LTRC-Medicom at MEDIQA-SYNUR 2026: Schema-Guided Clinical Information Extraction with Hybrid Clustering-SFT-Verification

Pasumarthy Deepak, Sushvin Marimuthu, Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad, India
pasumarthy.deepak@students.iiit.ac.in,sushvinmarimuthu@gmail.com,param.krishna@iiit.ac.in

Abstract

Extracting structured clinical data from unstructured patient transcripts is challenging due to large target schemas and inherent linguistic ambiguity. We address the extraction of 193 heterogeneous clinical attributes from nursing notes and clinician–patient dialogues, and demonstrate that zero-shot large language models (LLMs) are ineffective in this setting, achieving an F1 score below 0.15 due to context window saturation and hallucination. We propose a four-stage framework that combines semantic schema clustering, role-based chain-of-thought prompting, supervised fine-tuning of Llama-3.1-8B, and transcript-verified post-processing. Our approach achieves an F1 score of 0.66, representing a 4.4x improvement over the baseline, by balancing high recall from generative models with high precision from verification. These results highlight the effectiveness of hybrid pipelines for high-stakes clinical information extraction.

Keywords: clinical information extraction,large language models, schema clustering,supervised fine-tuning,healthcare NLP

1. Introduction

The increasing digitization of healthcare has intensified the demand for efficient clinical documentation, particularly in nursing workflows. Critical patient information is often conveyed during nurse–patient conversations, yet much of it is not systematically captured in structured Electronic Health Records (EHRs).

In mediqa-synur shared task at the Clinical NLP 2026 (George Michalopoulos, 2026), the goal is to advance automated methods for capturing clinically salient information from nursing dialogue data. By improving structured extraction from nurse–patient conversations, the task seeks to reduce nursing documentation burden and enhance the accuracy and completeness of patient records.

Unlike standard Named Entity Recognition (NER) (Xu et al., 2010). which targets generic entities (e.g., “Drug”, “Disease”), this task requires mapping text to a strict schema of 193 constrained fields. Each field enforces structured outputs such as predefined enumerations for “Mobility” (e.g., “Ambulates with assistance” vs. “Bedbound”) or bounded integers for “Pain Score” that trigger downstream protocols. By enabling accurate schema-level extraction from nurse–patient conversations, the task aims to reduce documentation burden and improve the completeness and reliability of patient records.

To address this challenge, we propose a hybrid system that evolves from retrieval-augmented generation (RAG) (Lewis et al., 2020) concepts to supervised learning. By clustering schema fields, we reduce cognitive load and improve task decomposition. Through fine-tuning, we internalize schema constraints within the model. Finally, we incorporate algorithmic verification against the source text to mitigate hallucinations and ensure faithful, schema-compliant extractions.

We systematically explored a progression of modeling strategies with increasing levels of supervision and structural incorporation. A zero-shot baseline performed poorly **(F1 < 15%)**, underscoring the complexity of direct schema-guided extraction. Incorporating schema-aware clustering improved performance to **33% F1**, while systematic prompt engineering further raised it to **45% F1**. Transitioning to supervised fine-tuning (SFT) resulted in a substantial gain, achieving **62% F1**. Finally, augmenting SFT with rule-based post-processing and validation yielded a peak performance of **66% F1**.

2. Related Work

Clinical NER and Information Extraction: Traditional approaches to clinical information extraction have relied heavily on Named Entity Recognition (NER) and relation extraction. Early systems used rule-based methods and classical machine learning (Xu et al., 2010), while modern approaches leverage pretrained transformers such as ClinicalBERT (Alsentzer et al., 2019) and domain-specific models (Yang et al., 2020). However, these methods typically target generic entity types (e.g., medications, diseases) rather than schema-guided extraction with strict constraints.

LLMs for Clinical Tasks: Recent work has explored using large language models for clinical documentation (Agrawal et al., 2022), demonstrating zero-shot and few-shot capabilities. However, these approaches struggle with high-dimensional

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 228–234

schemas and exhibit significant hallucination in clinical contexts, motivating our hybrid verification framework.

Retrieval-Augmented Generation: RAG-based methods (Lewis et al., 2020) have shown promise in grounding LLM outputs, but applying RAG to structured extraction from long clinical dialogues remains underexplored. Our semantic clustering can be viewed as a deterministic routing mechanism that shares conceptual similarities with retrieval-based context selection.

3. Dataset

The shared task utilizes SYNUR (Corbeil et al., 2025a), the first dataset specifically designed for structured observation extraction from nursing dictations. The dataset is accompanied by a schema specification in JSON format that defines the complete set of clinical observation fields. Each schema entry includes a unique identifier (id), a field name (name), and a designated value type (value_type). In total, the schema defines 193 distinct clinical variables.

The schema supports four value types: NUMERIC (e.g., vital signs), STRING (free-text notes), SINGLE_SELECT (categorical status variables), and MULTI_SELECT (symptom or condition lists). For categorical fields (SINGLE_SELECT and MULTI_SELECT), the schema further provides a predefined set of allowable values (value_enum), ensuring standardized normalization of extracted observations.

The training and development splits are provided in JSONL format. Each data instance contains a unique identifier (id), the corresponding nursing dictation (transcript), and a list of annotated observations (observations). Each observation entry includes a schema-aligned identifier (id), field name (name), designated value type (value_type), and the normalized field value (value).

The test split is also released in JSONL format and contains only the unique identifier (id) and the associated nursing dictation (transcript), requiring systems to predict the structured observations according to the predefined schema.

4. System Description

Our pipeline consists of four distinct phases, evolved iteratively to address specific failure modes observed in preliminary experiments.

4.1. Phase 1: Semantic Clustering (Hybrid Lexical-LLM Routing)

Motivation: Large Language Models suffer from "Context Overload" when presented with long lists of instructions, often referred to as the "Lost in the Middle" phenomenon (Liu et al., 2023). Our hypothesis was that by decomposing the 193-item schema into semantically cohesive sub-groups, we could maintain the model's attention mechanism on a smaller, relevant set of definitions.

Implementation: Contrary to traditional K-Means approaches which often create incoherent clusters, we implemented a **Two-Tier Hybrid Router** using `Gemma-2-9B-Instruction` (Team, 2024) as a semantic classifier.

- **Tier 1: High-Precision Lexical Routing:** We defined strict prefix rules for known clinical scales to ensure determinism. For example, any item starting with "Glasgow coma score" is automatically routed to the `Neurological` cluster.
- **Tier 2: LLM-Based Semantic Classification:** For items not caught by lexical rules, we employed a strict classification prompt with `Gemma-2-9B`. The model was tasked with assigning each schema item to one of 4 primary buckets (*Numeric, Single Select, Multi Select, String*) and then further into specific clinical domains (e.g., *Cardiovascular, Respiratory, Safety*).

Cluster Definition:

- *Numeric:* Vitals, Cardiac Measures, Fluid Balance.
- *Single Select:* Respiratory, Neurological (GCS), Skin (Braden).
- *Multi Select:* Symptoms, Interventions.
- *String:* Qualitative Observations, Notes.

Inference Architecture: The pipeline processes the transcript in sequential passes. For each pass, a dynamic system prompt is constructed containing only the definitions for that specific domain (e.g., "Neurological Single Selects"). This reduced the effective instruction token count from $\sim$8000 tokens to $<$1000 tokens per pass, preventing context overflow.

Observation: This phase improved F1 from **0.15 to 0.33**, largely by eliminating cross-domain hallucinations (e.g., confusing "Fall Risk" definitions with "Mobility" definitions).

4.2. Phase 2: Systematic Prompt Engineering (The 9-Step Protocol)

Motivation: General-purpose instruction following is insufficient for clinical rigor. Models tend to be "helpful" rather than "accurate", guessing values to complete the task. We required a methodology

to force strict adherence to clinical documentation standards.

Implementation: We designed a 9-Step Chain-of-Thought (CoT) System Prompt based on the persona of a "CDI Specialist", inspired by the winning approach in the MEDIQA-OE 2025 shared task (Corbeil et al., 2025b).

The Full 9-Step Template: shown in Table 1

Key Innovation: For high-confusion items identified during development, we injected field-specific hints to improve disambiguation. For ID 14 (Safety Equipment), originally defined only as "Equipment for safety", we appended: "Hint: Bed rails, Mats, Alarms ONLY. Do not extract walkers/canes." For ID 40 (Orientation), we added: "Hint: Look for 'Alert and Oriented x3' or 'A&O x4'." These targeted hints reduced ambiguity for the 12 most frequently confused fields.

Observation: This improved precision to 0.45 while maintaining recall at 0.45, achieving balanced performance. However, the model remained conservative and often missed rare or implicit clinical cues, such as inferring fall risk from behavioral descriptions. This limitation in capturing nuanced patterns motivated Phase 3.

4.3. Phase 3: Supervised Fine-Tuning (The Recall Engine)

Motivation: While prompting provides task-specific guidance at inference time, supervised fine-tuning embeds schema knowledge directly into model parameters. To capture the full nuance of 193 field definitions—especially rare patterns such as specific skin conditions or IV catheter types—we hypothesized that parameter updates would enable more robust extraction than prompting alone.

Implementation: We employed Supervised Fine-Tuning (SFT) using QLoRA (Quantized Low-Rank Adaptation).

Hyperparameters:

- Base Models: Qwen-3-8B (Team, 2025)
- Quantization: 4-bit (NF4).
- LoRA Rank (r): 128 (High rank chosen to capture complex schema relations).
- LoRA Alpha: 256.
- Target Modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.
- Learning Rate: 2e-4.
- Epochs: 3.

Input Format During Training: Each training example was constructed as a conversational turn with the following structure:

- **User message:** Contains the complete nursing transcript along with the *full schema specification* (all 193 field definitions with their types, enumerations, and descriptions).
- **Assistant message:** Contains the gold-standard JSON output with all annotated observations for that transcript.

Schema Presentation Strategy: During fine-tuning, we presented the complete 193-field schema in every training example, rather than using clustered subsets. This differs from our inference approach (Phase 1), where clustering is employed to reduce context load. The rationale is that fine-tuning allows the model to internalize schema knowledge through parameter updates, eliminating the need for clustering during training. By exposing the model to the full schema repeatedly across 237 examples over 3 epochs (711 total exposures), we enable comprehensive schema memorization.

Prompt Template: We applied the same 9-step Chain-of-Thought prompt (described in Phase 2) during fine-tuning. Each training example began with the complete 9-step system prompt, followed by the transcript and full schema, with the model trained to generate the corresponding JSON output. This ensures alignment between training and inference procedures.

Training-Inference Alignment: A key architectural consideration is the difference between our training and inference pipelines. During training, the model observes complete schema presentations to internalize all 193 fields through parameter updates. During inference (Section 5), we employ Phase 1 clustering to reduce context load, as the model has already internalized schema knowledge through fine-tuning. This hybrid approach combines the benefits of comprehensive training with efficient clustered inference, preventing context overflow while maintaining schema coverage.

Observation (Recall Explosion): The SFT model demonstrated a dramatic behavioral shift. It learned to identify implicit clinical cues that the prompted model missed.

- Prompted Model: Saw "patient trying to get OOB" $\rightarrow$ Extracted Nothing.
- SFT Model: Saw "patient trying to get OOB" $\rightarrow$ Extracted "Fall Risk: Impulsive" and "Mobility: Unsteady".
- Recall: Jumped to 0.75.
- Challenge: The model became "trigger-happy", often inferring conditions where none existed (Precision drop to 0.21).

#	Step Name	Prompt Content
1	Role Attribution	You are an expert Clinical Documentation Improvement (CDI) Specialist with 20 years of experience. You are precise, conservative, and strictly adhere to provided schemas.
2	Transcript Definition	The input text is a 'Transcript' of a clinical encounter. It may contain disfluencies, interruptions, and informal language.
3	Task Definition	Your task is to analyze the Transcript and extract specific clinical variables defined in the Schema.
4	Type Definitions	*[Dynamic Section: Inserts rules for* `SINGLE_SELECT`*,* `MULTI_SELECT`*, or* `NUMERIC` *based on current item]*
5	Output JSON Key Definitions	Return a JSON list of objects with keys: `"id"`, `"name"`, `"value"`, `"value_type"`.
6	Reasoning Guidelines (Negative Constraints)	• ABSOLUTELY NO inference of 'Normal' if not explicitly stated. • If the transcript is silent, SKIP the item. • Silence $\neq$ Normalcy.
7	Evidence Requirement	Before extracting, you must locate the specific mention in the transcript.
8	Overall Guidelines	• Accuracy is paramount. Do not round unless necessary. • Strict Enum Matching: The `value` must strictly match one of the provided options.
9	Eliciting JSON Output	`TRANSCRIPT`: {transcript} `SCHEMA`: {schema_subset} Answer with the JSON list only.

Table 1: The 9-Step Chain-of-Thought System Prompt Protocol used for the "CDI Specialist" persona (Corbeil et al., 2025b).

4.4. Phase 4: Transcript-Verified Post-Processing (The Precision Engine)

Motivation: Generative models are probabilistic and inherently prone to hallucination, particularly when trained to maximize recall. To achieve clinical-grade precision, we must constrain the output with deterministic rules. We treat the fine-tuned model as a candidate generator and employ post-processing as a precision-focused verifier.

Implementation: We developed a robust Python-based verification engine.

- **Algorithm 1: Transcript Verification**
  This algorithm ensures that every extracted string actually exists in the source. This filtered out pure hallucinations (e.g., model inventing "Pain: 5/10" when pain was not mentioned).
- **Algorithm 2: Schema-Based Phantom Filter**
  The model often defaults to "Normal" values (e.g., "Skin: Intact") even when the input is silent, violating the schema's requirement for "Null" values. We enforce strict Schema Adherence by utilizing a `STRICT_DENY_LIST`. If item is in a "Nullable" Schema Category (e.g., Skin, Mood) and Value is in a "Normative Set" (e.g., "Intact", "Normal"), the value is automatically rejected to ensure compliance with the Schema's "No Default" policy.
- **Algorithm 3: Enum Snapshot**
  We enforced strict schema compliance using Levenshtein distance.
  - **Model Output:** "Alert"
  - **Schema Options:** ["Alert and oriented x3", "Confused", "Unresponsive"]
  - **Action:** Match "Alert" to "Alert and oriented x3" (Distance < Threshold).
  - **Result:** Corrects lazy outputs to valid schema enumerations.

Observation: This phase was the most critical for precision, raising it from 0.21 to 0.61 without sacrificing the Recall gained in Phase 3.

5. Inference

Architecture: The inference engine iterates through the semantic clusters defined in Phase

1. For each pass, a dynamic system prompt is constructed containing *only* the definitions for that specific domain (e.g., "Neurological Single Selects"). This reduced the effective instruction token count from $\sim$8000 tokens to $<$1000 tokens per pass, preventing context overflow.

$$P(\text{Output} \mid \text{Transcript}, \text{Schema}) \approx \prod_{i=0}^{k} P(\text{Output}_i \mid \text{Transcript}, \text{Cluster}_i) \quad (1)$$

6. Evaluation

We evaluated the pipeline on a held-out test set of clinical transcripts using standard Information Extraction metrics.

- **Precision**: Measure of correctness (Are extracted items real?).
- **Recall**: Measure of completeness (Did we find all items?).
- **F1 Score**: Harmonic mean of Precision and Recall.

We specifically focused on the **Recall-Precision Trade-off**, analyzing how SFT improved Recall while Verification recovered Precision.

7. Results

7.1. Development Set Results

We first present performance on the development set, which guided our design decisions and hyperparameter selection for each phase.

Phase	Method	F1	Rec	Prec
0	Zero-Shot	0.14	0.12	0.16
1	Clustering	0.31	0.69	0.20
2	Prompting	0.43	0.43	0.43
3	SFT (Qwen-3-8B)	0.60	0.73	0.51
4	+ Verification	**0.64**	**0.70**	**0.59**

Table 2: Performance Evolution by Phase on Development Set

7.2. Test Set Results

Final evaluation on the held-out test set confirms the generalization of our approach.

Phase	Method	F1	Rec	Prec
0	Zero-Shot	$<$0.15	Low	Low
1	Clustering	0.33	0.72	0.21
2	Prompting	0.45	0.45	0.45
3	SFT	0.62	0.75	0.53
4	Post-Proc	**0.66**	**0.72**	**0.61**

Table 3: Performance Evolution by Phase

7.3. Analysis and Discussion

Why Clustering Improved Performance (Phase 1): The clustering phase (F1: 0.14 $\rightarrow$ 0.31) directly addresses the "Lost in the Middle" phenomenon documented by Liu et al. (Liu et al., 2023). By decomposing the 193-field schema into 15 semantic clusters, we reduced the effective context from approximately 8,000 tokens to fewer than 1,000 tokens per pass. This prevents catastrophic forgetting and eliminates cross-domain hallucinations—for example, the model no longer confused "Fall Risk" assessments with "Mobility" status. However, clustering alone achieved high recall (0.69) but poor precision (0.20), as the model still lacked explicit guidance on extraction constraints and when to abstain from inference.

Why Prompting Balanced Precision and Recall (Phase 2): The systematic prompting phase (F1: 0.31 $\rightarrow$ 0.43) introduced explicit constraints through our 9-step Chain-of-Thought protocol. The key innovation was the negative constraint explicitly forbidding inference of "Normal" values when the transcript was silent. This principle—"silence does not equal normalcy"—balanced precision and recall at 0.43 each, correcting the precision deficit from Phase 1. However, the prompted model remained conservative and failed to capture implicit clinical cues. For example, it could not infer "Fall Risk" from behavioral descriptions like "trying to get out of bed" without explicit fall risk language, limiting recall potential.

Why SFT Dramatically Boosted Recall (Phase 3): Fine-tuning (F1: 0.43 $\rightarrow$ 0.62) embedded schema knowledge directly into model parameters through 711 training exposures (237 examples $\times$ 3 epochs). The model internalized implicit clinical reasoning patterns from annotated examples, enabling it to recognize subtle indicators. Recall jumped to 0.73 (+70% relative improvement), as the model learned associations like "trying to get OOB" $\rightarrow$ "Fall Risk: Impulsive" and "appears uncomfortable" $\rightarrow$ potential pain indicators. However, this sensitivity came at a cost: precision increased only marginally to 0.51 (+19% relative), as the model occasionally over-inferred conditions not explicitly stated. This demonstrates the fundamental precision-recall trade-off inherent in generative

extraction systems.

Why Post-Processing Recovered Precision (Phase 4): The verification phase (F1: 0.62 → 0.66) applied deterministic rules to filter false positives while preserving true positives. The three-algorithm pipeline increased precision from 0.51 to 0.59 (+16% relative) with only a minor recall decrease (0.73 → 0.70, -4% relative). This demonstrates the effectiveness of treating fine-tuned models as candidate generators followed by rule-based precision verification. The modest recall loss occurs because some valid implicit inferences lack direct textual evidence (e.g., inferring alertness from conversational engagement), causing them to be filtered by transcript verification. However, the substantial precision gain outweighs this minor recall cost.

Key Trade-offs and Optimal Strategy: Our experiments reveal a systematic precision-recall trade-off at each phase. Clustering sacrifices precision for recall by reducing context overload. Prompting balances both but limits recall due to conservatism. Fine-tuning maximizes recall through pattern learning but increases false positives. Post-processing recovers precision with minimal recall loss through verification. The optimal strategy combines all four phases sequentially, achieving both high recall (0.72) and high precision (0.61) by leveraging the complementary strengths of each approach.

7.4. Qualitative Error Analysis

We analyzed the remaining error modes in the final Phase 4 output:

1. **Semantic Ambiguity**: Confusion between related fields (e.g., "Mobility" vs "Activity Level"). The model typically picks the more common one.

2. **Implicit Negation**: The text says "Patient denies pain", but the model extracts "Pain Severity: 0". While clinically true, if the schema requires "Null" for "No Pain", this counts as a False Positive using strict evaluation.

3. **Complex Temporal Logic**: In long transcripts, the patient status changes (e.g., "Alert" → "Confused"). The model sometimes struggles to "overwrite" the earlier state with the latest one.

8. Conclusion

We have presented a robust framework for high-dimensional clinical data extraction. By combining **Semantic Clustering**, **Chain-of-Thought Prompting**, **Supervised Fine-Tuning**, and **Deterministic Verification**, we overcame the limitations of off-the-shelf LLMs. Our system provides a blueprint for deploying Generative AI in high-stakes healthcare environments: **Trust (the model), but Verify (the transcript).**

9. Code Availability

The complete implementation, including inference and training scripts, schema definitions, training logs, and detailed documentation, is available at: GitHub Repository: ltrc-medicom-mediqa-synur-clinicalnlp-2026

10. Limitations & Ethical Considerations

- **Data Bias**: Our fine-tuning dataset was derived from a specific institution's transcript style. The model may underperform on dialects or documentation styles not represented in the training distribution.
- **Privacy**: All experiments used de-identified datasets. However, deploying Generative AI in healthcare requires strict adherence to HIPAA/GDPR. The "Hallucination" risk, while mitigated, is never zero; thus, this system is designed as a "Human-in-the-Loop" assistive tool, not an autonomous coder.
- **Compute Costs**: While inference is optimized, the SFT process requires GPU resources that may be inaccessible to low-resource settings.

11. References

Monica Agrawal et al. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

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. 2025a. 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, Suzhou (China). Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and Francois Beaulieu. 2025b. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. *arXiv preprint arXiv:2510.26974*.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Gemma Team. 2024. Gemma.

Qwen Team. 2025. Qwen3 technical report.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.