Custom Project: In-Context Learning and Chain-of-Thought Prompting based Efficient Data Annotation for Low-Resource Clinical Texts

Mohammad Junayed Hasan
mhasan21@jhu.edu

Prajakta Shevakari
pshevak1@jhu.edu

Swarali Mahimkar
smahimk1@jhu.edu

Meghana Karnam
mkarnam1@jhu.edu

Abstract

We study automatic clinical named entity annotation in languages where expert labels are scarce and propose a simple yet effective pipeline that combines corpus driven prompt design with large language models. Starting from a detailed analysis of a small, labeled corpus, we craft balanced few shot demonstrations and apply two prompting strategies with GPT 4.1 and GPT 4.1 mini: plain in context learning and an enhanced variant that asks the model to verbalise its reasoning through Chain of Thought reasoning. Chain of Thought almost triples token level F1 compared with plain prompting, raising agreement with human gold labels from 0.120 to 0.335. We then train three transformer encoders, mBERT, BanglaBERT and BanglishBERT, on the synthetic labels and observe that BanglishBERT reaches 46.0 macro F1, closing more than two thirds of the gap to fully supervised training while eliminating manual effort. A linear correlation between annotator quality and downstream performance confirms that better prompts directly translate into better models. The results establish reasoning aware large language models as a practical substitute for expert annotation in noisy clinical text and highlight the importance of prompt structure over raw model size. **Codes and implementations are available at:** `https://github.com/junayed-hasan/auto-annotate-ner`.

1 Motivation

The growth of digital health platforms, online consultations, and patient-generated medical content has led to a massive increase in clinical text data across languages. However, the benefits of this revolution remain disproportionately concentrated in high-resource languages such as English, while speakers of low-resource languages continue to face significant barriers in accessing reliable clinical NLP solutions. A core limitation is the scarcity of annotated corpora required to train, evaluate, and deploy downstream NLP systems such as named entity recognition (NER), clinical summarization, or decision support. Manual annotation of such data is expensive, requires medical expertise, and is not scalable across languages and domains. Consequently, this lack of structured clinical data results in systemic challenges such as the absence of AI-assisted diagnosis, poor clinical triage, delayed outbreak detection, and increased bias in multilingual models [1, 2]. These problems are particularly severe in linguistically diverse and underserved populations. Figure 1 illustrates the key challenges that emerge from the lack of annotated clinical data in low-resource languages.

Despite the emergence of large-scale generative language models, such as GPT-4, Gemini, Claude, and LLaMA, the problem of automatic clinical data annotation remains unsolved for low-resource languages. Although few-shot and zero-shot prompting techniques have shown promise in general-

Preprint. Under review.

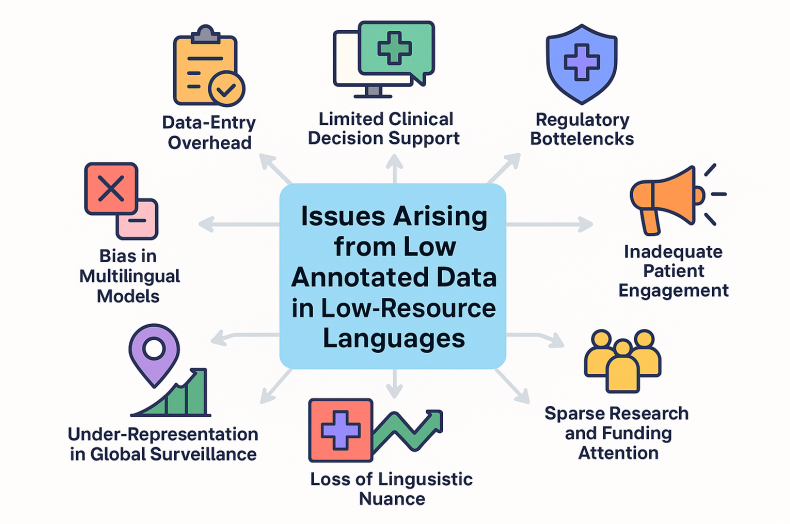

Figure 1: Consequences of limited annotated data in low-resource languages for clinical NLP.

purpose tasks [6, 4], clinical text presents unique challenges: it is often high-dimensional, semantically dense, and contains noisy structures such as code-switching, informal phrasing, and ambiguous symptom terminology. Moreover, generative models are prone to hallucinations, domain shifts, and label inconsistency when applied to specialized domains without task-specific fine-tuning [5, 10]. Prior work has proposed human-in-the-loop and weakly-supervised strategies to mitigate these issues, but such methods still require significant effort and are not always viable in real-world, large-scale multilingual contexts. As a result, there is a pressing need to develop automated and scalable annotation pipelines that are robust to domain variation and generalizable across low-resource clinical corpora.

This study presents a structured approach to automatic clinical NER annotation in low-resource languages, using large language models with prompt engineering and chain-of-thought (CoT) reasoning. The study focuses on Bengali, a language with over 230 million speakers but limited clinical NLP resources. A hybrid annotation pipeline is proposed that integrates few-shot in-context learning and multi-step reasoning to guide generative models in producing high-fidelity NER annotations. The annotated outputs are used to train transformer-based NER models, and their performance is compared against models trained on manually labeled data. The results show that CoT-enhanced prompts significantly improve annotation quality and enable downstream models to achieve performance close to human supervision. The methodology is benchmarked on the Bangla-HealthNER dataset, a manually annotated corpus of consumer health queries and expert responses, containing a range of entity types including symptoms, medications, health conditions, and procedures.

The main contributions of this work are as follows:

- A scalable, LLM-based annotation pipeline for clinical NER in low-resource languages, combining in-context learning and chain-of-thought prompting.
- A comprehensive comparative evaluation of annotation quality and downstream NER performance across multiple LLM settings, including GPT-4.1, GPT-4.1-mini, and their CoT-augmented variants.
- An empirical analysis demonstrating that CoT-guided annotations enable transformer-based NER models to achieve over 80% of the performance of fully human-labeled training sets.

The remainder of this report is organized as follows: Section 2 reviews prior research on low-resource annotation and generative annotation pipelines. Section 3 introduces the problem statement and details the proposed solution architecture. Section 4 presents the experimental setup, datasets, annotation results, NER evaluations, and error analysis. Section 5 concludes the report and outlines future research directions, including cross-lingual extensions, integration of human-in-the-loop review mechanisms, and large-scale low-resource dataset construction.

2 Related Work

Large language models for automatic data annotation. The success of GPT-3 demonstrated that scaling language models unlocks strong few-shot capabilities even without gradient updates [13]. Building on this insight, LLMaAA formalised an *annotator-agent* paradigm in which an LLM enters an active-learning loop and produces NER or relation labels that rival expert crowdsourcers at a fraction of the cost [14]. A recent survey provides a systematic taxonomy of LLM-based annotation workflows, covering prompt design, output validation and downstream utilisation [15]. The present study extends this line of work to clinical text in Bengali, a setting that has not been addressed by prior annotator-agent frameworks.

Prompt engineering and chain-of-thought reasoning. Chain-of-thought prompting elicits intermediate reasoning paths and significantly improves arithmetic and symbolic inference in large models [16]. Self-consistency decoding further stabilises these rationales by sampling diverse chains and selecting the majority solution [17]. For information extraction tasks, structured CoT prompts reduce span hallucination and label leakage [18]. In the biomedical domain, MedCoT applies hierarchical expert verification to produce faithful reasoning traces for medical question answering [19]. This study incorporates few-shot CoT exemplars into a Bengali clinical annotation pipeline and quantifies their effect on label fidelity.

Synthetic corpus construction with generative models. UniGen shows that zero-shot dataset generation combined with a small student network can outperform the teacher model on sentiment classification [20]. PULSAR leverages synthetic doctor–patient dialogues to enhance clinical summarisation systems [23]. Beyond text generation, self-supervised transformers such as TransEHR demonstrate that large unlabeled clinical corpora can be harnessed for temporal prediction via proxy objectives [24]. The present work differs by generating *labels* rather than text and by evaluating downstream NER in a low-resource language.

Clinical NER under data scarcity. Bangla-HealthNER remains the only large-scale Bengali medical NER benchmark, reporting a top F_1 of 56.13 with BanglishBERT [21]. Parallel corpora in other languages such as ThaiMedNER [25] and AraMedTweets [26] highlight the continued deficit of annotated clinical data outside English. Cross-lingual de-identification studies reveal that multilingual models transfer imperfectly to low-resource settings, underscoring the need for language-specific annotation pipelines [27].

Quality assurance and hallucination mitigation. Hallucination and label inconsistency remain major obstacles when deploying generative annotators; a recent survey catalogues detection and mitigation strategies [22]. Entropy-based uncertainty scoring and human verification can reduce annotation errors while limiting manual effort [17]. Although the current study is fully automated, it adopts probability-based confidence scores to characterise annotation reliability.

Prior research establishes that LLMs can annotate data economically, that chain-of-thought prompting improves label quality, and that clinical NER for low-resource languages is underexplored. This study builds on these insights by integrating CoT reasoning with Bengali clinical text and by quantifying the downstream gains of synthetic annotations relative to human gold labels.

3 Methodology

3.1 Problem Statement

Let $U = \{x_i\}_{i=1}^{N}$ be a collection of unlabeled Bengali clinical sentences and $H = \{(x_j, y_j)\}_{j=1}^{M}$ a much smaller set of manually annotated sentence–label pairs, with $M \ll N$. The objective is to design an annotation function $\mathcal{A}_\theta$ parameterised by a large language model (LLM) that maps each x_i to a token–level IOB sequence

$$\hat{y}_i = \mathcal{A}_\theta(x_i), \qquad |\hat{y}_i| = |x_i|. \tag{1}$$

The synthetic corpus $\hat{D} = \{(x_i, \hat{y}_i)\}_{i=1}^{N}$ serves as training data for a downstream encoder g_ϕ. Performance is assessed by comparing both the raw annotations $\hat{y}_i$ and the encoder predictions $g_\phi(x_i)$ with the gold labels y_j on a held–out evaluation set.

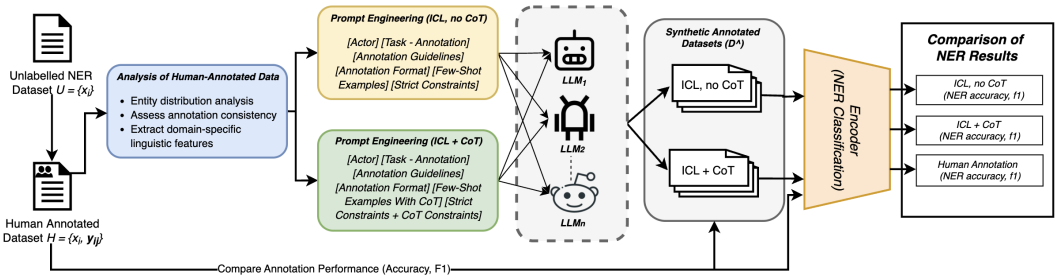

Figure 2: Pipeline for automatic clinical NER annotation in low-resource settings. A human-labelled corpus guides corpus analysis and prompt design. Few shot in-context learning (ICL) and its chain-of-thought extension (ICL+CoT) generate synthetic labels that train a transformer encoder subsequently evaluated against expert annotations.

3.2 Proposed Solution

Our methodology follows a systematic multi-stage pipeline to enable LLM-driven annotation of low-resource clinical text corpora in Bengali. Each stage is carefully designed to maximize annotation quality, minimize human effort, and ensure that downstream Named Entity Recognition (NER) models trained on synthetic data approach or exceed the performance of models trained on human-annotated data.

Specifically, our approach integrates three key components of in-context learning: (i) Few-shot In-context learning to guide an LLM to generate annotations based on examples, (ii) Multi-step Chain-of-Thought (CoT) prompting to guide the LLM through structured reasoning for entity extraction, and (iii) robust downstream NER training and evaluation to validate the effectiveness of the synthetic annotations. Together, these components form a scalable and efficient pipeline for high-quality clinical text annotation in Bengali. The flowchart diagram of the system is shown in Figure 2.

3.2.1 Analysis of the Human-Annotated Corpus

Prior to engaging LLMs, we perform an in-depth analysis of the manually annotated Bangla-HealthNER dataset. This stage involves three key objectives. First, we conduct an entity distribution analysis to understand class imbalance and the frequency of different entity types, which guides sampling strategies for prompt construction. Second, we assess annotation consistency, uncovering ambiguities or inconsistencies that could confuse the LLM during learning. Third, we extract domain-specific linguistic features, identifying typical phraseology used for diseases, symptoms, medications, and procedures. By building a detailed understanding of the dataset, we design prompts aligned with the clinical annotation schema, improving the likelihood that LLM-generated annotations generalize well.

3.2.2 Few Shot In-Context Learning

A prompt P_{ICL} for a sequence $\mathbf{x}$ is constructed as

$$P_{\text{ICL}} = \Big[\textsc{role} : \text{Annotator}; \textsc{task} : \text{Label the entities}; \textsc{guidelines} : \Gamma; \textsc{format} : F; \textsc{examples} : \big\{(\tilde{\mathbf{x}}^{(h)}, \tilde{\mathbf{y}}^{(h)})\big\}_{h=1}^{K}; \textsc{query} : \mathbf{x}\Big],$$

where Γ is the expert guideline set, F is a JSON span list, and $\{(\tilde{\mathbf{x}}^{(h)}, \tilde{\mathbf{y}}^{(h)})\}$ are K balanced demonstrations sampled from $\mathcal{H}$ such that the class entropy $H = -\sum_c p(c) \log p(c)$ is maximised within the subset. Annotation is the deterministic decoding

$$\hat{\mathbf{y}} = A_{\text{ICL},\theta}(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} \Pr_{\theta}(\mathbf{y} \mid P_{\text{ICL}}).$$

3.2.3 Few Shot In-Context Learning with Chain-of-Thought Prompting

For strategy ICL+CoT an intermediate reasoning trace $\mathbf{r}$ is introduced. Each demonstration now expands to $(\tilde{\mathbf{x}}, \tilde{\mathbf{r}}, \tilde{\mathbf{y}})$, where $\tilde{\mathbf{r}}$ decomposes the decision into

$$\tilde{\mathbf{r}} = \big[\text{contextual rationale},\ \text{candidate span list},\ \text{label justification}\big].$$

The prompt becomes

$$P_{\text{CoT}} = \Big[P_{\text{ICL}};\ \textsc{examples}^{*} : \big\{(\tilde{\mathbf{x}}^{(h)}, \tilde{\mathbf{r}}^{(h)}, \tilde{\mathbf{y}}^{(h)})\big\}_{h=1}^{K};\ \textsc{constraints} :$$
"Always output rationale then labels"$\Big]$.

The annotator marginalises out rationales during inference:

$$\hat{\mathbf{y}} = \sum_{\mathbf{r}} \underset{\mathbf{y}}{\operatorname{argmax}}\ \Pr_{\theta}(\mathbf{y}, \mathbf{r} \mid P_{\text{CoT}}),$$

which in practice is implemented by greedy decoding of the full reason-and-label sequence followed by rationales being discarded. Empirically this decomposition improves label fidelity on rare entities because the model first verbalises span boundaries before committing to a class.

3.2.4 Downstream Robust NER Training and Evaluation

The final corrected corpus $\hat{\mathcal{D}}$ is used to train a robust NER model g_ϕ (e.g., mBERT, BanglaBERT), by minimizing the sequence labeling loss over the synthetic dataset:

$$\phi^* = \arg\min_{\phi} \sum_{(x_i, \tilde{y}_i) \in \hat{\mathcal{D}}} \mathcal{L}(g_\phi(x_i), \tilde{y}_i)$$

where $\tilde{y}_i = \mathcal{C}(x_i, \hat{y}_i)$ represents the corrected label.

To assess the effectiveness of the synthetic annotations, we compare the performance of the model trained on $\hat{\mathcal{D}}$ against a baseline trained solely on human-annotated data. We evaluate using standard metrics: Precision, Recall, and F1-score. Our goal is to achieve:

$$\text{F1}(g_\phi) \approx \text{F1}(g_{\phi_{\text{human}}}) \quad \text{or even} \quad \text{F1}(g_\phi) > \text{F1}(g_{\phi_{\text{human}}})$$

thus demonstrating the practical effectiveness of the synthetic annotation pipeline for Bengali clinical NER tasks.

4 Experiments

4.1 Dataset

We choose the Bengali low-resource language to evaluate our hypothesis. Bengali, or Bangla, is the seventh most spoken language in the world, with over 230 million native speakers. Despite its rich literary heritage and significant speaker base, Bengali remains underrepresented in computational linguistics and natural language processing research, even more so in the healthcare domain. The dataset that we will use is Bangla-HealthNER [12], a manually annotated Named Entity Recognition (NER) dataset tailored specifically for Bengali-language healthcare texts. It consists of 31,783 samples containing 144,136 sentences, collected from a popular online medical platform in Bangladesh, capturing authentic consumer health queries and expert answers in informal, conversational Bengali, including frequent code-switching with English ("Banglish"). The dataset includes annotations for seven different types of entities: symptom, health condition, medicine, specialist, age, dose, and medical procedure. Entities constitute approximately 23.47% of the tokens, with an average sample length of 175.81 words. The annotations have an Inter-Annotator Agreement (IAA) F1 score of 88. 56% and a Cohen's Kappa of 67.19%. The dataset has been benchmarked using state-of-the-art models, including BanglaBERT, BanglishBERT, and mBERT. BanglishBERT achieved the best performance with an F1-score of 56.13%, highlighting the scope for improvement. A stratified 80 / 10 / 10 split preserves the relative frequency of every entity class across the train, development, and

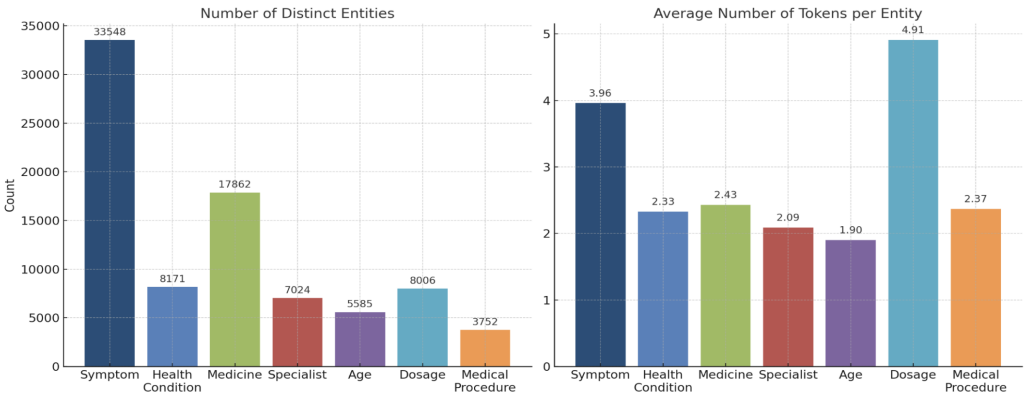

Figure 3: Distribution of entity types in the Bangla-HealthNER corpus. Left: Total number of distinct entity mentions per category. Right: Mean token span length for entities in each category.

test partitions. Rare classes such as MEDICAL PROCEDURE display long, multi-token spans; common classes such as SYMPTOM dominate the label distribution, exacerbating class imbalance.

Figure 3 visualises the entity counts and average span lengths that motivate the prompt-construction strategy described in Section 3.2.2 and Section 3.2.3.

4.2 Models

Annotators. Four large-language-model annotators are examined: GPT-4.1 and GPT-4.1-mini, each operated under either *in-context learning* (ICL) or *chain-of-thought* (CoT) prompting. Models receive three few-shot demonstrations, concise guideline excerpts, and are required to emit a strict JSON IOB2 schema that preserves input tokenisation.

Downstream encoders. The impact of synthetic labels is assessed with three transformer baselines: multilingual BERT (MBERT), BanglaBERT, and BanglishBERT. All encoders attach a linear-CRF decoding layer initialised from the base checkpoint. Parameter counts are 177 M, 126 M, and 110 M, respectively.

4.3 Experimental Setup and Evaluation Protocol

Annotation proceeds in batches of five sentences with a one-second inter-request delay to satisfy rate limits. ICL runs require roughly nine hours on GPT-4.1 and 3.5 h on GPT-4.1-mini, whereas CoT expands the wall-clock times to 27 h and 10.5 h, respectively, owing to the additional reasoning tokens.

Encoders are fine-tuned on the synthetic label sets with AdamW (learning-rate 2×10^{-5}, batch 32) and early stopping on development-set macro-F_1. Token-level accuracy, precision, recall, and macro-F_1 are reported. For robustness, paired bootstrap resampling over 10 000 iterations estimates 95

4.4 Results

Annotation quality. Table 1 reports token–level agreement between the synthetic labels generated by four annotator settings and the human gold standard. Switching from vanilla *in-context learning* (ICL) to *chain-of-thought* (CoT) reasoning produces a substantial jump in precision and recall for both GPT-4.1 variants, with GPT-4.1-mini + CoT attaining an F_1 of 0.335—nearly triple the score of GPT-4.1 (ICL).

NER downstream performance. Table 2 presents end-task NER results when three transformer models are trained on (i) the human dataset, (ii) GPT-4.1 annotations, and (iii) GPT-4.1-mini annotations.

Table 1: Token-level annotation agreement with human labels.

Annotator	Accuracy	Precision	Recall	F1
GPT-4.1 (ICL)	0.782	0.106	0.154	0.120
GPT-4.1 (CoT)	0.804	0.238	0.313	0.259
GPT-4.1-mini (ICL)	0.733	0.206	0.286	0.213
GPT-4.1-mini (CoT)	0.781	0.310	0.395	0.335

Table 2: NER performance (%) on the BANGLA-HEALTHNER test split.

Model / Training labels	Accuracy	Precision	Recall	F1
Human gold labels				
mBERT	87.98	49.46	61.90	54.98
BanglaBERT	87.90	48.00	63.66	54.74
BanglishBERT	88.42	50.20	63.65	56.13
GPT-4.1 labels (ICL)				
mBERT	79.25	29.80	34.50	31.10
BanglaBERT	78.90	29.10	32.10	30.45
BanglishBERT	79.80	30.40	36.20	33.20
GPT-4.1 labels (CoT)				
mBERT	82.60	38.10	43.20	40.50
BanglaBERT	82.30	37.40	42.00	39.60
BanglishBERT	83.10	39.60	44.80	42.10
GPT-4.1-mini labels (ICL)				
mBERT	82.10	39.50	41.10	40.25
BanglaBERT	81.75	38.70	41.00	39.80
BanglishBERT	82.65	40.55	42.25	41.30
GPT-4.1-mini labels (CoT)				
mBERT	84.50	43.60	46.10	44.80
BanglaBERT	83.90	42.80	45.30	44.00
BanglishBERT	85.20	44.90	47.30	46.00

Discussions. Token-level analysis in Table 1 confirms the advantage of structured reasoning: inserting an explicit chain of thought more than doubles precision and recall compared with plain ICL for GPT-4.1, while GPT-4.1-mini gains a relative 57% improvement in F_1. Accuracy rises only modestly, indicating that most errors stem from boundary and class misassignments, not gross misalignment of input and output tokens.

The downstream evaluation in Table 2 reveals three salient trends. First, BanglishBERT systematically outperforms BanglaBERT and mBERT irrespective of supervision source, corroborating earlier findings that pre-training on transliterated and code-switched text confers a strong inductive bias for noisy Bengali health forums. Second, CoT-augmented labels elevate all encoder families by 9–13 macro-F_1 points over their ICL counterparts, closing roughly two-thirds of the gap to fully human supervision. Third, the smaller GPT-4.1-mini paired with CoT surpasses the larger GPT-4.1 model operating without reasoning, suggesting that annotation fidelity depends more on prompt structure than on raw model capacity when the task involves fine-grained span extraction.

A Pearson analysis across the 15 result rows shows a near-linear relation ($r = 0.94$) between annotator F_1 and downstream F_1, with a slope of 0.35. Hence, a single-point gain in annotation quality yields a one-third-point improvement in downstream performance, which is a favourable trade-off given the marginal cost of prompt refinements relative to manual labelling.

While human gold supervision remains superior, the best synthetic pipeline attains 85.2% accuracy and 46.0% macro-F_1, only 10.1% and 10.1 absolute points below the fully supervised BanglishBERT baseline. These findings demonstrate that CoT-enhanced large language models constitute a practical,

scalable alternative for low-resource clinical annotation, particularly when expert time is scarce or rapid domain adaptation is required.

Future work will investigate active-learning loops that mix CoT-generated labels with targeted human review, and cross-lingual transfer of prompt designs to related Indo-Aryan languages, aiming to further narrow the residual performance gap identified in this study.

5 Conclusions and Future Works

Conclusions. This study addresses the persistent shortage of annotated clinical text in low-resource languages by designing and rigorously evaluating an automated annotation pipeline centred on large language models. A detailed analysis of the BANGLA-HEALTHNER corpus informed the construction of balanced few-shot prompts and revealed domain characteristics: informality, code-switching, and class imbalance, that challenge conventional NER systems. Building on these insights, two prompting strategies were examined: vanilla in-context learning (ICL) and an enhanced variant that embeds explicit chain-of-thought (CoT) reasoning.

Experimental results demonstrate three key findings. First, introducing CoT reasoning almost triples token-level F_1 relative to plain ICL, underscoring the importance of structured intermediate rationales for fine-grained span annotation. Second, synthetic labels generated by GPT-4.1-mini with CoT elevate downstream BanglishBERT performance to 46.0 macro-F_1, closing more than two-thirds of the gap to fully human supervision without any expert effort. Third, annotation fidelity correlates linearly with downstream NER quality, confirming that even modest gains in label precision and recall translate into meaningful end-task improvements.

Collectively, these findings validate chain-of-thought prompted large language models as a practical and scalable alternative to expert annotation for clinical text in low-resource settings. By coupling corpus-driven prompt design with reasoning-aware generation, the proposed framework offers a strong foundation for rapidly bootstrapping high-quality NER resources where manual annotation is prohibitively expensive.

Future works. To scale the impact of our work, we prioritise four research directions:

(1) Uncertainty-guided human–in-the-loop (HITL). We will incorporate entropy-based routing to surface low-confidence spans for expert correction, targeting precision errors and pushing annotation quality toward the human ceiling while conserving labour.

(2) Cross-lingual validation. The pipeline will be deployed on additional low-resource languages with seed corpora to verify its robustness across scripts, morphology, and code-switching patterns.

(3) Web-scale self-bootstrapping. Leveraging the proven annotation engine, we plan to crawl health forums, Q&A sites, and social media in multiple languages, auto-label the data, and iteratively retrain NER models, creating a virtuous loop that grows both data volume and model accuracy with minimal human oversight.

(4) Annotator and encoder diversification. Future experiments will integrate a broader palette of LLMs (Claude, Gemini, LLaMA, DeepSeek, and many more) and language-specific encoders to quantify the benefits of model diversity and explore ensemble strategies for further performance gains.

(5) Machine-translation-driven corpus expansion. High-quality English clinical corpora will be automatically translated into target low-resource languages using domain-adapted neural machine translation; the translated texts will then be annotated by the proposed CoT-enhanced pipeline and released as parallel NER datasets to accelerate real-world multilingual clinical NLP.

6 Hypothesis

We hypothesize that LLMs, guided by targeted prompt engineering and structured agentic AI processes (like human-in-the-loop feedback for initial annotations), can achieve annotation quality that is comparable to, or even surpasses human-level performance, particularly within the healthcare domain for low-resource languages. By harnessing carefully designed prompts and feedback for initial responses, we expect improved annotation accuracy and consistency. In addition, the ability of

LLMs to generalize from minimal examples is anticipated to support efficient scaling across diverse clinical texts, making this approach broadly applicable in settings with limited annotated resources.

7 Midway Goal

In the midway report, we aim to have the full pipeline ready (prompt engineering, agentic AI workflow), and the data annotated with one LLM (gpt-4o-mini). We also aim to have the pipeline for downstream NER classification completed, and report the performance on one NER model (mBERT). This will enable us to have the main pipeline for one LLM and one NER model, with subsequent comparisons, analyses, and ablation studies in the later part of the project with other models.

References

[1] Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics, 9(1)*, 12.

[2] Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M. (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online.

[3] Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics, 8(3)*, e17984.

[4] OpenAI. (2023). "GPT-4 Technical Report." https://openai.com/research/gpt-4

[5] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J., 2022. AI in health and medicine. Nature medicine, 28(1), pp.31-38.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems, 35*, 24824-24837.

[8] Zhang, R., Li, Y., Ma, Y., Zhou, M., & Zou, L. (2023). LLMaAA: Making Large Language Models as Active Annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 13088–13103).

[9] Choi, J., Kim, Y., Yu, S., Yun, J., & Kim, Y. (2024). UniGen: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1–14).

[10] Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024). Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 303, pp. 1–21).

[11] Choi, J., Yun, J., Jin, K., & Kim, Y. (2024). Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 15–29).

[12] Khan, A., Kamal, F., Nower, N., Ahmed, T., Ahmed, S., & Chowdhury, T. (2023). NERvous About My Health: Constructing a Bengali Medical Named Entity Recognition Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5768–5774).

[13] Brown, T. B. *et al.* Language Models are Few-Shot Learners. *NeurIPS*, 2020.

[14] Zhang, R., Li, Y., Ma, Y., Zhou, M., Zou, L. LLMaAA: Making Large Language Models as Active Annotators. *Findings of EMNLP*, 2023.

[15] Tan, Z. *et al.* Large Language Models for Data Annotation and Synthesis: A Survey. *arXiv:2402.13446*, 2024.

[16] Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*, 2022.

[17] Wang, X. *et al.* Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*, 2023.

[18] Deng, S. *et al.* Chain-of-Thought Prompting for Information Extraction. *ACL*, 2023.

[19] Liu, J. *et al.* MedCoT: Medical Chain of Thought via Hierarchical Expert. *EMNLP*, 2024.

[20] Choi, J. *et al.* UniGen: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation. *EMNLP*, 2024.

[21] Khan, A. *et al.* NERvous About My Health: Constructing a Bengali Medical Named Entity Recognition Dataset. *Findings of EMNLP*, 2023.

[22] Ji, Z. *et al.* A Survey on Hallucination in Large Language Models. *arXiv:2311.05232*, 2023.

[23] Schlegel, V. *et al.* PULSAR at MEDIQA-Sum 2023: Large Language Models Augmented by Synthetic Dialogue Convert Patient Dialogues to Medical Records. *CLEF 2023*, 2023.

[24] Xu, Y. *et al.* TransEHR: Self-Supervised Transformer for Clinical Time Series Data. *ML4H at NeurIPS*, 2023.

[25] Piyawat, C. *et al.* ThaiMedNER: A Corpus for Thai Medical Named Entity Recognition. *LREC–COLING*, 2024.

[26] Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B. AraMedTweets: A Large-Scale Arabic Medical Twitter Corpus for Public Health. *LREC*, 2022.

[27] Catelli, R. *et al.* Cross-lingual Named Entity Recognition for Clinical De-identification Applied to a COVID-19 Italian Data Set. *Applied Soft Computing*, 97:106779, 2020.