Night Shift Nerds at MEDIQA-SYNUR 2026: Pushing Small Large Language Model Capability for Clinical Observation Extraction and Normalization from Nurse Dictation using RLVR

Bayu Aryoyudanta[1], Maria Yuliana[1], Mikie Rachman[2], I Made Agus Setiawan[1]
[1]University of Pittsburgh, [2]Washington University in St. Louis
{baa181, mgy9, made.agus}@pitt.edu, rachmanmikie@wustl.edu

Abstract

We presented a small decoder-only language model for clinical observation extraction and normalization from nurse dictation developed using Reinforcement Learning with Verifiable Rewards (RLVR). We fine-tune Qwen3-1.7B model using a two-stage pipeline: (1) supervised fine-tuning (SFT) with an augmented chain-of-thought (CoT) dataset generated by a teacher model to mitigate RL cold-start, followed by (2) GRPO-based RLVR with multi-component reward functions that verify output format, concept presence, value type, and value correctness using the shared-task ontology (193 concepts) as a verifier. On the development set, SFT+GRPO substantially outperforms GRPO-only **(F1 0.803 vs. 0.620)**. After the test holdout was released, our final system achieved **0.700** precision, **0.785** recall, and **0.740** F1. Error analysis shows remaining challenges in concept over-detection and missed concepts, as well as boundary errors in categorical and multi-select value types extraction. Our results demonstrates that small language models can enable accurate, cost-effective, and privacy preserving automated clinical documentation for nurse dictation, supporting scalable deployment in low-resource healthcare settings to reduce nurses' documentation burden.

Keywords: Clinical Extraction Task, LLM, SFT, RLVR

1. Task Overview

Clinical documentation refers to the process of recording patient's conditions and communicating actions, decision and clinical reasoning among members of the care team. Its primary aim is to provide concise, history-rich notes that summarize collected information and support clinical impression, diagnosis, treatment and recommended follow-up (Kuhn et al., 2015). Despite the advantages of clinical documentation, the overwhelming workload associated with creating clinical documentations during patient encounters creates a significant burden among clinicians especially nurses where they spent nearly 30% of their time with EHR documentation which leads to less time and attention to patient interaction (Sasseville et al., 2025).

In response, the rapid advancement of Artificial Intelligence (AI) technologies in modern healthcare has positioned AI medical scribe technology as a promising approach for improving clinical documentation workflow and reducing clinician burden. AI medical scribes listen to clinician-patient interactions and generate not only verbatim transcripts but also structured clinical documentation mapped to existing clinical concepts (Mess et al., 2025). While transcription quality from AI medical scribe system has achieved a strong performance, the extraction and normalization of the clinical observation from these transcripts remain a key limitation.

To address this challenge, the MEDIQA-SYNUR shared task focuses on the extraction and normalization of clinical observation from patient and nurse conversational transcripts and mapping them to a large ontology of clinical concepts. The goal is to reduce nursing burden and improve the accuracy and completeness of patient records by automatically capturing clinically salient information from nurse-patient conversations. MEDIQA-SYNUR shared task is part of the 2026 Workshop on Clinical Natural Language Processing (Clinical NLP).

Given that many healthcare settings, including rural hospitals and organizations in low-resource countries, have limited computational budget and infrastructure, there is a need for AI tools that can work within these constraints. This project examines whether the reinforcement learning technique, especially reinforcement learning with verifiable reward (RLVR) (Lambert et al., 2025) can improve the performance of small large language models (LLM), fewer than 2B parameters, for extracting and normalizing clinical observations from patient-nurse conversation transcripts. By focusing on small models, we aim to make this approach more accessible and applicable to a wider range of clinical settings.[1]

2. Related Work

Extracting clinical observations from biomedical texts with natural language processing (NLP) approach has been a long-standing area of research.

[1]Full disclosure: our team includes organizers from this shared task to our code-repository. All codes for data pre-processing, augmentation, fine-tuning, inference, post-processing, and evaluation are available at: `https://github.com/yudanta/nsn-mediqa-synur`

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 163–173
16 May 2026.

Clinical named entity recognition (Clinical NER), typically formulated as a sequence labeling task has been widely studied for extracting clinical information using rule-based systems, recurrent neural networks and transformer-based models trained for NER downstream task(Neumann et al., 2019; Lee et al., 2020). While sequence labeling approaches are effective at identifying biomedical entities, it does have several limitations particularly with respect to entity type rigidity and value normalization. These models are usually trained on a fixed set of entity types, and normalization to standardized clinical concepts often requires additional post-processing modules which beyond sequence labeling architectures capability.

With the emergence of LLM especially the generative pre-trained model, LLM have been increasingly applied to clinical NER task through prompt engineering, often achieving improved performance over the traditional sequence labelling approach(Hu et al., 2024). In addition, LLM can be fine-tuned for Clinical NER task with instruction type following dataset. For instance, Liu et al 2025 (Liu et al., 2025b), fine-tuned an LLM (Llama 3.1 8B) with supervised fine-tuning (SFT) technique with structured schema (i.e. JSON output format) and achieve 90% exact match accuracy witch less annotated data compared to the traditional machine learning approach. Although, SFT-based LLM demonstrated improved schema aware generation and better contextual understanding compared to sequence modelling, SFT primarily teach models to generate outputs given input and output from its training tokens. As a result, SFT-trained model often struggle with unseen cases or scenarios requiring logical verification, which are common in clinical concept extraction task. For instance, the clinical entity might explicitly mentioned but negated or it may need to be inferred implicitly from the surrounding context. These limitations constrain the generalizability of SFT-based LLM for clinical entity extraction and normalization task.

Reinforcement learning (RL) has recently emerged as a powerful paradigm for improving the capability of LLM. For tasks with with verifiable outcomes, (Guo et al., 2025), introduced Group Relative Policy Optimization (GRPO) an RL method that eliminates the need for a critic network by estimating advantages from grouped response comparisons from verifiable reward functions. In addition, their work also demonstrates that RL can substantially improve model reasoning capabilities. Building on this foundation, Tulu 3 (Lambert et al., 2025) introduced Reinforcement Learning with Verifiable Rewards (RLVR), a multi-stage post-training framework that combines SFT with subsequent RLVR training pipeline, achieving state of the art results and empirically outperforming SFT only approaches. In the biomedical domain, Fleming-R1(Liu et al., 2025a) applies RLVR with GRPO for medical reasoning tasks with three complementary innovations: 1) reasoning-oriented data curation, 2) Chain-of-Thought (CoT) cold start training before RL training and 3) two-stage RLVR training focus on different reasoning skills sets data resulting in near-parity performance with GPT-4o model. Building upon these prior works, our projects extend the RLVR framework for clinical observation extraction and normalization task from patient-nurse conversational transcripts.

Field	Content
ID	14
Transcript	*[Clinician]* Alright, let's see here...uh, the patient's high risk for falls. They have a, uh, history of falls and, uh, joint deformity, so, uh, we're providing, um, moderate assistance with mobility...
Observations	Cognitive status: Alert with general confusion; . . . ; Patient safety: Yes

Table 1: Example of clinical documentation entry from our dataset.

3. Dataset

The 2026 MEDIQA-SYNUR shared-task dataset `https://sites.google.com/view/mediqa2026/mediqa-synur` aimed to extract medical observation from nurse dictation data. This dataset consists of patient-nurse dictation and normalized extracted clinical information mapped into predefined clinical schema. The patient-nurse dictations were generated through a controlled multi-agent simulation pipeline, and the gold-standard annotations were produced by expert nurses using an open-source, large scale clinical schema (193 concepts) (Corbeil et al., 2025). The dataset is partitioned into training, development and test subsets. In this manuscript, we mainly focus on train and development subsets of MEDIQA-SYNUR dataset. Examples of nurse-patient transcript and its extracted clinical information and clinical concept map can bee seen from Table 1 and 2 while dataset statistics and its clinical observation distribution can be found in Table 3.

4. Methods

We extend the RLVR framework to the task of extracting and normalizing clinical observations from patient - nurse conversational transcripts and mapping them to a large ontology of clinical concepts.

Field	Value
ID	3
Name	Respiratory interventions
Value Type	MULTI_SELECT
Value Enum	• raising head of bed • turn cough • deep breathe • incentive spirometer usage

Table 2: Example clinical concept map for respiratory interventions.

Statistic	Train	Dev	Test
Total examples	122	101	199
Avg. obs. per example	13.8	13.0	12.8
Min. obs. per example	6	7	4
Max. obs. per example	34	33	39
Single-select obs.	1200	951	1853
Multi-select obs.	195	157	344
String obs.	108	70	257
Numeric obs.	182	137	98

Table 3: Summary statistics of the MEDIQA-SYNUR dataset across training, development, and test splits.

Our training pipeline employs a cold start CoT fine-tuning stage followed by RL with GRPO as illustrated in Figure 1. To construct the CoT dataset, we leverage a larger open-source model to generate reasoning traces that aid model's exploration capability by learning concept identification and value extraction through CoT dataset before GRPO training.

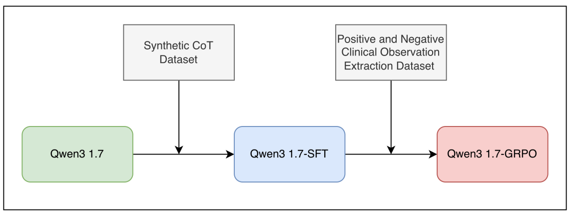

Figure 1: Current Approach Training Pipeline.

4.1. Dataset

The raw dataset from the 2026 MEDIQA-SYNUR shared task was processed to align with the requirements for SFT and RLVR model fine tuning. Each A clinical concept will be paired with a clinical transcript as a context with an output of normalized extracted value. These pair of dataset will be sampled and used as a seed for generating CoT dataset for the SFT fine-tuning process.

4.1.1. Dataset pre-processing and augmentation

The train datasets from this shared-task underwent several pre-processing steps, focusing on dataset formating and negative example augmentation. We created two classes of data: positive and negative examples. The positive dataset consist of consist of nurse dictation transcript paired with clinical concept and its expected extracted value. The Negative dataset consist of nurse dictation transcript paired with clinical concepts that do not appear in the transcript with a null extracted value.

During negative dataset generation, we also generate negative pairs of similar or potentially confusing clinical concepts to model distinguish between related concepts that are not present in the transcript. To support this process, we first created a mapping of clinical concepts and their related or commonly confused concepts as shown in Table 4. Examples of positive and negative samples are shown in Table 5 and Table 6. Summary statistics of the training dataset are presented in Table 7

Concept	Similar / related concepts
Blood pressure	Mean arterial pressure (MAP)
Mean arterial pressure (MAP)	Blood pressure

Table 4: Example of similar clinical concept for blood pressure.

4.1.2. Synthetic dataset augmentation for unseen concepts

During our exploratory data analysis, we found out that the train dataset examples doesn't cover all clinical concepts from the provided ontology. Those unseen clinical concepts from train dataset are: **'Bladder scan volume unit', 'Emesis volume unit', 'Hearing aid', 'Assistance with toileting', 'Seizure activity', 'Excoriated tissue', 'Voice quality', 'Prosthetic use', 'Food allergies', 'Foreign object removal', 'Pupil size', 'Speech content', 'Stool consistency', 'Caloric intake unit', 'Temperature source', 'Tracheostomy status', 'Vaginal discharge', 'Ear examination', 'Cranial nerve function', 'Voiding function', 'Body art', 'Tube feeding type', 'Patient belongings', 'Broset violence checklist - physically threatening', 'Chest expansion', 'Positioning frequency', and 'Motor reflexes'**. To overcome this unseen concept, we generated synthetic datasets to fill the unseen concept gaps in our dataset. During the

Field	Content
Transcript	[Clinician] Patient presents with moderate dyspnea. Uh, they're using accessory muscles—um, you can see the increased effort with their breathing. We've been, uh, managing this with a few respiratory interventions...
Concept	Dyspnea
Value Type	SINGLE_SELECT
Value Enum	['None', 'Mild', 'Moderate', 'Severe']
Value	Moderate

Table 5: Example of positive train dataset.

Field	Content
Transcript	[Clinician] Patient presents with moderate dyspnea. Uh, they're using accessory muscles—um, you can see the increased effort with their breathing. We've been, uh, managing this with a few respiratory interventions...
Concept	Emesis color
Value Type	SINGLE_SELECT
Value Enum	['dark green', 'green']
Value	None

Table 6: Example of negative train dataset.

process, we utilize OpenAI GPT-4o to generate this synthetic examples.

4.1.3. CoT data augmentation

In addition to positive and negative examples for model fine-tuning, we also generated a CoT dataset to support the SFT training process. The CoT dataset examples were generated using OpenAI GPT OSS 120B as teacher models. As seed for generation, we sampled the pre-processed dataset decribed in Section 4.1.1. The output of this process is a collection of CoT dataset consisting of nurse dictation transcript, clinical concepts and their associated reasoning process as illustrated in Table 8. At the end of CoT data augmentation process, we selected 3066 examples for SFT training as summarized in Table 9

4.2. SFT with CoT

In this study, to address the cold start problem and improve overall model performance as described in RLVR training pipeline(Lambert et al., 2025), we first fine-tuned the Qwen3 1.7B base model with the CoT dataset generated in Section 4.1.3. This CoT-initialized model is expected to enhance base model exploration capability during the reinforcement learning process, thereby mitigating the cold start issue and improving overall task performance.

Category	Count
Total clinical concepts covered	193
Negative samples	5818
Positive samples	3036
SINGLE_SELECT	6109
STRING	1147
NUMERIC	939
MULTI_SELECT	659

Table 7: Summary of the training dataset used for model development.

4.3. RLVR Fine-Tuning

During GRPO training, instead of using answer correctness as a single verification mechanism, we employ multi-component reward functions that verify several aspects of the model output:

- **Response format**: verifies that the response contain both reasoning and answer section enclosed within <think></think> and <answer></answer> tags.
- **Presence of clinical concept**: verifies whether the target clinical concept is present in the patient-nurse transcript or not.
- **Correctness of clinical concept's value type**: verifies that the extracted value type matches the concept value type defined in the clinical concept ontology.
- **Correctness of extracted clinical value**: verifies that the extracted value matches the gold standard annotation values.

In this experiment we employed non-binary reward values, which typically assigning **+1** for correct response and **(0 or -1)** for incorrect responses, to provide more granular reward control and penalties based on model behavior. The combination of these reward functions and non-binary reward values is intended to address the unique challenges of this clinical observation, extraction, normalization and ontology mapping where errors can occur in multiple aspects.

In our experiment, we iteratively adjusted reward and penalty values for clinical concept detection based on error analysis to balance false positive and false negative clinical concept detection. Because our developed model tended to produce more false positive observations, we set higher

Field	Content
Transcript ID	160
Concept	Heart rate
Value type	NUMERIC
Label	Positive
Ground-truth value	88
Transcript	[Clinician] Patient is, uh, currently in a pos...
CoT reasoning	The transcript contains the exact phrase **“Heart rate is being monitored at 88 beats per minute”** (line... “Heart rate is being monitored at 88 beats per minute, with heart sounds...”). This wording directly mentions the clinical concept “Heart rate,” satisfying the requirement that the concept be explicitly named in the text. Because the phrase supplies a numeric measurement (“88”) together with the unit “beats per minute,” it unambiguously provides the quantitative value for the heart-rate concept, and no other heart-rate numbers appear elsewhere in the note. Consequently, the correct extracted value for the “Heart rate” concept is **88**, matching the ground-truth specification. Given the clear, verbatim statement and the absence of any contradictory data, we are highly confident (near-certain) that the heart-rate value is 88 bpm.

Table 8: Example instance from the augmented chain-of-thought (CoT) dataset used for supervised fine-tuning.

Category	Count
Total clinical concepts covered	193
Negative samples	2042
Positive samples	1024
SINGLE_SELECT	2091
STRING	379
NUMERIC	338
MULTI_SELECT	258

Table 9: Summary of the augmented chain-of-thought (CoT) dataset used for supervised fine-tuning.

penalty to false positive cases compared to false negative cases. Using non-binary reward values provided greater flexibility in steering model behavior during RL training.

In this study, we performed a full-parameter model fine-tuning for both SFT and RLVR training pipeline by considering that our experiment is limited to small LLM with the base models constrained to maximum of 2 billion of model parameters. Both SFT and RLVR training process were implemented using Huggingface's TRL library.

4.4. Model Evaluation and Performance Metrics

Following the evaluation from the 2026 MEDIQA-SYNUR shared task in clinical extraction, we will employ **Precision**, **Recall** and **F1** evaluation metrics to assesses the performance of the proposed approach. The evaluation script can be access through this link: `https://github.com/abachaa/MEDIQA-SYNUR-2026/blob/main/mediqa_synur_eval_script.py`

5. Results

Table 10 presents the evaluation metrics across training configurations. Our model gained substantial advantages from the SFT training with the augmented CoT dataset compared to the model trained solely using GRPO-RL, which align with training recipe suggested in prior work from(Lambert et al., 2025; Guo et al., 2025). Furthermore, the accumulated reward for the model trained on SFT + GRPO-RL was higher than trained directly using GRPO-RL. As shown in Figure 2 the model initialized with SFT starts with higher reward values, helping overcome the cold-start problem commonly observed in RLVR model training.

By the time the test holdout was published, our fine-tuned model achieved a **precision of 0.700**, **recall of 0.785** and **F1-scores of 0.740**.

6. Error Analysis

To better understand the strengths and limitations of our developed system, we conducted an error analysis on the **development (dev) dataset** outputs from our best-performing model. Overall, the mean number of clinical observations per transcript in the development dataset is 13.019, while the model predicts an average of 14.316 observations,

Method	Dataset	Precision	Recall	F1-score
GRPO-only	Train	0.496	0.828	0.621
	Dev	0.494	0.830	0.620
SFT+GRPO (Best Model)	Train	0.772	0.832	0.801
	Dev	0.770	0.838	0.803

Table 10: Models (Qwen3-1.7B) performance comparison.

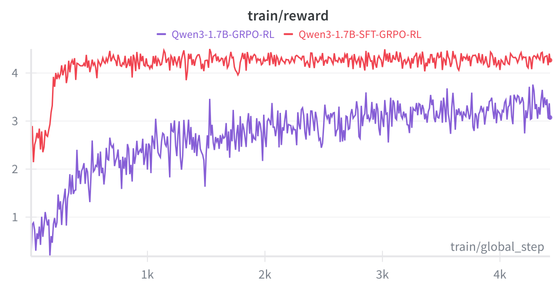

Figure 2: Overall reward function between GRPO-RL only and SFT-Followed by GRPO-RL Training.

indicating that our model tends to extract slightly more concepts than present in the reference annotations. On average, 2.772 extracted observations per transcript correspond to false positives. While, our model correctly extracts a mean of 10.762 observations. Among these correctly identified observations, the mean number of value extraction errors is 0.782, suggesting that our fine-tuned model is generally effective at extracting and normalizing clinical observation values although some errors remain.

6.0.1. Medical Concept Detection

Further error analysis on medical concept detection shows that our fine-tuned model still struggles with both over-detection (false positives) and missed observation (false negatives). Across development dataset, our model produced a total of 280 false-positive clinical observations, indicating an over-extraction tendency, where clinical concepts are predicted even when they are not present in the transcript. These false-positive errors frequently occurred in symptoms or assessment related concepts such as: gastrointestinal symptoms, fall risk identification, cognitive status as well as unit-based concepts such as oxygen saturation unit, heart rate unit, and bladder scan volume unit. This patterns suggest that our model sometimes infer plausible observations from transcript but the real observation is not exist. Examples for both false-positive and false-negative observation can be seen from Table 11.

In the other hand, our fine-tuned model also produced a total of 66 false-negative observations, where clinical concepts actually present in the transcript but not detected by our model. The most commonly missed clinical concepts including secondary analysis, patient safety, mobility, orientation and patient identification. These false-negative cases indicate that our model still lacks robustness in detecting clinical concepts, even when they are expressed explicitly in the transcript. Taken together, this medical concept detection analysis suggests that while our model demonstrates strong overall clinical concept detection, it still shows a tendency toward both over and under detection, which warrant further improvements.

6.0.2. Extracted Concept Values

Error analysis on extracted values shows that among correctly extracted concepts, 97 instances contain incorrect values, with most errors occurring in SINGLE_SELECT (40) and STRING (19), followed by MULTI_SELECT (16) and a smaller number of NUMERIC errors (4). For MULTI_SELECT clinical concepts, our model often over or under select options, for instance predicting "dark" and "cloudy" observation where only "cloudy" is annotated. For SINGLE_SELECT concept, our model show grading or label-boundary confusion such in predicting "moderate" instead of "severe" in Dyspnea concept. For STRING type value, model often extract more phrase compared to the annotated value for instance our model extract "2+ bilaterally" instead of "2+" for Peripheral pulse concept. For NUMERIC type value, error occurs in numeric values that is stated in a compound form, for example is blood pressure observation which written as 160/95mmHg. Overall these pattern suggest that model still need improvement primarily for MULTI_SELECT observation and capability in categorical boundary decision.

7. Conclusion

We presented a fine-tuned decoder only LLM for clinical observation extraction based on Qwen3-1.7B trained on the MEDIQA-SYNUR shared-task dataset using an SFT followed by GRPO-RL pipeline. In addition to the shared-task dataset, we incorporated an augmented CoT dataset during SFT, which improved model performance after GRPO-RL training.

Our experimental results suggest that clinical observation detection, extraction, and normalization given a predefined clinical concepts ontology can be formulated as a verifiable problem, making RLVR suitable for this task. The clinical concept ontology can serve as a verifier for the RLVR reward functions, enabling the model to learn reasoning behaviors that improve overall task performance, even for relatively small LLM such as Qwen3-1.7B. Therefore, this resource-efficient model is not only accessible, but also cost-effective and privacy-preserving, making it suitable for low-resource setting and scalable to broader clinical settings to help reduce nurse burden in clinical documentation.

8. Future Work

Error analysis highlights that our developed model still suffer from false positive and false negative observations, as well as incorrect value extraction for correctly identified clinical concept from patient-nurse dictation data. These limitation warrant future study to improve clinical concept detection, value extraction and normalization from patient-nurse dictation data.

In addition to improving RLVR reward function design with ontology-aware and concept-level penalties to better balance both false-positive and false-negative detection behaviors, new approaches such as interactive agentic AI system that support clarification for ambiguous clinical concepts are wort exploring.

9. Ethical Consideration

Failures in LLM-based clinical extraction system could exacerbate existing health disparities and increase the risk of patient safety. Given that the LLM may exhibit demographic biases and lower accuracy in non-English language, a systematic extraction errors could affect certain populations. In addition, missing critical observations like confusion or pain assessment for patients could create or further increase clinical documentation disparities that may affects to individuals healthcare outcomes.

For user data privacy, the 2026 MEDIQA-SYNUR shared task datasets were generated through multi-agent interaction in closed environment without using any real person dataset. Therefore, issues that might arise related to personal health information (PHI) can be minimized.

10. Acknowledgment

We gratefully acknowledge Health and Rehabilitation Informatics (HARI) Lab `hari.pitt.edu` for providing computational resources and infrastructure support for the model fine-tuning experiments.

11. Bibliographical References

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering Healthcare Practitioners with Language Models: Structuring Speech Transcripts in Two Real-World Clinical Applications. ArXiv:2507.05517.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan,

Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Thomson Kuhn, Peter Basch, Michael Barr, Thomas Yackel, and for the Medical Informatics Committee of the American College of Physicians*. 2015. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper From the American College of Physicians. *Annals of Internal Medicine*, 162(4):301–303.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. ArXiv:2411.15124.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. 2025a. Fleming-R1: Toward Expert-Level Medical Reasoning via Reinforcement Learning. ArXiv:2509.15279 version: 1.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng Lai, Thomas Gracie, Ashwin Kishtagari, Peter R. Carroll, Alexander G. Bick, Anobel Y. Odisho, Maggie Chung, and Adam Yala. 2025b. Human level information extraction from clinical reports with finetuned language models. *Scientific Reports*, 15(1):45239.

Sarah A. Mess, Alison J. Mackey, and David E. Yarowsky. 2025. Artificial Intelligence Scribe and Large Language Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations. *Plastic and Reconstructive Surgery - Global Open*, 13(1):e6450.

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, Florence, Italy. Association for Computational Linguistics.

Maxime Sasseville, Farzaneh Yousefi, Steven Ouellet, Florian Naye, Théo Stefan, Valérie Carnovale, Frédéric Bergeron, Linda Ling, Bobby Gheorghiu, Simon Hagens, Samuel Gareau-Lajoie, and Annie LeBlanc. 2025. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare*, 13(12):1447.

12. Appendices

Field	Content
False Positive Example	
ID	87
Transcript	[Clinician] Alright, let me gather my notes here for the patient assessment. So, um, we have a patient who's showing, uh, let's see, a mix of urinary and gastrointestinal symptoms. The patient reported, uh, vomiting, and, uh, the emesis was, uh, recorded at about 500 cc. It's that dark green color, which could suggest, um, some gastrointestinal distress or maybe something going on there, like a blockage. [Clinician] On the urinary side, um, the patient is experiencing difficulty urinating and, uh, has this urgency. So, yeah, it might be, um, a sign of some bladder irritation or maybe obstruction, possible stone there. Uh, mobility-wise, the patient is slightly limited, um, but continent, which is good. Uh, this limited mobility could be due to discomfort or pain from these issues. [Clinician] Now, um, checking the oral mucosa, it's dry, which could hint at dehydration, especially with the vomiting and possibly not taking enough fluids. Uh, vital signs are stable-ish, I'd say, with an oxygen saturation at 92% and a heart rate of 85 bpm. So, yeah, this paints a picture of acute distress from these gastrointestinal and urinary issues, not critical but definitely needs attention and maybe, um, further diagnostics to pinpoint the exact cause.
Gold Concept	None (Gastrointestinal symptoms not annotated)
Predicted Concept	Gastrointestinal symptoms
Explanation	The model incorrectly inferred the higher-level concept "Gastrointestinal symptoms" from the mention of vomiting. However, vomiting is annotated under its own concept in the ontology, leading to a false-positive concept detection.
False Negative Example	
ID	10
Transcript	[Clinician] Alright, so, uh, here's what we've got. This patient, elderly, has been, um, experiencing a few issues over the last 24 hours. They're currently, uh, bedridden due to a recent surgery, which limits their, uh, mobility quite a bit. Now, uh, during the assessment, some things stood out. There's evidence of a, uh, constipation episode, which we can see with the perineal edema present. It's, um, making the patient quite uncomfortable. [Clinician] Now, the patient is also dealing with, uh, persistent nausea. They are retching but, uh, haven't had any actual vomiting, so it's more of a gastrointestinal distress without the vomiting. Skin assessment shows, uh, tented skin turgor, indicating possible dehydration. This might, uh, be linked to the constipation they're experiencing.[Clinician] Cardiovascular-wise, the heart rate's a bit high, sitting at around 100 bpm. This could be, um, partially due to dehydration and maybe some pain. The patient, uh, reports pain at about 6 out of 10 and is also having, um, painful swallowing, which complicates, uh, oral intake. Clinician] Overall, it looks like a multifactorial approach is needed here. We'll need to, uh, address hydration management, constipation relief, and, uh, control the pain effectively.
Gold Concept	Mobility
Predicted Concept	Not detected
Explanation	Although the transcript explicitly describes reduced mobility ("bedridden" and "limits mobility"), the model failed to detect the Mobility concept, resulting in a false-negative observation.

Table 11: Examples of model errors from the development dataset. The first example illustrates a false-positive concept detection caused by ontology-level confusion, while the second shows a false-negative detection despite explicit mention in the transcript.

Prompt Template for SFT Training

```
Extract medical concept: {medical_concept_name}
Value Type: {value_type}
Allowed values (choose ONE): {value_enums}

Transcript: {patient_nurse_dictation_transcript}

Select exactly one value from the allowed list.

Expected output format:
<think>
reasoning process
</think>

<answer>
{"concept_present": true/false,
 "concept": "concept name",
 "value": "value" or null if concept_present is false}
</answer>

{completions}
```

Figure 3: Prompt template for SFT Training.

GRPO Fine-Tuning Example for Clinical Concept Extraction

System Prompt
You are a medical concept extraction assistant analyzing conversations between patients and clinician or nurses. Your task is to identify and extract key medical concepts including symptoms, conditions, medications, treatments, procedures, and relevant clinical observations.

1. First, analyze the conversation in the <think> section, identifying:

- Medical terminology and clinical concepts
- Patient-reported symptoms and concerns
- Clinician or nurse assessments and recommendations
- Temporal relationships and context
- Whether the target concept is present, absent, or ambiguous
- For the target concept: locate the specific value or measurement
- For unit concepts (e.g., 'Heart rate unit'): verify the base measurement is present before extracting the unit

2. Then, provide structured extracted concepts in the <answer> section.

Output format:
<think>
reasoning process
</think>

<answer>
{"concept_present": true/false,
"concept": "concept name",
"value": "value" or null if concept_present is false}
</answer>

IMPORTANT: Set concept_present to true ONLY if a specific value for the concept can be identified in the transcript. Never extract a unit concept (e.g., 'Temperature unit', 'Heart rate unit') unless the corresponding measurement itself is documented.

User Prompt
Extract medical concept: {medical_concept_name}
Value Type: {value_type}
Allowed values (choose ONE): {value_enums}
Transcript: {patient_nurse_dictation_transcript}
Select exactly one value from the allowed list.

Figure 4: Example instruction pair used during GRPO fine-tuning for clinical concept extraction from nurse dictation transcripts.