lookup
BRAT (Stenetorp et al., 2012)	-	-	-	no specific	STAV	general-purpose manual annotation tool with automatic structured annotation import features
MER - MetaMap+ (Ben Abacha and Zweigenbaum, 2011)	UMLS, Wikipedia, Health on the Net, Biomedical Entity Network	POS tagging & dict lookup	i2b2	Problem, Treatment, Test	MetaMap	No tool, method evaluation only
MER - TT-SVM (Ben Abacha and Zweigenbaum, 2011)	-	POS tagging + SVM	i2b2	Problem, Treatment, Test	libSVM	
MER - BioCRF (Ben Abacha and Zweigenbaum, 2011)	-	CRF	i2b2, Berkely	Problem, Treatment, Test	CRF++	
MER - MetaMap + BioCRF (Ben Abacha and Zweigenbaum, 2011)	UMLS	MetaMap + CRF	i2b2	Problem, Treatment, Test	MetaMap, CRF++	
Concept-level IE (D'Avolio et al., 2011)	UMLS	entity type selection + CRF	i2b2	UMLS entity types	cTAKES	
cTAKES (Savova et al., 2010)	UMLS	POS tagger & dict lookup	PTB, GENIA, Mayo + Clinic EMR	UMLS entity types	SPECIALIST Lexical Tools, OpenNLP	

Table 1: Feature-based comparison of semi-automatic (medical) annotation tools from the literature

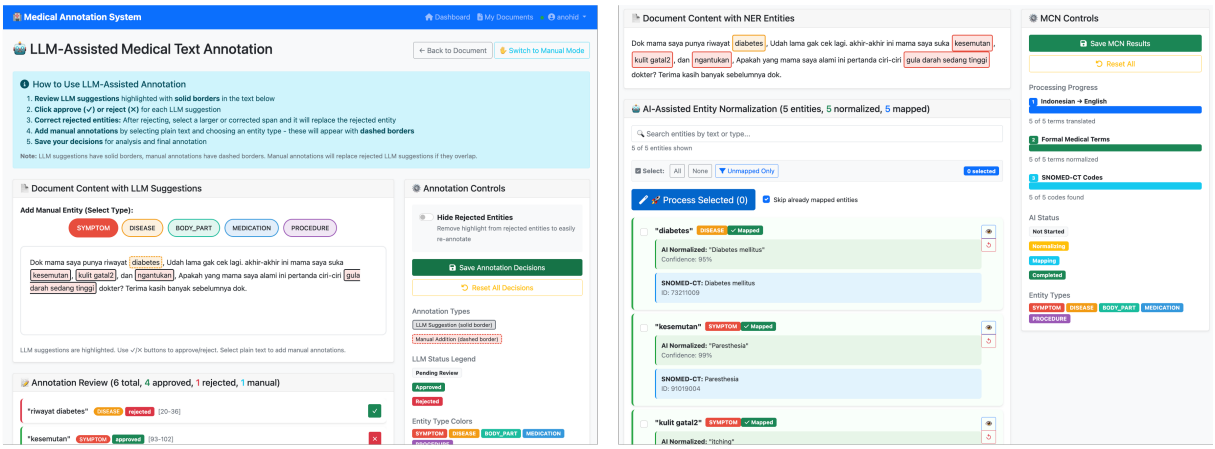

Figure 2: Annotator side interfaces. (Left) The LLM-assisted NER interface for entity extraction and manual verification. (Right) LLM-assisted MCN interface for mapping extracted entities to SNOMED-CT concepts.

by searching for the appropriate normalized term using our system's search feature. The search will return several candidate medical concepts along with their IDs, and the annotators can then select the most suitable one.

4 Evaluation

4.1 Experimental Setup

We conducted a user study with 21 annotators to assess how effectively the system supports the annotation task. The study compared LLM-assisted annotation against manual annotation to evaluate whether LLM assistance improves annotation quality. Our evaluation utilised 23 posts from Alodokter, an Indonesian medical Q&A platform, focusing on diabetes mellitus, tuberculosis, and stroke cases, as these conditions represent the primary drivers of mortality and morbidity in Indonesia[7]. Posts were selected to represent diverse lay medical terminology and varying linguistic complexity.

The annotation workflow consisted of two sequential tasks: First, a Named Entity Recognition (NER) task required annotators to identify lay medical terms and phrases within user questions, then classify them as predefined entity types, including symptoms, diseases, medications, procedures, and body parts. Second, a Medical Concept Normalisation (MCN) task required annotators to map each identified entity to its corresponding SNOMED-CT description and code.

4.2 User Study

4.2.1 Participants

We recruited 21 participants: one expert medical coder[8] and twenty undergraduate students enrolled in health information management study programs. All of the students had completed coursework in medical terminology and clinical coding courses; however, their practical annotation experience was limited. All participants were compensated for their time and contribution to the study.

Participants were randomly assigned to two groups of ten. The manual annotation group, which included an expert, labeled entities without assistance. The second, LLM-assisted group looked at the LLM model's predictions and decided whether to accept, reject, or change them, and added any entities that the model overlooked.

4.2.2 User Study Workflow

A three-hour workshop was conducted in a single session. The session began with an overview of the workshop objectives, followed by instruction on the annotation schema and entity definitions. We provided group-specific demonstrations of the system workflow and annotation procedures. Throughout the session, annotators were permitted to take breaks as needed and to ask questions regarding the interface and task instructions.

For the NER task, the manual group labeled entity spans directly on the user text without assistance, while the LLM-assisted group reviewed and verified LLM-generated pre-annotations. For the

[7] https://kemkes.go.id/eng/indonesia-health-profile-2024-moh

[8] The expert annotator in the control group established the benchmark data for all participants to ensure a fair comparison between student annotators.

MCN task, both groups used the same integrated SNOMED-CT search feature; however, the manual group was required to independently translate Indonesian entities into English, prior to searching [9], whereas the LLM-assisted group received LLM-generated normalized terms that were automatically mapped to SNOMED-CT candidates via the SNOMED-CT API, which they could then accept, reject, or change.

4.3 Evaluation Metrics

We evaluated annotation quality with two complementary sets of measures. First, we assessed the inter-annotator agreement with Cohen's κ to quantify consistency among annotators within each group. Second, we computed precision, recall, and F1-score to measure each annotator's performance against the expert's annotations, which served as the ground truth.

4.4 Analysis

The effectiveness of the LLM-assisted annotation system is evaluated along three dimensions. On one, we measure inter-annotator agreement by aggregating the pairwise consistency between students within each group and task. On the second, we assess how well the aggregated student annotations align with the expert ones. On the third dimension, we examine the agreement between the raw LLM-generated and the student annotations across each group and task.

4.4.1 Inter-Annotator Agreement (IAA)

We measure the IAA to evaluate the consistency among annotators in each group. The group of manual annotators (annotators 1–10) worked without LLM assistance, while the LLM-assisted group (annotators 11–20) received pre-annotations from Gemini 2.5 Flash that they could accept, reject, modify, or add new annotations manually.

We calculated pairwise Cohen's κ for both tasks. For the NER task, κ is computed at the character level, where each character position in the user post is assigned the entity label following the BIO tagging scheme. In contrast, entities are matched by exact span boundaries and label, then κ is computed over the assigned SNOMED-CT concept IDs. We reported the aggregate results in Table 2. The LLM-assisted group yields higher means κ for both

Group	Task	Mean κ	Median	Std
Manual	NER	0.71	0.71	0.16
	MCN	0.52	0.50	0.32
LLM-Assisted	NER	**0.76**	0.81	0.23
	MCN	**0.63**	0.63	0.28

Table 2: Inter-Annotator Agreement (Cohen's κ)

Group	Schema	Precision	Recall	F1
Manual	strict	0.55	0.54	0.53
	exact	0.58	0.57	0.56
	partial	0.72	0.72	0.71
	ent type	0.78	0.78	0.77
LLM-Assisted	strict	0.65	0.68	**0.66**
	exact	0.68	0.70	**0.68**
	partial	0.76	0.79	**0.76**
	ent type	0.76	0.80	**0.77**

Table 3: Pairwise Annotation Consistency of NER Task

NER (0.76 vs. 0.71) and MCN (0.63 vs. 0.52), indicating that LLM pre-annotation contributes to more consistent agreement among annotators within the LLM-assisted group, particularly for the medical concept normalization task. Both tasks show substantial agreement within each group, suggesting that the annotation guidelines were sufficiently clear to the annotators.

We acknowledge that character-level κ in NER may overrepresent the true agreement, as the majority of character positions are likely labeled as O (outside). Therefore, we further evaluated annotator consistency at the entity level by computing pairwise F1 using `nervaluate`[10], based on the SemEval 2013 evaluation metrics, across four evaluation schemas: *strict* (exact boundary and correct type), *exact* (exact boundary only), *partial* (overlapping boundary), and *ent_type* (correct type, boundary ignored). As shown in Table 3, the LLM-assisted group achieves higher F1 scores across all schemas, particularly for the strict (0.66) and exact (0.68) schemas.

Furthermore, for the MCN task, we adopt a multiclass performance evaluation: (1) entity spans are first matched between each annotator pair by exact boundary and label, then (2) the assigned SNOMED-CT concept IDs are compared. Span metrics measure entity detection consistency between annotator pairs (reported as F1-score), while *concept accuracy* measures the proportion of identical SNOMED-CT concept ID assignments among matched entity pairs. Table 4 shows that the LLM-assisted group achieves higher span F1 (0.70) and

[9]We recommended the use of DeepL https://www.deepl.com

[10]https://github.com/MantisAI/nervaluate

Group	Span P	Span R	Span F1	Concept Acc.
Manual	0.55	0.54	0.55	0.59
LLM-Assisted	0.71	0.69	0.70	0.70

Table 4: Pairwise Annotation Consistency of MCN Task

Stage	Schema	Precision	Recall	F1
Manual (no LLM)	strict	0.46	0.43	0.44
	exact	0.47	0.44	0.45
	partial	0.70	0.66	**0.67**
	ent type	0.77	0.72	**0.73**
LLM-Assisted (After Human Review)	strict	0.46	0.46	**0.45**
	exact	0.47	0.47	**0.46**
	partial	0.67	0.66	0.66
	ent type	0.72	0.71	0.70
Raw LLM	strict	0.41	0.37	0.39
	exact	0.42	0.38	0.39
	partial	0.63	0.57	0.59
	ent type	0.68	0.61	0.64

Table 5: Performance Against Expert of NER Task

concept accuracy (0.70) compared to the manual group (0.55 and 0.59, respectively).

Based on the results, we argue that LLM pre-annotation contributes to better boundary precision and concept normalization consistency among annotators in the LLM-assisted group compared to the manual group, as evidenced by higher F1 scores across all NER schemas and higher span F1 and concept accuracy in the MCN task.

4.4.2 Performance Against Expert Annotation

We compare each annotator's performance with the expert annotation on the same user posts. For each post, we treat the students' annotation as a prediction and the expert's ones as ground truth.

For NER (Table 5), the result shows that both groups perform similar to the expert across all schemas. The LLM-assisted group achieved slightly higher F1 score on the strict and exact schemas, indicating that LLM assistance may help annotators align more closely with the expert.

We also evaluate the raw performance of Gemini 2.5 Flash, comparing it against the expert annotation to establish a baseline for the LLM pre-annotation quality. The results show that it achieves strict F1 score of 0.39 against the expert, which improves to 0.45 after human review, confirming that human-in-the-loop review plays an important role in improving raw LLM output quality.

For the MCN task (Table 6), the LLM-assisted group achieves a slightly higher span F1 score, while the manual group demonstrates higher concept accuracy. The raw LLM achieves span F1 of 0.39 and a concept accuracy of 0.55, which

Stage	Span P	Span R	Span F1	Concept Acc.
Manual (no LLM)	0.51	0.38	0.43	**0.60**
LLM-Assisted (After Human Review)	0.53	0.41	**0.46**	0.54
Raw LLM	0.55	0.30	0.39	0.55

Table 6: Performance Against Expert of MCN Task

improves to 0.46 and slightly decreases to 0.54, respectively, after human review. Although the LLM pre-annotation helps the annotators detect more entities that are consistent with the expert, it may introduce normalization bias, where annotators tend to accept LLM-suggested SNOMED-CT concept IDs without critically evaluating the normalization output. This result is consistent with prior findings (Wang et al., 2024; Choi et al., 2024), which demonstrate that when LLM suggestions are incorrect, providing wrong labels negatively impacts human annotation accuracy; thus, the quality of the LLM pre-annotation directly influences the reliability of the final annotations. To mitigate this issue, future work will incorporate a minimum review time threshold per annotation task, where annotators spending less than this threshold will receive a pop-up prompt to discourage over-reliance on LLM-suggested SNOMED-CT concept IDs.

4.4.3 Annotation Time

We found that the LLM-assisted group did not annotate faster than the manual group. Table 7 shows that the LLM-assisted group needed slightly longer times on average compared to the manual group for both NER (133.2s vs. 117.8s) and MCN (275.1s vs. 262.8s). In particular, for the MCN task, we required the LLM-assisted annotators to manually initiate the LLM pre-annotation by clicking a button, which may contribute to the additional delay. Furthermore, we argue that the additional time may reflect the cognitive effort required to evaluate and verify LLM suggestions, rather than annotating from scratch, as annotators might spend additional time to evaluate the ambiguous or incorrect suggestions before accepting or rejecting them. Our findings align with Schroeder et al. (2025), where LLM-generated suggestions did not speed up annotation. To speed up the annotation, future work should consider automating the LLM pre-annotation rather than requiring manual input from annotators, especially for the MCN task, and investigate the relationship between annotation time and both LLM suggestion acceptance rate and correctness against expert annotations.

Group	Task	Mean (s)	Median (s)	Std	N
Manual	NER	117.8	90.0	85.2	220
	MCN	262.8	206.3	178.1	214
	Overall	**189.3**	**139.5**	**156.6**	434
LLM-Assisted	NER	133.2	101.0	99.7	200
	MCN	275.1	221.6	190.0	190
	Overall	**202.3**	**146.2**	**166.4**	390

Table 7: Estimated Annotation Time per User Post (sec)

5 Conclusion

This paper presents AnnoHID, a semi-automated annotation framework for medical texts in low-resource languages that combines large language model (LLM)-based pre-annotation with human-in-the-loop validation for named entity recognition (NER) and medical condition normalization (MCN) tasks. A user study of 21 participants using Bahasa Indonesia medical social media data shows that LLM assistance enhances inter-annotator agreement but may introduce normalization bias in MCN. Contrary to the hypothesis, annotation time did not decrease. Future work will focus on mitigating normalization bias, reducing annotation overhead, and extending the framework to larger datasets and additional low-resource languages beyond Bahasa Indonesia to validate the multilingual adaptability of the system.

Ethical Considerations

This study involves human annotators and publicly available user-generated medical texts. All annotators participated voluntarily and were compensated for their time and contribution. Participants were informed about the study objectives and annotation procedures prior to the experiment. The textual data used in this work were collected from a publicly accessible medical Q&A platform. To protect user privacy, we removed or anonymized any personally identifiable information (PII) that could directly or indirectly reveal user identities before annotation and analysis. The dataset was used solely for research purposes.

The proposed framework incorporates LLM to assist annotation. While LLM-generated suggestions can improve efficiency and consistency, they may also introduce biases or incorrect medical normalization outputs. To mitigate these risks, our system follows a human-in-the-loop design in which annotators review, accept, or correct model predictions, ensuring that final annotations remain under human decision. We acknowledge potential risks related to the misuse of medical NLP systems, including overreliance on automated outputs in clinical contexts. Our framework is intended strictly for research and dataset creation purposes, not for clinical decision-making.

Acknowledgments

This work was supported by the OeAD - Austria's Agency for Education and Internationalisation on behalf of ASEA-UNINET (Project ASEA 38-2024) and TU Wien.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Medical Entity Recognition: A Comparaison of Semantic and Statistical Methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA. Association for Computational Linguistics.

Alexander S Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The llm effect: Are humans truly using llms, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054.

Leonard W D'Avolio, Thien M Nguyen, Sergey Goryachev, and Louis D Fiore. 2011. Automated concept-level information extraction to reduce the need for custom software and rules development. *Journal of the American Medical Informatics Association*, 18(5):607–613.

Guy Divita, Qing T Zeng, Adi V. Gundlapalli, Scott Duvall, Jonathan Nebeker, and Matthew H. Samore. 2014. Sophia: A Expedient UMLS Concept Extraction Annotator. *AMIA Annual Symposium Proceedings*, 2014:467–476. Sophia.

Dmitriy Dligach, Steven Bethard, Lee Becker, Timothy Miller, and Guergana K Savova. 2014. Discovering body site and severity modifiers in clinical texts. *Journal of the American Medical Informatics Association*, 21(3):448–454.

A. M. Fage-Butler and M. Nisbeth Jensen. 2016. Medical terminology in online patient–patient communication: evidence of high health literacy? *Health Expectations*, 19(3):643–653.

Genevieve Gorrell, Xingyi Song, and Angus Roberts. 2018. Bio-YODIE: A Named Entity Linking System for Biomedical Text. *arXiv preprint*. ArXiv:1811.04860 [cs].

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*,

pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics. INCEpTION.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Fei Li, Weisong Liu, and Hong Yu. 2018. Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning. *JMIR Medical Informatics*, 6(4):e12159. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Z. Miftahutdinov and E. Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proc. of the 57th Annual Meeting of the ACL: Student Research Workshop*, pages 393–399, Florence, Italy. ACL.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. ArXiv:1902.07669 [cs].

Annisa Maulida Ningtyas, Alaa El-Ebshihy, Florina Piroi, and Allan Hanbury. 2024. Improving laypeople familiarity with medical terms by informal medical entity linking. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 113–126, Cham. Springer Nature Switzerland.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513. CTAKES.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336. CLAMP.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics. BRAT.

Xiaoyan Wang, A.D Dave, G Ruaño, and J Kost. 2022. Automated Extraction of Pain Symptoms: A Natural Language Approach using Electronic Health Records. *Pain Physician*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard Jb Dobson. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. *Journal of the American Medical Informatics Association*, 25(5):530–537.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

A LLM Pre-Annotation Prompt for NER Task

The following prompt was used for LLM-based pre-annotation in the NER task, employing Gemini-2.5 Flash to identify medical entities, including overlapping spans in user posts.

Prompt: NER Pre-Annotation

System: You are an expert medical text annotator specializing in detecting overlapping medical entities. Your task is to identify ALL medical entities in the text, including entities that overlap or share the same text spans.

Entity Types:

- `SYMPTOM`: Medical symptoms, complaints, or clinical presentations
- `DISEASE`: Diseases, disorders, conditions, or diagnoses
- `BODY_PART`: Anatomical structures, organs, or body regions
- `MEDICATION`: Drugs, medications, or therapeutic substances
- `PROCEDURE`: Medical procedures, treatments, or interventions

Critical Instructions for Overlapping Entities:

1. Entities can overlap — the same text can belong to multiple entities
2. Look for modifier-noun combinations (e.g., "severe chest pain" contains `BODY_PART` + `SYMPTOM` with modifier)
3. Look for compound entities (e.g., "metformin 500mg" is `MEDICATION` with dosage included)
4. Identify both atomic and compound entities
5. Include all meaningful medical entities, even if they overlap
6. Include modifiers (severity, frequency, dosage) as part of the entity text
7. For body parts within larger medical terms, extract both the body part and the full term

Examples of Overlapping Entities:

- "severe chest pain": `BODY_PART`("chest"), `SYMPTOM`("pain"), `SYMPTOM`("chest pain"), `SYMPTOM`("severe chest pain")
- "metformin 500mg daily": `DISEASE`("diabetes"), `MEDICATION`("metformin 500mg"), `MEDICATION`("metformin")
- "cardiac catheterization": `BODY_PART`("cardiac"), `PROCEDURE`("cardiac catheterization")

Output Format (JSON array, no markdown):

```
[{
  "text": "exact text span",
  "start_offset": character_position,
  "end_offset": character_position,
  "entity_type": "TYPE",
  "confidence": confidence_score_0_to_1,
  "reasoning": "brief explanation",
  "overlaps_with": [
    "other overlapping entity texts"
  ]
}]
```

Be exhaustive in finding overlapping entities. Include confidence scores and reasoning for complex or ambiguous cases. Ensure character positions are accurate.

B LLM Pre-Annotation Prompt for MCN Task

The following prompt was used for LLM-based pre-annotation in the MCN task. We employ Gemini-2.5 Flash to perform a three-step normalization pipeline for each identified entity.

Prompt: MCN Pre-Annotation

System: You are a medical AI assistant specializing in Indonesian medical terminology translation and normalization.

Task: Process the following Indonesian medical term through a 3-step pipeline.

Input:

- Indonesian Term: `{text}`
- Entity Type: `{entity_type}`
- Document Context: `{context}`

Step 1 – Indonesian to English Translation:

- If the term is already in English, mark it as such
- Provide the most accurate medical English translation
- Consider the entity type and context
- Provide confidence score (0.0–1.0)
- List alternative translations if applicable

Step 2 – Normalize to Formal Medical English:

- Convert informal/colloquial English to formal medical terminology
- Use standard medical nomenclature
- e.g., "belly" → "abdominal region", "headache" → "cephalgia"
- Provide confidence score (0.0–1.0)

Step 3 – SNOMED-CT Search Term:

- Provide the best search term for SNOMED-CT lookup
- Should be the most standardized medical term
- Optimize for SNOMED-CT database search

Entity Type Guidelines:

- `SYMPTOM`: Focus on clinical findings and patient complaints
- `DISEASE`: Focus on pathological conditions and diagnoses
- `BODY_PART`: Use anatomical terminology
- `MEDICATION`: Use generic pharmaceutical names
- `PROCEDURE`: Use standard medical procedure terminology

Output Format (JSON only, no markdown):

```
{
  "step1_translation": {
    "original": "{text}",
    "english": "<translated term>",
    "is_already_english": false,
    "confidence": 0.95,
    "alternatives": ["<alt1>", "<alt2>"],
    "notes": "<any relevant notes>"
  },
  "step2_normalization": {
    "informal": "<from step 1>",
    "formal": "<formal medical term>",
    "confidence": 0.90,
    "reasoning": "<brief explanation>"
  },
  "step3_snomed_search": {
    "search_term": "<optimized term>",
    "alternative_terms": ["<alt1>", "<alt2>"],
    "semantic_tag":
      "<disorder|finding|procedure|etc>"
  }
}
```

Return ONLY the JSON object, no additional text or markdown formatting.