AnotherOne at MEDIQA-SYNUR 2026: Detect, Extract, Normalize - Knowledge-Grounded LLM Pipeline for Clinical Observation Extraction

Jerrin John Thomas, Parameswari Krishnamurthy
International Institute of Information Technology, Hyderabad
jerrin.thomas@research.iiit.ac.in, param.krishna@iiit.ac.in

Abstract

We present a system for the MEDIQA-SYNUR 2026 shared task on extracting structured clinical observations from nurse dictation transcripts. The transcripts contain spoken-style clinical language with disfluencies, filler words, and hesitations. Our approach is a four-stage LLM inference pipeline preceded by an offline knowledge enhancement step: (1) knowledge-enhanced concept detection using medical domain clustering, (2) evidence-grounded value extraction, (3) schema-constrained value normalization, and (4) deterministic post-processing with fuzzy matching and unit pairing. In the offline step, we use the task ontology and training examples to generate per-concept clinical definitions and extraction rules, and group the 193 concepts into 19 non-exclusive medical domain clusters. These are injected into all downstream prompts as domain priors. All LLM stages use gpt-oss-120b with structured JSON output and chain-of-thought reasoning. The task requires exact matching on concept ID and value pairs across a 193-concept ontology, making precision particularly challenging. We iteratively refine concept definitions and prompt guidelines based on error analysis of the training data. Our system achieves an F1 score of 0.806 on the test set.

Keywords: clinical NLP, information extraction, nursing documentation, large language models, ontology-constrained extraction

1. Introduction

Nurses spend a disproportionate amount of time on electronic health record (EHR) documentation rather than direct patient care (Tierney et al., 2024). Automating the conversion of spoken nursing assessments into structured flowsheet entries can substantially reduce this documentation burden (Quiroz et al., 2019).

The MEDIQA-SYNUR 2026 shared task (Ben Abacha et al., 2026) addresses this problem directly: given a raw nurse dictation transcript containing filler words, hesitations, and disfluencies, systems must extract all clinical observations, map each to one of 193 predefined concepts in a structured ontology, and normalize values to the correct type. The task builds on the SYNUR dataset introduced by Corbeil et al. (2025), who demonstrated that LLMs can structure clinical speech transcripts but that the task remains challenging due to the strict exact-match evaluation criterion.

Several factors make this task difficult: (a) noisy spoken-style input with disfluencies and filler words; (b) a 193-concept ontology with strict exact-match evaluation on concept ID and value pairs; (c) limited training data (122 samples); (d) high annotation density, with approximately 14 observations per transcript; (e) negation handling where "no edema" must map to Edema: No edema; and (f) linking each numeric measurement to its corresponding unit concept.

Rather than a single end-to-end LLM call, we decompose the task into specialized stages - detect, extract, normalize, validate, each operating within constrained output spaces. We cluster concepts by medical domain, since transcripts documenting one concept in a clinical area typically document others from the same area (e.g., a respiratory assessment typically covers breath sounds, oxygen saturation, and cough description together).

Our contributions are: (1) a decomposed pipeline separating concept detection, value extraction, and normalization into distinct stages; (2) medical-domain concept clustering that exploits co-occurrence patterns in clinical assessments; (3) an offline knowledge enhancement step generating per-concept definitions and extraction rules as LLM priors; and (4) an iterative refinement process using training-set error analysis and model uncertainty signals to improve concept definitions, extraction rules, and prompt guidelines.

2. Related Work

Clinical NLP and Information Extraction. Early work in clinical information extraction includes the i2b2 challenges on concept, assertion, and relation extraction from clinical text (Uzuner et al., 2011). More recently, LLM-based extraction works have shown that providing concept definitions in prompts improves biomedical NER by 0.15 F1 score points (Hu et al., 2024b), and that prompt engineering alone can make LLMs competitive with fine-tuned models for clinical NER (Hu et al., 2024a).

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 251–256
16 May 2026.

MEDIQA Shared Tasks. The MEDIQA series has addressed clinical question answering (Ben Abacha et al., 2019), dialogue summarization (Ben Abacha et al., 2023), medical error detection (Ben Abacha et al., 2024), and medical order extraction (Ben Abacha et al., 2025). MEDIQA-SYNUR extends this to nursing observation extraction, building on the SYNUR dataset introduced alongside the closely related SIMORD dataset for medical order extraction (Corbeil et al., 2025).

Processing Clinical Dictation. Nurse dictation transcripts contain disfluencies, filler words, and informal language that pose challenges distinct from polished clinical notes. Corbeil et al. (2025) evaluated both open- and closed-weight LLMs on the SYNUR task, finding that larger models handle spoken-style clinical text more reliably. The documentation burden problem (Tierney et al., 2024; Quiroz et al., 2019) has driven work on structuring clinical speech.

Chain-of-Thought Reasoning. Wei et al. (2022) demonstrated that chain-of-thought prompting improves complex reasoning in LLMs. We adopt this by requiring explicit reasoning fields in every LLM output, forcing the model to articulate extraction logic before producing structured answers.

Negation in Clinical Text. Negation detection is critical in clinical IE. For example, "no facial droop" must map to Facial droop: no facial droop. The ConText algorithm (Harkema et al., 2009) introduced patterns for negation, experiencer, and temporal status. We incorporate these principles as prompt guidelines rather than rule-based preprocessing.

3. Task Description

The MEDIQA-SYNUR 2026 shared task (Ben Abacha et al., 2026) requires systems to extract structured clinical observations from synthetic nurse dictation transcripts. Given a raw transcript containing spoken-style clinical language with filler words, hesitations, and disfluencies, the system must identify which clinical concepts are documented, extract their values, and normalize the output to match a predefined ontology schema.

Each observation specifies a concept ID, concept name, value type, and value. The ontology defines 193 clinical observation concepts spanning vital signs, respiratory, cardiac, neurological, gastrointestinal, genitourinary, musculoskeletal, skin, safety, and behavioral domains.

Split	Samples	Gold Obs.	Avg. Obs.
Train	122	1,774	14.5
Dev	101	1,390	13.8
Test	199	2,721	13.7

Table 1: Dataset statistics. Avg. Obs. is the mean number of gold observations per transcript.

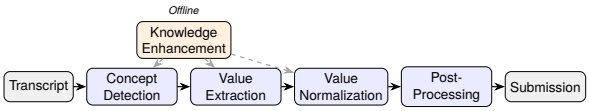

Figure 1: System architecture. Dashed arrows indicate domain knowledge (concept definitions, extraction rules, and cluster assignments) injected into LLM prompts. Post-processing is deterministic.

3.1. Dataset

The MEDIQA-SYNUR 2026 dataset (Ben Abacha et al., 2026; Corbeil et al., 2025) consists of synthetic nurse dictation transcripts in English with gold-standard structured observations. Table 1 summarizes the dataset statistics.

The 193 concepts span four value types: approximately 130 SINGLE_SELECT concepts with enumerated options, 25 STRING concepts requiring free-text values, 20 NUMERIC concepts with associated unit observations, and 15 MULTI_SELECT concepts. An additional 15 unit observation concepts accompany the numeric ones.

Evaluation. Performance is measured by instance-level Precision, Recall, and F1 score with exact matching on concept ID and value pairs. For MULTI_SELECT, values are unrolled into individual ID-value pairs. Numeric predictions are cast to a common numeric type before comparison.

4. System Description

4.1. Overview

Our system is a modular pipeline with four inference stages preceded by an offline knowledge enhancement step. Figure 1 illustrates the architecture. Each stage operates on the output of the previous one, progressively refining the extraction from broad concept detection to schema-compliant normalized observations. All LLM stages output structured JSON with chain-of-thought reasoning.

4.2. Knowledge Enhancement (Offline)

Two offline LLM-assisted steps prepare the domain knowledge injected into all downstream prompts.

Concept Definitions and Extraction Rules. For each of the 193 schema concepts, we prompt the LLM with the ontology and training examples to generate: (a) a clinical definition, and (b) 3-5 general extraction rules. For example:

- **Oxygen saturation** (NUMERIC): *Definition*: "Percentage of hemoglobin binding sites occupied by oxygen, typically measured via pulse oximetry." *Rules*: "Extract only when explicitly documented in the transcript", "Extract the numeric value only; extract the unit as a separate observation", "Do not infer a value from qualitative descriptions such as 'hypoxemic' "
- **Nausea** (SINGLE_SELECT, Yes/No): *Definition*: "Presence or absence of the subjective sensation of an urge to vomit." *Rules*: "Extract the affirmative option when nausea is stated as present", "Extract the negative option only when explicitly denied (e.g., 'no nausea', 'denies nausea')", "Do not assume absence; a negative value requires explicit negation"

Rules focus on when to extract vs. not extract, and are deliberately general. Normalization is handled in a later stage. The prompt instructs the LLM to produce a structured mapping of definitions and rules for each concept.

Medical Domain Clustering. The 193 concepts are grouped into 19 non-exclusive clusters based on medical domain knowledge and co-occurrence patterns in the training data (Table 2). The guiding principle is that a transcript documenting one member of a cluster is likely to document others from the same cluster, reflecting the structure of nursing assessments (e.g., a respiratory assessment typically documents breath sounds, oxygen saturation, cough description, and oxygen delivery device together). Clusters are strictly medical categories, not output-type groups, and their union covers all 193 concepts. Cluster boundaries were determined by grouping concepts that co-occur in the same clinical assessment context (e.g., cardiovascular concepts that appear together in a cardiac exam), then splitting overly large groups to keep per-cluster prompts focused.

The clustering is configurable: both the number of clusters and their composition can be redefined. Our configuration of 19 clusters balances prompt focus with full coverage.

4.3. Stage 1: Concept Detection

The concept detection stage identifies which of the 193 concepts are explicitly mentioned in each transcript. For each transcript-cluster pair, a single LLM call determines which concepts within that cluster are present.

Cluster	Size
Misc. GI & Body	23
Cardiovascular & Hemodynamics	18
Misc. Functional Mobility	16
Measurements & Units	15
Genitourinary	14
Neurologic & Mental Status	14
Misc. Vitals & Respiratory	13
Misc. Neuro & Sensory	12
Respiratory	11
Scores, Scales & Assessments	11
Nutrition & Hydration	11
Devices, Lines & Tubes	10
Gastrointestinal	10
Skin, Wound & Pressure Injury	9
Behavioral & Violence Risk	7
Mobility, Falls & Functional	7
Nursing Care & Environment	7
Admin. & Education	4
Pain	4

Table 2: The 19 medical domain clusters. Size indicates the number of concepts assigned to each cluster. Clusters are non-exclusive: a concept may appear in multiple clusters.

The prompt provides guidelines for explicit-only detection, negation-as-present (e.g., "denies nausea" means the Nausea concept is present with a negative value), and synonym awareness (e.g., "SpO2" = "Oxygen saturation"). For each concept in the cluster, the prompt includes: the concept ID, name, value type, definition, rules, and valid options for SELECT types.

The LLM outputs structured reasoning, detected concepts with evidence quotes and confidence scores, and a list of uncertainties. Responses are aggregated per transcript and deduplicated by concept ID across clusters.

4.4. Stage 2: Value Extraction

A single LLM call per transcript extracts raw values for all concepts detected in Stage 1. Evidence quotes from the detection stage are passed forward to ground the extraction.

The prompt provides type-specific extraction rules:

- NUMERIC: Extract the number only (e.g., "85" not "85%"); the unit is extracted as a separate observation using a predefined mapping from numeric concepts to their unit concepts.
- SINGLE_SELECT: Match one of the valid options.
- MULTI_SELECT: Return a list of matching options.
- STRING: Extract the minimal clinical phrase, removing filler words.

Explicit negation handling guidelines instruct the LLM to interpret "no jugular venous distention" as

JVD: “Absent” and “denies pain” as a negative indicator. As in Stage 1, structured reasoning and evidence are output alongside each extraction.

4.5. Stage 3: Value Normalization

A single LLM call per transcript normalizes all extracted values to match the schema exactly. The prompt provides strict normalization rules:

- SINGLE_SELECT: Output must exactly match one valid option (case-sensitive).
- MULTI_SELECT: Output a JSON array where each element matches a valid option.
- NUMERIC: Output a bare number (integer or float), stripping units.
- STRING: Clean the phrase by removing filler words.

For each concept, the full list of valid options is provided in the prompt. The LLM must also produce separate unit observations for numeric values and explain each normalization decision.

4.6. Stage 4: Post-Processing

A deterministic post-processing stage applies five operations:

Schema Validation. Each observation is checked against the ontology schema. Unknown concept IDs are rejected and value types are enforced.

Fuzzy Matching. For SELECT types, exact matching is attempted first. If no exact match is found, a fuzzy matching score is computed against all valid options, accepting the closest match above a similarity threshold. This recovers near-miss variants produced by the LLM.

Numeric Coercion. Unit suffixes are stripped (e.g., “%”, “bpm”, “mmHg”), blood pressure formats are preserved, and numbers embedded in text are extracted via regular expressions.

Deduplication. Duplicate ID-value pairs are removed.

Unit Pairing. A predefined mapping associates each numeric concept with its unit concept (e.g., Oxygen saturation → Oxygen saturation unit). If a numeric observation is present but its unit is missing, a default unit is inserted (e.g., “%” for O_2 saturation, “bpm” for heart rate), matched against the unit concept’s valid options.

System	Split	P	R	F1
Initial	Train	0.656	0.855	0.743
	Dev	0.641	0.863	0.736
	Test	0.645	0.886	0.747
Refined	Dev	0.761	0.839	0.798
	Test	0.755	0.865	0.806

Table 3: System performance across splits. Initial is the submitted system; Refined incorporates domain knowledge improvements from training-set error analysis (Section 5.3).

	Initial			Refined		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
SINGLE_SEL.	0.716	0.898	0.797	0.815	0.868	0.841
MULTI_SEL.	0.846	0.945	0.893	0.859	0.959	0.906
NUMERIC	0.742	0.984	0.846	0.929	0.965	0.947
STRING	0.020	0.102	0.033	0.020	0.061	0.030
Overall	**0.645**	**0.886**	**0.747**	**0.755**	**0.865**	**0.806**

Table 4: Per-type performance on the test set. Initial is the submitted system; Refined incorporates domain knowledge improvements.

4.7. Infrastructure

All LLM stages use gpt-oss-120b with structured JSON output. Calls are batched across transcripts for efficiency. gpt-oss-120b is a cost-effective model compared to current frontier models.

5. Results and Discussion

5.1. Main Results

Table 3 presents performance across data splits for both the initial submitted system and the refined system (Section 5.3).[1]

The initial system scores are consistent across splits, with F1 ranging from 0.736 on dev to 0.747 on test. The refined system improves F1 to 0.798 on dev and 0.806 on test, with the gains concentrated in precision.

5.2. Analysis by Value Type

Table 4 breaks down test set performance by value type for both systems.

MULTI_SELECT achieves the highest F1 (0.893 initial, 0.906 refined) because the constrained option lists reduce ambiguity. NUMERIC shows near-perfect recall in the initial system (0.984); the refined system trades a small amount of recall for a large precision gain (0.742→0.929), primarily by

[1]The official submission scored F1=0.703 due to a concept ID formatting mismatch (leading zeros). All scores in this paper use normalized IDs.

resolving the confusion between Oxygen saturation and Pulse oximetry concepts. SINGLE_SELECT, the largest category, improves from 0.797 to 0.841.

STRING observations remain the clear weakness, with F1 ~0.03 in both systems. The LLM generates plausible clinical phrases that are semantically correct but lexically different from the gold standard, and exact-match evaluation penalizes these mismatches.

5.3. Knowledge Refinement

Each LLM stage outputs an uncertainty list alongside its predictions, enabling systematic error analysis. We ran the initial system on the training set (F1=0.743) and analyzed both the error patterns and the uncertainty signals from each module. This analysis used only the training set; no development or test data informed the refinements.

The training-set analysis identified five systematic issues: (1) *concept confusion* between Oxygen saturation and Pulse oximetry, where the model extracted the same SpO2 reading under both concept IDs; (2) *over-inference of behavioral concepts*, particularly the Broset Violence Checklist confusion item being triggered by any mention of "confusion" rather than requiring an explicit BVC reference; (3) *STRING hallucination* for concepts like Patient identification and Safety equipment, where the model generated plausible observations not present in the gold annotations; (4) *unit hallucination* where unit observations were extracted without a corresponding numeric value; and (5) *overly restrictive definitions* for Mobility and Activity level, which referenced Braden Scale subscales and missed general functional assessments.

We broadened concept definitions where they were too restrictive, added concept-specific rules to disambiguate confusable concepts, tightened prompt guidelines for STRING types and null value handling, and added gating rules that require explicit tool or scale references for checklist items. The pipeline architecture, clustering, and post-processing remained unchanged.

The refined system improves F1 by 6.2 points on the development set (0.736→0.798) and 5.9 points on the test set (0.747→0.806). The gains are driven by precision (+12 on dev, +11 on test) with modest recall reduction (–2.4 on dev, –2.1 on test). NUMERIC precision improved from 0.742 to 0.929 after resolving the Oxygen saturation / Pulse oximetry confusion.

5.4. Error Analysis

Table 5 shows the top 5 false positive and false negative concepts in the initial system, alongside their counts after refinement.

Cat.	Concept	Initial	Refined
FP	Broset - confusion	102	0
	Patient identification	86	8
	Patient safety	81	35
	Delirium symptoms	80	18
	Pulse oximetry	67	4
FN	Mobility	23	14
	Dyspnea	18	20
	Fall risk ident.	18	21
	Delirium symptoms	17	17
	Temperature unit	15	15

Table 5: Top 5 false positive and false negative concepts from the initial system, with counts after refinement.

False Positives. The refinement reduced false positives across all five worst-offending concepts. Broset violence checklist confusion dropped from 102 to 0 after requiring explicit BVC references. Pulse oximetry dropped from 67 to 4 after disambiguating it from Oxygen saturation. Patient identification dropped from 86 to 8 after restricting extraction to specific identifiers. Patient safety and Delirium symptoms were also reduced, though they remain active sources of error.

False Negatives. False negative patterns were more stable across refinement. Mobility improved from 23 to 14 false negatives after broadening its definition beyond Braden Scale subscales. Dyspnea and Fall risk identification showed slight increases in false negatives (18→20 and 18→21). Delirium symptoms and Temperature unit were unchanged.

5.5. Discussion

The system extracts observations well from constrained value spaces. Refining definitions and rules from training-set error analysis further improves precision without changing the pipeline architecture. The largest gains came from resolving concept confusion between Oxygen saturation and Pulse oximetry, gating over-inferred behavioral concepts, and broadening restrictive definitions for functional assessments.

STRING extraction remains the bottleneck. The generative nature of LLMs produces semantically correct but lexically different text, which exact-match evaluation penalizes. Potential strategies include span selection from the transcript rather than free generation, fuzzy or semantic similarity matching in the evaluation, and few-shot prompting with gold annotation examples to calibrate output style.

5.6. Future Work

We plan to investigate: (a) confidence calibration to improve precision by filtering low-confidence predictions; (b) fine-tuning on the training data; (c) ensemble approaches across models, given headroom for model scaling; and (d) STRING-specific strategies such as span extraction.

6. Conclusion

We presented a modular LLM pipeline for the MEDIQA-SYNUR 2026 shared task that decomposes clinical observation extraction into four specialized stages. Medical domain clustering, knowledge-enhanced prompts, and deterministic validation yield strong results on structured value types. Training-set error analysis guided iterative refinement of definitions and rules, improving precision without sacrificing recall and achieving an F1 of 0.806 on the test set.

7. Bibliographical References

Asma Ben Abacha, Jean-Philippe Corbeil, et al. 2025. Overview of the MEDIQA-OE 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Asma Ben Abacha, Jean-Philippe Corbeil, et al. 2026. Overview of the MEDIQA-SYNUR 2026 shared task on clinical observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*. Forthcoming.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the BioNLP Workshop at ACL 2019*. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Meliha Yetisgen, and Fei Xia. 2024. Overview of the MEDIQA-CORR 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin 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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870, Suzhou, China. Association for Computational Linguistics.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2024a. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1949–1961.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Qianqian Xie, Yujia Zhou, Zhiyong Lu, Hua Xu, and Qiao Jin. 2024b. On-the-fly definition augmentation of LLMs for biomedical NER. *arXiv preprint arXiv:2404.00152*.

Juan C. Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digital Medicine*, 2(1):114.

Aaron A. Tierney, Gavin Gayre, Brian Hoberman, Bruce Mattern, Robert Muller, Marie Schmid, et al. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.