Contextual Clinical Extraction: Integrating Foundation Models with Domain-Specific Validation Rules

Ankit Singh, Siva Satyanarayana Raju
Team Gladiators MEDIQA-SYNUR @ ClinicalNLP / LREC 2026
singhankit16@gmail.com, pusapati.badri@gmail.com

Abstract

We present a hybrid system for schema-guided clinical information extraction from synthetic nursing dictations, developed for the MEDIQA-SYNUR shared task at ClinicalNLP 2026. Our approach combines a large language model (Claude Opus 4) with comprehensive prompt engineering, multi-layered post-processing pipelines, and rule-based extraction using over 400 domain-specific regex patterns. The system extracts structured observations across 193 clinical concepts spanning vital signs, neurological assessments, respiratory status, and functional evaluations. On the development set, our system achieves 74.27% precision, 84.32% recall, and 78.98% F1 score. Detailed error analysis reveals that secondary diagnosis over-extraction (32 false positives) and cognitive status classification (12 value errors) represent primary challenges. We provide comprehensive methodology documentation to enable reproducibility, including prompt templates, validation rules, and regex patterns. We discuss limitations including cost-performance tradeoffs of proprietary LLMs and deployment considerations for healthcare settings.

Keywords: clinical NLP, information extraction, large language models, nursing documentation, schema-guided extraction, MEDIQA, hybrid systems

1. Introduction

Clinical documentation represents a significant burden in modern healthcare, with nurses spending substantial time on administrative tasks that detract from direct patient care. Studies estimate that nurses spend up to 35% of their shift on documentation (Baumann et al., 2018), and the transition to Electronic Health Records (EHRs) has not reduced this burden as initially hoped (Joukes et al., 2018). Voice-based documentation systems offer potential efficiency gains by allowing clinicians to dictate observations naturally, but converting these narratives into structured EHR entries requires sophisticated natural language understanding.

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) addresses this challenge by focusing on extracting structured clinical observations from synthetic nursing dictation transcripts. The task presents several technical challenges: (1) a large schema space of 193 clinical concepts across diverse domains; (2) four distinct value types (NUMERIC, STRING, SINGLE_SELECT, MULTI_SELECT) requiring different extraction strategies; (3) strict ontological constraints where extracted values must exactly match enumerated options; and (4) the need to distinguish explicitly stated observations from implicit or speculative information.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in clinical NLP tasks, including medical entity recognition (Agrawal et al., 2022), clinical question answering (Singhal et al., 2023), and medical text summarization. However, direct application of LLMs to structured clinical data extraction faces challenges including hallucination of plausible but incorrect observations, inconsistent output formatting, and failure to adhere to strict ontological constraints (Ji et al., 2023). These issues are particularly concerning in clinical settings where extraction errors could affect patient care.

We propose a hybrid approach that combines the reasoning and generalization capabilities of foundation models with the precision and reliability of rule-based validation systems. Our contributions include:

(1) A comprehensive prompt engineering strategy incorporating complete schema specification, few-shot learning with diverse training examples, and domain-specific extraction guidelines for challenging clinical concepts.

(2) A multi-layered post-processing pipeline with six specialized filters addressing common LLM extraction errors, including speculative diagnosis removal, physiological range validation, and context-based filtering.

(3) A rule-based extraction component employing over 400 contextually-aware regex patterns organized by clinical domain, with validation mechanisms to prevent false positives.

(4) An intelligent merging strategy that balances LLM comprehensiveness with rule-based precision, along with detailed error analysis quantifying system performance across error types and clinical concepts.

2. Related Work

2.1 Clinical Information Extraction

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 183–191

Clinical information extraction has a rich history spanning rule-based, statistical, and neural approaches. Early systems relied heavily on medical ontologies and hand-crafted rules. MetaMap (Aronson and Lang, 2010) maps biomedical text to UMLS concepts using lexical and syntactic analysis, achieving high precision for concept identification but requiring extensive manual curation. The clinical Text Analysis and Knowledge Extraction System (cTAKES; Savova et al., 2010) provides a comprehensive pipeline for clinical NLP including named entity recognition, assertion classification, and relation extraction, built on rule-based and machine learning components. These systems established important baselines but struggled with the linguistic variability of clinical narratives.

Negation detection represents a critical subtask in clinical NLP, as clinicians frequently document absent findings. The NegEx algorithm (Chapman et al., 2001) uses simple regular expressions to identify negation triggers and their scope, achieving strong performance on discharge summaries. Extensions like ConText (Harkema et al., 2009) broadened scope to include temporality and experiencer identification. Our system incorporates similar negation handling through both rule-based pattern matching and LLM prompt instructions.

2.2 Transformer-Based Clinical NLP

The transformer architecture (Vaswani et al., 2017) revolutionized clinical NLP through domain-specific pretraining. BioBERT (Lee et al., 2020) pretrained BERT on PubMed abstracts and PMC full-text articles, demonstrating improved performance on biomedical NER and relation extraction. ClinicalBERT (Alsentzer et al., 2019) further adapted BERT to clinical text using MIMIC-III clinical notes, showing gains on clinical NLP benchmarks. PubMedBERT (Gu et al., 2021) demonstrated that pretraining exclusively on domain-specific text outperformed mixed-domain approaches, achieving state-of-the-art results on multiple biomedical NLP tasks.

However, these encoder-based models require task-specific fine-tuning and struggle with large schema spaces like the 193 concepts in MEDIQA-SYNUR. Fine-tuning separate classifiers for each concept would be impractical and would not leverage cross-concept dependencies. Our approach addresses this by using a generative LLM that can handle the full schema in a single prompt.

2.3 Large Language Models for Clinical Tasks

The emergence of large language models has opened new possibilities for clinical NLP. Brown et al. (2020) demonstrated that GPT-3 could perform few-shot learning across diverse NLP tasks without fine-tuning. Subsequent work applied this capability to clinical domains: Agrawal et al. (2022) showed that LLMs are effective few-shot clinical information extractors, achieving competitive performance with minimal task-specific examples. Singhal et al. (2023) demonstrated that LLMs encode substantial clinical knowledge and can perform medical question answering at expert level.

However, hallucination remains a significant concern for LLM-based extraction (Ji et al., 2023). LLMs may generate plausible but incorrect clinical observations, particularly when prompted to extract information that is not explicitly stated. This is especially problematic for secondary diagnoses where the model may infer conditions from symptoms. Our post-processing pipeline specifically addresses this through speculative language filtering and schema validation.

2.4 Schema-Guided Extraction

Schema-guided extraction constrains model outputs to conform to predefined ontologies. Rastogi et al. (2020) introduced the Schema-Guided Dialogue dataset, demonstrating that providing schema descriptions improves zero-shot generalization to new domains. Chen et al. (2021) extended this to multi-domain dialogue state tracking. Our approach adapts schema-guided principles to clinical documentation, where ontological constraints are stricter and extraction errors have higher stakes.

Hybrid approaches combining neural models with rule-based post-processing have shown promise in clinical NLP. Wang et al. (2018) surveyed clinical information extraction applications, noting that the best-performing systems often combine multiple approaches. Our system follows this paradigm, using LLM extraction for broad coverage and rule-based validation for precision.

3. Task Description and Dataset

3.1 MEDIQA-SYNUR Shared Task

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) focuses on extracting structured clinical observations from synthetic nursing dictation transcripts. The task simulates a real-world scenario where nurses verbally document patient assessments, and an NLP system must convert these dictations into structured EHR entries conforming to a clinical ontology.

The dataset was generated through a controlled multi-agent simulation pipeline to ensure diverse and clinically realistic scenarios. Gold-standard annotations were produced by expert nurses using an open-source, large-scale clinical ontology. The dataset comprises 122 training samples, 101 development samples, and a held-out test set for final evaluation. Each sample contains a transcript field with natural

nursing dictation and an observations field containing structured annotations.

3.2 Clinical Observation Schema

The observation schema defines 193 clinical concepts organized across multiple clinical domains. Table 1 summarizes the schema distribution by domain and value type. The domains span the full scope of nursing assessment, from vital signs and neurological status to functional capacity and safety considerations.

Domain	Concepts	Example Fields
Vital Signs	15	O2 sat, HR, BP, temp, MAP
Neurological	22	Orientation, GCS, pupils, cognition
Respiratory	18	Dyspnea, breath sounds, O2 delivery
Cardiovascular	14	Rhythm, edema, JVD, pulses
Gastrointestinal	16	Bowel sounds, nausea, vomiting
Genitourinary	12	Urine output, appearance, odor
Musculoskeletal	18	Motor strength, mobility, ROM
Skin/Wounds	20	Turgor, wounds, pressure injuries
Safety	15	Fall risk, bed alarm, restraints
Functional	25	ADL assistance, transfers, gait
Other	18	Pain, nutrition, IV therapy
Total	193	-

Table 1: Distribution of clinical concepts by domain in the SYNUR schema.

Each observation concept is associated with one of four value types, each requiring different extraction and validation strategies:

NUMERIC: Quantifiable measurements such as vital signs (heart rate, oxygen saturation, temperature) and volumes (urine output, emesis). These require extraction of numeric values and validation against physiologically plausible ranges.

STRING: Free-text observations such as pain descriptions, wound characteristics, and patient identification. These require minimal normalization but may contain verbose LLM outputs that need cleaning.

SINGLE_SELECT: Categorical observations with mutually exclusive enumerated options, such as dyspnea severity (None/Mild/Moderate/Severe) or cognitive status. These require exact matching against the schema's allowed values.

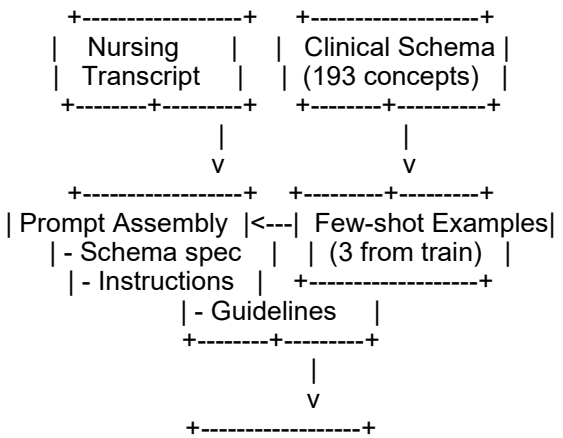
MULTI_SELECT: Observations allowing multiple concurrent values from an enumerated set, such as breathing patterns (labored, shallow, irregular) or respiratory interventions. These require parsing of conjunctions and list structures.

3.3 Evaluation Metrics

The official evaluation uses precision (correct predictions / total predictions), recall (correct predictions / total gold annotations), and F1 score. A prediction is considered correct only if both the observation ID and value exactly match the gold standard. For MULTI_SELECT fields, the extracted values must match as sets (order-independent). Numeric values are compared with type coercion (integer vs. float).

4. System Architecture

Our hybrid extraction system consists of three main components operating in sequence: (1) an LLM-based primary extractor using Claude Opus 4 with comprehensive prompt engineering; (2) a multi-layered post-processing pipeline with domain-specific filters and corrections; and (3) a rule-based supplementary extractor with contextual validation. Results from both extraction methods are merged through an intelligent strategy and validated against the clinical schema. Figure 1 illustrates the complete pipeline.



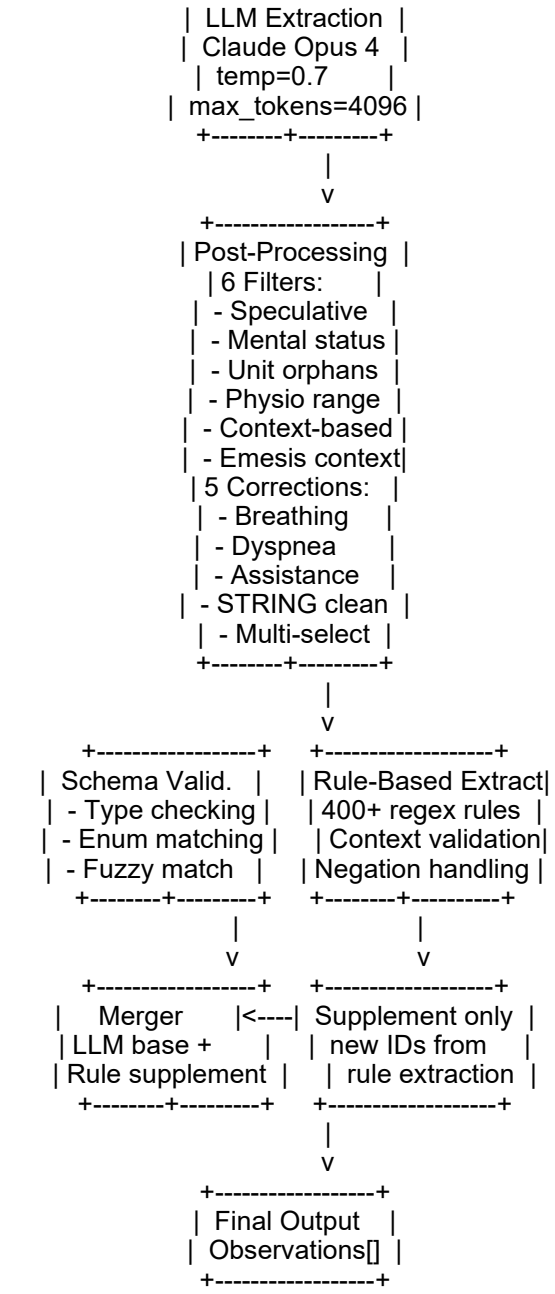


Figure 1: System architecture showing the complete hybrid extraction pipeline with all processing stages.

4.1 LLM-Based Extraction

We use Claude Opus 4 (model ID: claude-opus-4-5-20251101) as the primary extraction engine. The model is accessed via the Anthropic API with the following configuration: temperature=0.7 for balanced consistency and flexibility, and max_tokens=4096 constraining output length only. The input prompt, including schema specification, few-shot examples, and transcript, fits within the model's context window without truncation.

The prompt engineering strategy incorporates five key components designed to maximize extraction accuracy while minimizing hallucination:

Schema Specification: The complete schema of 193 observation concepts is provided in the prompt, including concept IDs, names, value types, and enumerated options for SELECT fields. This explicit schema grounding helps the model understand the extraction targets and valid output values.

Few-Shot Examples: We include 3 training samples as few-shot examples, selected to demonstrate diverse observation types. Examples are chosen to cover: (a) multiple value types in a single transcript, (b) challenging concepts like multi-select respiratory interventions, and (c) proper handling of negated or absent findings.

Extraction Rules: The prompt explicitly instructs the model to: (a) extract only observations that are explicitly stated in the transcript; (b) avoid speculation or inference from symptoms; (c) preserve exact wording for STRING values; (d) map to schema-compliant values for SELECT types; and (e) handle negated findings appropriately.

Domain-Specific Guidelines: Special instructions address challenging clinical concepts: (a) respiratory observations require explicit dyspnea severity mentions rather than inference; (b) cognitive status should distinguish between alert, confused, and disoriented states; (c) secondary diagnoses should only include confirmed conditions, not differentials.

Multi-Select Handling: Instructions for parsing conjunctions ('nausea and vomiting') and list structures ('raising head of bed, turn cough, deep breathe') to extract multiple values for MULTI_SELECT fields.

4.2 Post-Processing Pipeline

The LLM output undergoes extensive post-processing through six specialized filters and five correction mechanisms. These address systematic extraction errors identified through iterative development on the training and development sets.

4.2.1 Filtering Mechanisms

Speculative Diagnosis Filter: Removes secondary diagnosis extractions containing uncertain language markers: 'possibly', 'could be', 'might', 'suspected', 'likely', 'probably', 'suggestive of', 'consistent with', 'differential', 'rule out', 'r/o', 'versus', 'signs of'. This prevents extraction of differential diagnoses as confirmed conditions.

Mental Status Validation: The 'forgets limitations' mental status value is only retained if explicitly mentioned in the transcript. General forgetfulness or confusion does not trigger this extraction, preventing over-generalization from related cognitive findings.

Unit Field Orphan Removal: Removes unit fields (e.g., 'heart rate unit') when the corresponding value field was not extracted. This addresses LLM tendency to extract partial observation pairs.

Physiological Range Validation: Numeric values are validated against clinically plausible ranges: heart rate 40-200 bpm, respirations 8-40/min, oxygen saturation 50-100%, temperature 32-43 C, mean arterial pressure 40-150 mmHg. Values outside these ranges are filtered.

Context-Based Filtering: Cardiac values (heart rate) require cardiac context keywords within 50 characters ('heart rate', 'pulse', 'HR', 'bpm'). Respiratory values require respiratory context ('respiration', 'RR', 'breaths per minute'). This prevents numeric values from unrelated contexts being incorrectly mapped.

Emesis Context Filter: Prevents extraction of urine appearance values (e.g., 'dark') when the transcript context discusses emesis rather than urinary output. Triggered by emesis keywords ('vomit', 'emesis') in proximity to the extracted value.

4.2.2 Correction Mechanisms

Breathing Pattern Correction: Maps descriptions like 'labored breathing' to the correct MULTI_SELECT breathing pattern values. Handles variations like 'breath is labored', 'with labored breathing', 'breathing is shallow'.

Dyspnea Severity Mapping: Infers dyspnea severity from descriptions when explicit severity is not stated: 'labored breathing' maps to 'Severe', 'short of breath with exertion' maps to 'Mild', 'severe respiratory distress' maps to 'Severe'.

Assistance Field Normalization: Standardizes variations of assistance levels to schema-compliant values: 'requires assistance' to 'Assisted', 'fully dependent' to 'Dependent', 'able to perform independently' to 'Independent'.

STRING Field Cleaning: Removes units from numeric STRING fields (e.g., '150 mLs' to '150' for PO intake) and converts verbose yes/no responses ('No known allergies', 'NKDA') to simple 'No'. Normalizes whitespace and removes trailing punctuation.

Multi-Select Conjunction Parsing: Splits conjunction phrases for MULTI_SELECT fields: 'nausea and vomiting' becomes ['nausea', 'vomiting'], 'numbness and tingling' becomes separate values. Handles various conjunction patterns ('and', 'with', commas).

4.3 Rule-Based Extraction

The rule-based component employs over 400 regular expression patterns organized by clinical domain and value type. These patterns provide high-precision extraction for common clinical values and serve as supplementary extractions for observations missed by the LLM.

4.3.1 Pattern Organization

Vital Sign Patterns (~65 patterns): Target numeric extractions for oxygen saturation, heart rate, respirations, blood pressure, temperature, mean arterial pressure, urine output, emesis volume, bladder scan volume, oxygen flow rate, and fall risk scores. Patterns handle various unit notations and contextual phrasing.

Categorical Patterns (~180 patterns): Target SINGLE_SELECT fields including oxygen delivery devices, dyspnea severity, cognitive status, orientation levels, Glasgow Coma Scale components, bed position, breath sounds, cough descriptions, fall risk levels, nausea/vomiting, edema characteristics, mobility status, and assistance levels.

Multi-Select Patterns (~155 patterns): Target MULTI_SELECT fields including respiratory interventions, breathing patterns, sensory symptoms, delirium symptoms, urine appearance, and abdomen exam findings. These patterns extract individual values that are later combined.

Field	Pattern Example	Matches
O2 Sat	(?:O2 sat|SpO2)[^\d]*(\d{2,3})\s*%	O2 sat 94%
Heart Rate	(?:heart rate|HR)[^\d]*(\d{2,3})	heart rate 88
Dyspnea	severe\s+dyspnea	severe dyspnea
Cognition	alert\s+(?:but|with)\s+confusion	alert but confused
Breathing	breath(?:ing)?\s+(?:is\s+)?labored	breathing is labored

Table 2: Example regex patterns for different observation types.

4.3.2 Contextual Validation

Contextual validation prevents false positive extractions through several mechanisms:

Keyword Proximity: Cardiac extractions require cardiac keywords within 50 characters of the numeric value. Similarly, respiratory extractions require respiratory context. This prevents numeric values from unrelated measurements being incorrectly categorized.

Unit Compatibility: Values with incompatible units are rejected. For example, a heart rate extraction is rejected if followed by volume units

('cc', 'mL'), which would indicate a fluid measurement rather than a pulse rate.

Substring Safeguards: Patterns avoid matching within larger words. For example, 'labored' should not match within 'nonlabored'. This is implemented through word boundary assertions and negative lookbehind patterns.

Negation Detection: A simple negation check filters extractions where the observation appears in negated context. Negation markers include 'no', 'denies', 'without', 'absent', 'negative for', appearing within 30 characters before the term.

4.4 Schema Validation and Merging

Schema validation ensures all extracted observations conform to the SYNUR ontology before final output. The validation process differs by value type:

NUMERIC Validation: Values are extracted using regex, converted to appropriate numeric types (integer if whole number, float otherwise), and validated against physiological ranges as described in Section 4.2.1.

STRING Validation: Values are cleaned by removing extraneous whitespace, normalizing common patterns (verbose yes/no, unit removal from numeric strings), and truncating excessive length. Empty strings after cleaning are rejected.

SINGLE_SELECT Validation: Values are matched against the schema's enumerated options. Exact matching (case-insensitive) is attempted first. If no exact match, fuzzy matching compares lowercased values using substring containment (e.g., 'moderate' matches 'Moderate dyspnea'). This fuzzy matching can introduce false positives when multiple options share substrings, but empirically improves recall more than it hurts precision.

MULTI_SELECT Validation: Input values (as list or single value) are matched individually against enumerated options. Substring matching is applied with safeguards against partial matches (e.g., 'labored' should not match options containing 'nonlabored').

The merging strategy combines LLM and rule-based extractions as follows: (1) LLM extractions form the base result set; (2) rule-based extractions supplement for observation IDs not already extracted by the LLM; (3) a seen-ID set prevents duplicate extractions; (4) LLM results take priority when both methods extract the same field. This approach leverages LLM comprehensiveness while supplementing with high-precision rule-based findings.

5. Experimental Setup

We evaluate our system on the MEDIQA-SYNUR benchmark using the official evaluation script provided by the shared task organizers. The 122 training samples are used exclusively for few-shot example selection; no fine-tuning or parameter updates are performed on any model component.

All development and error analysis reported in this paper is conducted on the 101-sample development set. This set was used for iterative system refinement, including prompt engineering, filter development, and regex pattern creation. Test set results are submitted separately to the shared task evaluation system and are not included in this paper to ensure fair comparison with other participants.

The complete system is implemented in Python 3.x using the Anthropic API for LLM access, standard library regex for pattern matching, and JSON/JSONL for data handling. Processing the full development set requires approximately 15-20 minutes using sequential API calls with default rate limiting.

6. Results

6.1 Overall Performance

Table 3 presents our results on the development set. The system achieves 74.27% precision, 84.32% recall, and 78.98% F1 score. The system generated 1578 predictions against 1390 gold observations, achieving 1108 true positives.

Metric	Value	Count
Precision	74.27%	1108/1578
Recall	84.32%	1108/1390
F1 Score	78.98%	-
True Positives	1108	-
False Positives	364	-
False Negatives	206	-
Gold Observations	1390	-
System Predictions	1578	-

Table 3: Development set results computed using the official MEDIQA-SYNUR evaluation script.

6.2 Error Analysis

Table 4 provides a quantitative breakdown of errors by type. Of 479 total errors, 'Extra' predictions (false positives) account for 273 (57.0%), followed by 'Missing' observations (false negatives) at 115 (24.0%), and 'Wrong Value' errors at 91 (19.0%).

Error Type	Count	%	Description
Missing (FN)	115	24.0%	Gold obs not extracted
Extra (FP)	273	57.0%	Extracted but not in gold
Wrong Value	91	19.0%	Correct ID, wrong value
Total	479	100%	-

Table 4: Error breakdown by type on development set.

6.3 Field-Level Error Distribution

Table 5 shows the most error-prone fields across all error types. Secondary diagnosis emerges as the most challenging concept, contributing errors across all three categories.

Field	Missing	Extra	Wrong	Total
Secondary diagnosis	6	32	16	54
Gastrointestinal symptoms	0	22	0	22
Cognitive status	6	0	12	18
Patient safety	11	0	0	11
Dyspnea	0	0	10	10

Table 5: Top error-prone fields with breakdown by error type.

The majority of errors (57%) are false positives, dominated by over-extraction of secondary diagnoses (n=32). Despite the speculative language filter, the LLM tends to extract conditions mentioned in clinical context that are not explicitly documented as current diagnoses. For example, a transcript mentioning 'history of diabetes' may trigger extraction of diabetes as a secondary diagnosis when the gold standard only includes active conditions.

Wrong value errors concentrate in cognitive status (n=12) and dyspnea (n=10). Cognitive status errors often involve mapping between similar options such as 'Alert' versus 'Alert and oriented' versus 'Alert with general confusion'. Dyspnea errors reflect difficulty inferring severity levels when not explicitly stated.

Missing observations (n=115) frequently involve patient safety (n=11) and mobility (n=7). These concepts often require multi-step inference from implicit information - for example, inferring patient safety measures from mention of bed alarms or fall precautions without explicit safety documentation.

7. Discussion

7.1 Clinical Utility Assessment

Our system achieves 84.3% recall, suggesting that the majority of clinical observations would be captured in a real deployment scenario. This high recall is valuable for reducing documentation burden, as nurses would not need to manually enter most observations. However, 74.3% precision implies that approximately 1 in 4 extracted observations would require correction.

In practice, this performance profile suggests a semi-automated workflow where the system pre-populates observations for nurse verification rather than fully autonomous documentation. Nurses can quickly confirm correct extractions while correcting errors, potentially providing significant time savings compared to manual entry while maintaining documentation accuracy.

The error distribution has implications for clinical deployment. False positives (over-extraction) are generally easier to catch during review than false negatives (missed observations), as nurses can quickly dismiss incorrect pre-populated values. Missing observations require more careful review against the original dictation. Our system's bias toward recall over precision aligns with this workflow consideration.

7.2 Deployment Considerations

The reliance on a proprietary LLM (Claude Opus 4) presents deployment challenges for healthcare settings. Data privacy regulations (HIPAA in the US, GDPR in Europe) may restrict transmission of patient data to external APIs. While we used the Anthropic API for development, production deployment would likely require: (a) on-premises model hosting, (b) privacy-preserving architectures with data de-identification, or (c) use of open-source alternatives.

Cost considerations are significant for high-volume healthcare documentation. Claude Opus 4 pricing at the time of development was approximately $15/1M input tokens and $75/1M output tokens. Processing the 101-sample development set cost approximately $8-10 USD. For a hospital processing thousands of nursing assessments daily, API costs could be substantial. Smaller, faster models (Claude Haiku, GPT-4o-mini) may offer better cost-performance tradeoffs.

Our rule-based component provides a potential fallback when LLM access is unavailable or cost-prohibitive. While we did not evaluate standalone rule-based performance, the 400+ patterns cover the majority of common vital signs and categorical observations. A tiered approach using rule-based extraction for common cases and LLM for complex narratives could optimize cost-effectiveness.

8. Limitations

Our study has several limitations that should be considered when interpreting results and planning future work:

Synthetic Data: We evaluate exclusively on synthetic nursing dictations generated through a simulation pipeline. Real-world transcripts may exhibit greater linguistic variability, speech recognition errors, background noise artifacts, and documentation inconsistencies. The generalization of our approach to authentic clinical settings remains to be validated.

Proprietary Model: Our use of Claude Opus 4, a proprietary closed-source model, limits reproducibility and raises concerns about deployment in regulated healthcare environments. The model's behavior may change with API updates, and there is no guarantee of long-term availability or consistent pricing.

Development Set Evaluation: All reported results are on the development set, which was used for iterative system refinement including prompt engineering, filter development, and regex pattern creation. This introduces potential overfitting to development-specific patterns. Test set performance, evaluated separately through the shared task submission system, provides a more rigorous assessment of generalization.

No Ablation Studies: We do not report ablation studies isolating the contribution of individual components (LLM-only vs. post-processing vs. rule-based extraction). Such studies would better quantify the value added by each pipeline stage and guide future optimization efforts.

Independent Extraction: Our approach extracts observations independently without modeling inter-observation dependencies. Clinical findings often correlate (e.g., tachycardia with fever, edema with heart failure), and a more holistic approach might improve both extraction accuracy and clinical validity checking.

Single LLM Evaluation: We evaluate only Claude Opus 4 without comparison to other LLMs (GPT-4, Llama, Mistral) or domain-specific models (BioBERT, ClinicalBERT). Comparative evaluation would inform model selection for different deployment scenarios.

9. Conclusion

We presented a hybrid system for schema-guided clinical observation extraction from nursing dictations, combining LLM-based extraction with domain-specific post-processing and rule-based augmentation. On the MEDIQA-SYNUR development set, our approach achieves 78.98% F1 score with 74.27% precision and 84.32% recall.

Our error analysis reveals that secondary diagnosis over-extraction and cognitive status value mapping are primary challenges. The high recall suggests the system could reduce documentation burden in a semi-automated workflow, though precision limitations necessitate human review. We provide detailed methodology documentation to enable reproducibility.

While our results demonstrate the potential of LLMs for clinical documentation automation, practical deployment requires addressing cost, privacy, and validation concerns. Future work should include ablation studies, evaluation on authentic clinical data, comparison across LLM architectures, and exploration of open-source alternatives for deployable solutions.

10. Reproducibility

To facilitate reproducibility, we document key implementation details:

Model Configuration: Claude Opus 4 (claude-opus-4-5-20251101), temperature=0.7, max_tokens=4096. API accessed via anthropic Python package version 0.40+.

Prompt Structure: System prompt includes complete schema (193 concepts with IDs, names, types, enums), 3 few-shot examples from training set, extraction instructions emphasizing explicit-only extraction, and domain-specific guidelines for respiratory, cognitive, and diagnosis fields.

Post-Processing: Six filters (speculative diagnosis, mental status, unit orphans, physiological range, context-based, emesis context) and five corrections (breathing pattern, dyspnea severity, assistance normalization, STRING cleaning, multi-select parsing).

Rule-Based Patterns: 400+ regex patterns organized by domain (vital signs, categorical, multi-select) with contextual validation (keyword proximity, unit compatibility, substring safeguards, negation detection).

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In Proceedings of EMNLP 2022, pages 1998-2022.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Proceedings of the Clinical NLP Workshop, pages 72-78.

Aronson, A.R. and Lang, F.M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229-236.

Baumann, L.A., Baker, J., and Elshaug, A.G. (2018). The impact of electronic health record systems on clinical documentation times: A systematic review. Health Affairs, 37(4):655-663.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877-1901.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., and Buchanan, B.G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34(5):301-310.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borber, I., and Chang, S. (2021). Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In Proceedings of NAACL 2021, pages 1124-1134.

Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1):1-23.

Harkema, H., Dowling, J.N., Thornblade, T., and Chapman, W.W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics, 42(5):839-851.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1-38.

Joukes, E., de Keizer, N.F., de Bruijne, M.C., Abu-Hanna, A., and Cornet, R. (2018). Impact of electronic versus paper-based recording on the quality of clinical documentation in nursing. Studies in Health Technology and Informatics, 247:261-265.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234-1240.

Michalopoulos, G., Corbeil, J.P., Bader, C., Bodenstab, N., and Ben Abacha, A. (2026). Overview of the MEDIQA-SYNUR 2026 Shared Task on Synthetic Nursing Observation Extraction. In Proceedings of ClinicalNLP @ LREC 2026.

Rastogi, A., Zang, X., Srivastava, S., Guez, R., and Dilek, Z. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In Proceedings of AAAI 2020, pages 8689-8696.

Savova, G.K., Masanz, J.J., Ogren, P.V., et al. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5):507-513.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620:172-180.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998-6008.

Wang, Y., Wang, L., Rastegar-Mojarad, M., et al. (2018). Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 77:34-49.