SQUCS at MEDIQA-SYNUR 2026: A Multi-Agent Open Source LLM System for Nursing Observation Extraction

Riham Jeeballah[1], **Adhari AlZaabi**[2], **and Abdulrahman K. AAlAbdulsalam**[1]
[1]Sultan Qaboos University, Department of Computer Science, Oman
[2]Sultan Qaboos University, Department of Clinical and Human Anatomy, Oman
a.aalabdulsalam@squ.edu.om

Abstract

Clinical nursing documentation contains detailed observational information that is essential for patient monitoring and clinical decision-making, yet this information is predominantly recorded in free-text form. The MEDIQA-SYNUR shared task addresses this challenge by requiring systems to extract structured nursing observations from clinical transcripts under strict constraints on evidence grounding and value normalization. In this work, we present a multi-agent large language model (LLM)–based system for the MEDIQA-SYNUR task. We utilize the Llama3 open source LLM for this purpose for ease of local deployment within hospital digital infrastructure. Our system decomposes the extraction process into specialized agents responsible for schema-guided extraction, rule-based validation, and precision-oriented filtering. Starting from a baseline multi-agent pipeline, we conduct a systematic error analysis over the entire development set, examining all false positive and false negative predictions. Our final configuration, selected after extensive exploration and error analysis, combined transcript segmentation, the precision agent, and a suppression table derived from development-set analysis. On the development set, this setup achieved an F1 score of 0.6930 (precision = 0.6427, recall = 0.7518). Applying the same configuration directly to the test set, without any additional tuning, yielded an F1 score of 0.5923 (precision = 0.5292, recall = 0.6725). These results represent the most effective balance of precision and recall achieved through our iterative refinements and reflect the final state of the system as submitted for the competition.

Keywords: Clinical Natural Language Processing, Nursing Observation Extraction, Multi-Agent Systems, Large Language Models, MEDIQA-SYNUR, Evidence-Grounded Extraction

1. Introduction

Extracting information from clinical text reports has been an active area of research within the last two decades (Wang et al., 2018; Meystre et al., 2008). Converting unstructured narrative text into structured schema is important for medical research and effective clinical decision making (Kreimeyer et al., 2017).

The majority of studies focused on extracting information from clinical documentation produced by physicians, and recently more publications investigated extracting information specifically from nursing notes (Mitha et al., 2023). Most recent studies focused on identifying one particular information from text authored by nurses such as ICU length of stay and mortality from intensive care units (Huang et al., 2021), risk of fall (Bjarnadottir and Lucero, 2018), sexual orientation and gender, and presence of urinary catheter and related symptoms (Gundlapalli et al., 2017).

The SYNUR (Synthetic NURsing) dataset, on the other hand, was proposed to cover a variety of comprehensive clinical observations often found within nursing dictations (Corbeil et al., 2025a). In this study, we describe an approach that extracts structured observations from the SYNUR dataset as part of the MEDIQA-SYNUR shared task (George Michalopoulos, 2026). The best submission achieved an F1 score of 0.59 with recall of 0.67 and precision of 0.53 on the test set. In this study we describe the development of multi-agent LLM-based system for this task.

2. Related Work

Structured extraction from clinical free text has been approached through rule-based systems, machine learning, and, more recently, transformer-based models. A systematic review of 127 papers covering NLP methods applied to EHRs found that named entity recognition and clinical note classification now dominate the application landscape, with deep learning and transfer learning architectures increasingly applied across these tasks (Hossain et al., 2023). While these advances have improved extraction from physician-authored documents such as radiology reports and discharge summaries, nursing documentation poses distinct challenges that limit the direct transfer of existing methods. Unlike physician notes, which favour concise problem-oriented abstraction, nursing narratives are illustrative and descriptive, rely on institution-specific phrasing, and document a broader range of observations per encounter (Hyun et al., 2009). An integrative review of 43 NLP studies using nursing notes further confirmed that methodological approaches remain heterogeneous and that nursing-specific standard terminolo-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 125–135

gies are used in fewer than one in five published studies (Mitha et al., 2023). A scoping review of NLP applied in post-acute care nursing settings found that existing systems target a narrow range of risk outcomes, and many studies do not employ standardized terminologies (Scharp et al., 2024). Taken together, these gaps point to the absence of comprehensive, schema-grounded benchmarks for nursing observation extraction prior to MEDIQA-SYNUR dataset.

The emergence of large language models (LLMs) has introduced a new paradigm for clinical information extraction, enabling schema-guided structured output without task-specific fine-tuning. Multi-step LLM pipelines that combine an LLM extraction stage with rule-based post-processing have shown particular promise. (Wang et al., 2024) applied such an architecture to medical record entity extraction, using LLM-generated question templates followed by rule-based filtering to suppress extraction inaccuracies, demonstrating that decomposing the extraction task across specialised stages improves schema alignment and reduces spurious predictions. (Adam et al., 2025) further showed that targeted prompting strategies combined with post-hoc validation heuristics can substantially suppress hallucinated numeric extractions from clinical text, a finding that directly informs the design of our validation and precision filtering agent. A critical practical constraint for hospital deployment is data privacy: proprietary cloud-hosted LLMs require transmitting protected health information to external servers, raising significant legal and ethical barriers. (Wiest et al., 2024) demonstrated that a locally deployed Llama 2 pipeline achieves strong zero-shot clinical extraction performance, with sensitivity and specificity exceeding 90% for key features, while keeping all patient data on-premise. This motivates our selection of LLaMA-3.3 for local deployment within hospital infrastructure.

Within the MEDIQA shared task series, which has benchmarked clinical NLP systems annually since 2019 across tasks including question answering, dialogue-to-note summarization, and medical order extraction (Abacha et al., 2019, 2021; Bjarnadottir et al.; Abacha et al., 2023; Saeed, 2024; Corbeil et al., 2025b), top-performing systems have progressively shifted from fine-tuned encoder models toward generative LLM-based pipelines. The MEDIQA-SYNUR 2026 task is distinctive within this series in requiring verbatim evidence grounding and value normalization across 192 nursing-specific schema categories constraints that no prior shared task has combined (George Michalopoulos, 2026).

3. Materials and Methods

3.1. Dataset

We use the MEDIQA-SYNUR dataset released on Codabench, which consists of clinical nursing transcripts paired with structured annotations. The dataset includes 122 training records, 101 development records, and 199 test records.

The transcripts follow a conversational clinical style, introducing ambiguity and variability that make structured extraction challenging. All annotations are strictly grounded in the transcript text, and systems are required to extract only explicitly stated observations with verbatim evidence spans.

Additional dataset statistics and distribution analyses are provided in the Appendix.

3.2. Observation Schema

The MEDIQA-SYNUR task defines a comprehensive observation schema consisting of 192 distinct nursing observation categories, each identified by a unique concept ID. These categories are designed to capture a wide spectrum of nursing assessments, interventions, physiological measurements, and patient status descriptors commonly documented in clinical care settings. Each observation category is associated with a predefined value type, which constrains the allowable output format and guides system extraction. The schema includes four primary value types: `SINGLE_SELECT` (e.g., Skin turgor: {Tented, Normal}); `MULTI_SELECT` (e.g., Urinary symptoms: {urgency, difficulty urinating}); `NUMERIC` (e.g., Urine output, Oxygen saturation); and `STRING` (e.g., Pain description, Heart sounds).

The schema spans multiple clinical domains, including but not limited to cognitive and neurological status (e.g., Cognitive status, Orientation, Glasgow Coma Scale components), respiratory and cardiovascular assessments (e.g., Oxygen delivery device, Work of breathing, Heart rate), mobility and functional status (e.g., Mobility, Level of assistance), genitourinary and gastrointestinal observations (e.g., Urinary symptoms, Urine appearance, Bowel movement description), skin and wound assessments (e.g., Skin turgor, Edema, Pressure injury stage), as well as nursing interventions and safety-related checks (e.g., Intravenous therapy, Bed safety, Fall risk assessment).

In addition, the schema explicitly models measurement units as separate observation categories for several numeric concepts (e.g., Urine output unit, Oxygen saturation unit, Temperature unit). This design requires systems to independently extract both the numeric value and its corresponding unit when present in the transcript, increasing task granularity and precision.

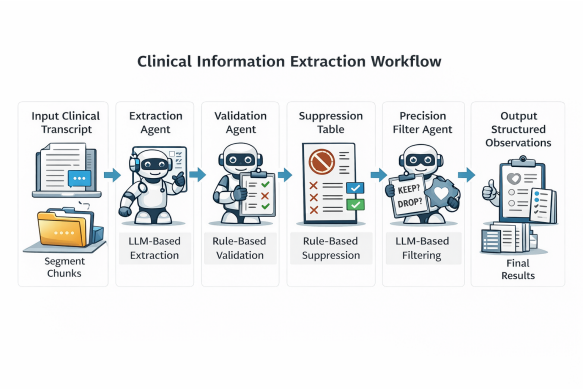

Figure 1: Overview of the proposed multi-agent clinical information extraction pipeline. The input clinical transcript is first segmented into smaller chunks for scalable processing. Each segment is processed by an LLM-based extraction agent to generate candidate observations, followed by a rule-based validation agent. A deterministic suppression table, derived from error analysis, is applied to reduce systematic false positives. A final LLM-based precision filter performs conservative KEEP/DROP decisions before aggregating results into structured observations.

3.3. Multi-Agent System Architecture

The proposed system follows a multi-stage, multi-agent architecture designed to extract structured nursing observations from free-text clinical transcripts while strictly enforcing evidence-based constraints. Each agent addresses a specific failure mode observed during development, including hallucinated extractions, implicit inference, and systematic false positives. The pipeline consists of four main components: an extraction agent, a validation agent, a precision filtering agent, and a deterministic suppression mechanism.

3.3.1. Overview of the Pipeline

Figure 1 illustrates the overall workflow of the proposed multi-agent system.

The system follows a sequential multi-stage design in which each component addresses a specific failure mode observed during development. Starting from transcript segmentation, each chunk is processed independently by the extraction agent, followed by validation, suppression, and precision filtering stages. The final outputs from all chunks are aggregated into structured observations.

Transcript Segmentation

Clinical transcripts in MEDIQA-SYNUR may be lengthy and information-dense. To reduce context overload and improve coverage, transcripts are optionally segmented into smaller text chunks based on character length (Corbeil et al., 2025c). Each segment is processed independently, allowing the extraction agent to focus on localized context and improving recall for observations mentioned sparsely or late in the transcript.

Schema Segmentation

A. Schema Segmentation and Batching (Static Schema Partitioning)

The MEDIQA-SYNUR observation schema contains 192 distinct categories, making it impractical to present the full schema to the model in a single prompt. To handle this, we implement a static batching strategy, where the schema categories are divided into fixed batches. Each batch is processed independently by the extraction agent, and only the subset of schema definitions belonging to that batch is included in the prompt.

B. Schema Retrieval via Embeddings (Dynamic Top-K Selection)

Each schema category is embedded once using a text-embedding model. At inference time, the transcript is also embedded, and cosine similarity is used to retrieve the top-K most relevant schema categories. Only these retrieved definitions are injected into the prompt for extraction.

This dynamic retrieval mechanism reduces prompt length, focuses the model on clinically relevant categories, and adapts naturally to transcript variability. However, experimental results showed that this retrieval-based strategy provided limited performance gains compared to full schema batching. Therefore, it was not included in the final system configuration.

3.3.2. Extractor Agent

The Extractor Agent is responsible for the initial identification of candidate nursing observations from the transcript using a large language model (LLM). To manage the large schema size (192 observation categories), the schema is dynamically chunked into smaller batches, and each batch is processed independently against segmented transcript chunks when segmentation is enabled. The extraction prompt explicitly enforces strict evidence rules. Each extracted observation must include a valid schema identifier, a value consistent with the schema type (numeric, single-select, multi-select, or free text), and an evidence field that is an exact verbatim substring copied from the transcript. The prompt explicitly prohibits inference, speculation, and abstraction. Negative values (e.g., “No”, “Absent”) are only allowed when explicit negation cues appear in the transcript. The output of the extraction agent is intentionally permissive, prioritizing recall while deferring strict correctness checks to downstream agents.

3.3.3. Validation Agent

The Validator Agent applies deterministic, rule-based validation to the raw outputs produced by the extraction agent. Its primary role is to remove structurally invalid, weakly grounded, or speculative observations before further refinement. Validation rules include rejecting observations whose evidence does not appear verbatim in the transcript, filtering out evidence containing hedging or speculative language (e.g., "possibly", "suggests", "likely"), enforcing explicit negation rules for negative values, applying schema-specific anchor constraints for error-prone categories (e.g., vomiting, pain, Glasgow Coma Scale), requiring the presence of domain-specific lexical cues. ensuring values conform exactly to the allowed schema enumeration when applicable. This agent significantly reduces hallucinated and semantically invalid predictions while preserving correct extractions supported by explicit textual evidence.

3.3.4. Filtering Agent

Despite strict validation, development analysis revealed that certain observations remained over-predicted due to subtle semantic drift or overly permissive extraction. To address this, a Precision oriented Filter Agent is applied as an additional refinement stage. The precision filter operates at the level of individual observations. For each validated observation, the agent re-evaluates whether the extraction should be retained or discarded by considering the full transcript context, the schema definition and allowed values, and the extracted evidence span and value. The agent produces a binary decision (KEEP or DROP) for each observation, along with a brief rationale. The precision filter is intentionally conservative and is configured with deterministic decoding settings (temperature = 0.0) to ensure reproducibility. Its role is not to introduce new information, but to eliminate residual false positives that survive earlier validation stages.

Prompt Design Across Agents The prompting strategy differs across agents but follows a consistent evidence-grounded design. The extractor agent employs a schema-aware prompt that enforces strict constraints, requiring verbatim evidence spans and explicitly prohibiting inference or speculative reasoning. The precision filter operates using a binary decision prompt, where each candidate observation is evaluated independently and labeled as KEEP or DROP based on the schema definition, extracted value, evidence span, and transcript context. In contrast, the validation and suppression stages are fully rule-based and do not rely on prompting.

3.3.5. Conservative Suppression Table

In addition to agent-based filtering, a lightweight deterministic suppression mechanism is applied. The suppression table is implemented as a rule-based post-processing step derived from systematic error analysis on the development set. It is constructed by identifying observation categories that were consistently over-predicted, semantically redundant, or weakly grounded in the input text.

The suppression rules are divided into two groups: (i) an always-suppressed set, where specific observation categories are removed regardless of context due to persistent false-positive behavior (e.g., high-level or abstract clinical concepts), and (ii) a negation-sensitive set, where negative values (e.g., "No", "Absent") are retained only if explicit negation cues (such as "no", "denies", "without", or "absent") are present in the evidence span.

This mechanism is intentionally conservative and does not introduce any new predictions; it only removes unreliable ones. It is applied uniformly across development and test data without any test-time tuning, ensuring consistent behavior and reducing systematic false positives while minimizing the risk of discarding valid observations.

3.4. Model Selection

For all experiments, we selected LLaMA-3.3, a 70B-parameter decoder-only transformer model, as the core language model in our multi-agent pipeline. This choice was informed by results from our prior study, where the model demonstrated strong instruction-following abilities and robust performance on structured information extraction tasks. Importantly, we used the model strictly in inference-only mode, with no fine-tuning or task-specific adaptation on the MEDIQA-SYNUR dataset.

3.5. Implementation Details

All experiments were conducted using LLaMA-3.3 via the Ollama framework, enabling fully local inference to satisfy clinical data privacy constraints. In early development stages, we also evaluated DeepSeek as the underlying language model; however, due to unstable performance and significantly lower recall, all subsequent experiments were carried out using LLaMA-3.3. The extraction pipeline follows a multi-agent design, where transcript segmentation is applied to divide long clinical notes into manageable chunks. Each segment is processed independently by the extraction agent, which iteratively extracts candidate observations across schema batches.

The prompting strategy differs across agents but follows a consistent evidence-grounded design. The extractor agent uses a schema-aware prompt

that includes a subset of observation categories and instructs the model to output only observations explicitly supported by verbatim text spans from the transcript. The prompt enforces strict constraints, including prohibiting implicit inference, speculative reasoning, and requiring exact substring evidence. Negative values are only allowed when explicit negation cues are present in the transcript.

The validation stage is fully rule-based and does not rely on prompting. It applies deterministic checks to ensure that each extracted observation is supported by exact evidence in the transcript, does not contain hedging language, satisfies negation constraints, and conforms to schema-specific requirements, including value types and anchor-based rules for selected concepts.

The precision filtering agent operates as a final refinement step and uses a separate prompt to re-evaluate each observation individually. Given the observation, schema definition, and full transcript, the model produces a strict binary decision (KEEP or DROP). This stage is configured with deterministic decoding (temperature = 0.0) to ensure reproducibility and is designed to remove residual false positives that pass earlier validation stages.

Schema handling is performed through static schema segmentation, where the full set of observation categories is divided into fixed batches, and each batch is processed independently. This approach enables scalable processing of the full schema without exceeding prompt limitations while maintaining full coverage across all observation categories.

In addition to agent-based filtering, a deterministic suppression mechanism is applied as a lightweight post-processing step. The suppression table is derived from systematic error analysis on the development set and includes two categories: (i) always-suppressed observations that were consistently over-predicted or semantically redundant, and (ii) negation-sensitive observations, where negative values are retained only if explicit negation cues are present in the evidence.

To assess the contribution of each component in the extraction pipeline, we conducted a series of controlled experiments by selectively enabling or disabling key modules, including transcript segmentation (for handling long transcripts), the suppression table (to reduce systematic false positives), and the precision agent (which applies an additional LLM-based refinement step).

In addition, we explored alternative strategies, such as embedding-based schema matching, where schema items are aligned with transcript representations to guide extraction under varying relevance thresholds. However, these approaches were evaluated only as part of the experimental analysis and were not included in the final system configuration. This systematic evaluation allows us to quantify how each module, both individually and in combination, impacts overall extraction performance.

4. Error Analysis

To better understand the strengths and limitations of the proposed multi-agent clinical information extraction system, we conducted a comprehensive error analysis on the development set. Since ground-truth annotations are hidden during the test phase, all qualitative and quantitative analyses were performed using false positive (FP) and false negative (FN) predictions observed on the development data. The objective of this analysis is not only to quantify errors, but also to identify systematic failure modes and assess how different components of the pipeline contribute to precision–recall trade-offs.

4.1. False Positive Analysis

Figure 2 presents the top 30 observation categories most frequently associated with false positive predictions. Several high-frequency false positives correspond to semantically broad or overlapping clinical concepts, including gastrointestinal symptoms, gait and transferring, urinary symptoms, oxygen intake, and dyspnea. These categories often appear in narrative nursing documentation in descriptive or contextual forms, which increases the likelihood of over-extraction even under strict evidence constraints.

In many cases, false positives arose when: (i) the model extracted implicit or contextual mentions rather than explicitly stated observations; (ii) multiple related concepts were mentioned in proximity (e.g., mobility, transfer technique, gait), leading to semantic boundary confusion; and (iii) measurement-related concepts (e.g., pulse oximetry unit, MAP unit) were mentioned as part of documentation templates rather than clinical findings.

These findings motivated the introduction of rule-based suppression tables and a precision filter agent, specifically targeting categories with historically high FP rates.

4.2. False Negative Analysis

Figure 3 shows the top 30 observation categories most frequently missed by the system (false negatives). False negatives were most prominent in categories such as orientation, secondary diagnosis, cognitive status, skin condition, and dyspnea. These concepts often exhibit: (i) high lexical variability (e.g., multiple ways of describing cognitive or mental status); (ii) distributed mentions across the transcript rather than a single explicit phrase; and (iii) subtle negation or partial descriptions that are

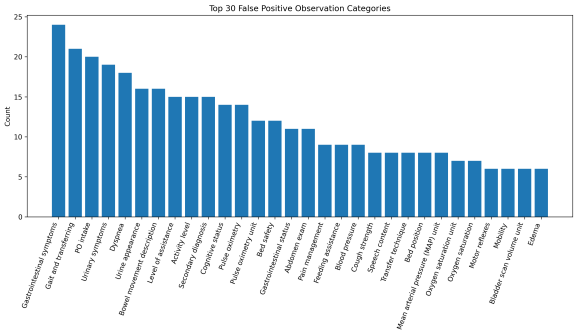

Figure 2: Top 30 observation categories most frequently associated with false positive predictions in the development set.

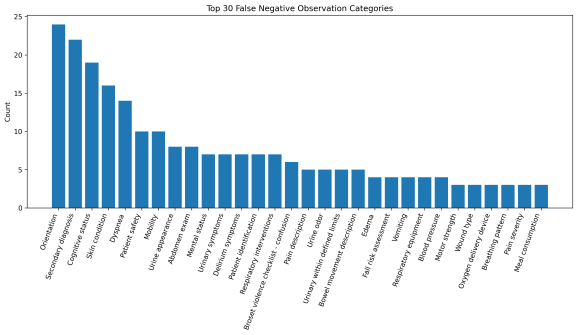

Figure 3: Top 30 observation categories most frequently associated with false negative predictions in the development set.

difficult to capture under strict no-inference rules. Additionally, some false negatives resulted from conservative validation and suppression strategies, where ambiguous or weakly anchored extractions were intentionally discarded to preserve precision.

4.3. Error Cause Distribution

To further characterize system behavior, we grouped all FP and FN instances into high-level error cause categories, summarized in Figure 4. The dominant error patterns include:

(i) **Other FP causes**: the largest share of false positives. This category captures residual errors not covered by predefined rule-based categories, and primarily arises from three sources: (a) *semantic ambiguity*, where phrases partially match a schema concept but lack sufficient specificity (e.g., vague mentions of discomfort interpreted as pain); (b) *template-driven artifacts*, where structured clinical templates or repeated phrasing trigger spurious matches without true clinical evidence; and (c) *borderline evidence spans*, where extracted text technically satisfies verbatim matching constraints but does not strongly support the assigned observation value. These errors reflect the limitations of strict substring-based evidence matching when applied

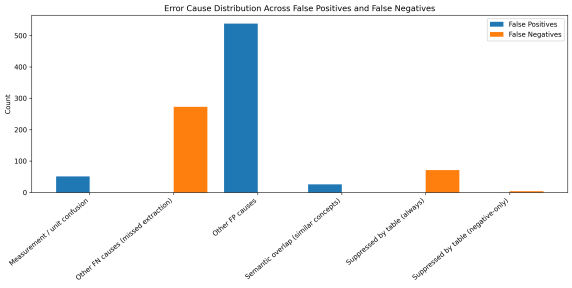

Figure 4: Distribution of high-level error causes for false positive and false negative predictions, aggregated over all development records.

to loosely structured clinical language.

(ii) **Other FN causes (missed extraction)**: the most common false negative category, typically due to strict evidence requirements or insufficient lexical anchoring;

(iii) **Measurement / unit confusion**: errors where numeric values or units were present but not clearly attributable to a specific observation category;

(iv) **Semantic overlap (similar concepts)**: errors arising from closely related schema categories (e.g., mental status vs. orientation vs. delirium symptoms); and

(v) **Suppression-related effects**: a smaller subset of false negatives attributable to conservative suppression rules, particularly for negative-only observations.

Importantly, these trends reflect an intentional precision-oriented design trade-off, where the system favors high-confidence, explicitly grounded extractions over recall in ambiguous cases.

5. Results

All results reported in this section were obtained before the official competition deadline No additional post-deadline experiments were conducted.

5.1. Development Set Exploration

All development-set experiments were conducted through extensive exploratory trial-and-error to better understand the behavior of the extraction pipeline. Our initial trials employed DeepSeek as the underlying model. Although the precision-filtered variant produced a high precision of 0.6984, its recall was extremely low (0.0633), resulting in an F1 score of only 0.1161. Efforts to improve recall by injecting internal evidence into the extraction process further degraded performance, with recall dropping to 0.0115 and the F1 score to 0.0224. These findings indicated that DeepSeek could not generalize well to this task, particularly in retrieving

relevant information from long and detailed clinical transcripts. Consequently, DeepSeek was excluded early in development, and all subsequent experiments were carried out using LLaMA 3.3.

Before settling on a stable configuration, we also examined a rule-based evidence–augmented pipeline, which achieved a very high recall of 0.7978 but substantially reduced precision, yielding an overall F1 score of 0.6039. With LLaMA 3.3 as the foundation, system refinement became significantly more effective. Incorporating a disambiguation agent alongside the precision filter and suppression table produced a notable improvement, reaching a precision of 0.6268, a recall of 0.7201, and an F1 score of 0.6702. This phase highlighted the importance of resolving ambiguous or overlapping observations in the transcripts. Further ablation studies demonstrated that both the suppression table derived from systematic error analysis, and the precision-filtering step were essential to maintaining balanced performance. When either component was removed, the system's F1 score dropped sharply to 0.1902, mainly due to a substantial decline in recall.

The decisive improvement occurred when transcript segmentation was integrated with the precision agent and the suppression table within the LLaMA 3.3 pipeline. This configuration yielded the strongest development-set performance, with a precision of 0.6427, a recall of 0.7518, and an F1 score of 0.6930. As illustrated in Figure 5, this setting demonstrated the most stable and balanced performance among all pre-deadline experiments and was therefore adopted as the final configuration for the system.

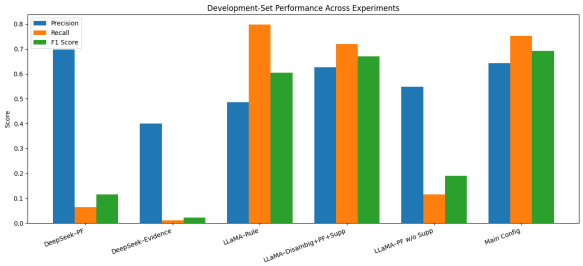

Figure 5: Development-set performance across major experimental configurations. "DeepSeek–PF" and "DeepSeek–Evidence" represent early baselines using the DeepSeek model with precision filtering or evidence-enriched prompting. "LLaMA–Rule" denotes rule-based evidence extraction with LLaMA 3.3. "LLaMA–Disambig+PF+Supp" includes the disambiguation agent, precision filtering, and the suppression table. "LLaMA–PF w/o Supp" shows the effect of disabling the suppression table. "Main Config" represents the final development-set configuration selected for test-set evaluation.

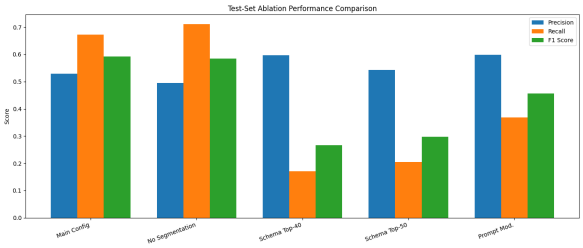

Figure 6: Test-set performance of the main configuration compared with ablation variants. "Main Config" refers to the full system using transcript segmentation, the precision agent, and the suppression table. "No Segmentation" removes transcript segmentation while retaining the precision agent. "Schema Top-40" and "Schema Top-50" restrict extraction to the top schema-matched concepts using embedding similarity. "Prompt Mod." applies prompting adjustments without segmentation or the suppression table. The main configuration achieves the best overall F1 score on the test set.

5.2. Test Set Results

The system that achieved the highest score on the development set was carried forward and evaluated directly on the test set. This configuration, combining transcript segmentation, the precision agent, and the suppression table, served as our primary setup. On the test set, it achieved precision = 0.5292, recall = 0.6725, and F1 = 0.5923 (ref_obs = 2721; hyp_obs = 3458), confirming its strong and balanced performance across metrics.

After establishing this main system, we conducted a series of additional experiments in an effort to further improve extraction quality. These included removing segmentation, enabling or disabling the precision agent, experimenting with the suppression table, and introducing schema-guided extraction using embedding similarity with different schema subset sizes (top-40, top-50, and top-70 items).

However, none of these enhancement attempts outperformed the main configuration. As shown in Figure 6, disabling segmentation while keeping the precision agent active resulted in a slightly lower F1 score of 0.5838. Schema-based variants, despite achieving high precision in some cases, suffered substantial recall losses, leading to significantly lower F1 scores (e.g., 0.2662 with top-40 items and 0.2969 with top-50 items). Even the broader schema variant with extraction prompt modification, tested without segmentation, with precision agent, without suppression table, reached only F1 = 0.4560.

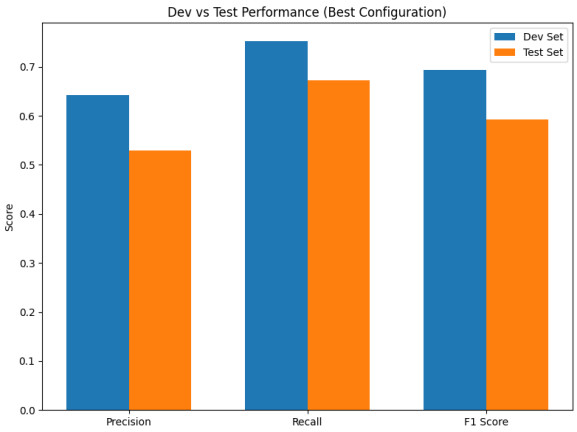

Figure 7: Comparison of development-set and test-set performance using the best configuration of the proposed system. The model integrates transcript segmentation, the precision agent, and the suppression table, which together achieved the highest score during development. Performance is shown across Precision, Recall, and F1 Score. While the test set shows a consistent drop across all metrics, as expected due to dataset shift, the configuration still maintains strong generalization, confirming it as the optimal setting for final evaluation.

5.3. Main Results: Best Configuration

After extensive exploration, the final configuration submitted to the competition consisted of Transcript segmentation (enabled), Precision agent (enabled), and Suppression table (enabled and derived from dev-set analysis). This setup achieved strong performance on the development set, with an F1 score of 0.6930 (precision = 0.6427, recall = 0.7518). The same configuration was then applied directly to the test set, without any further changes, resulting in an F1 score of 0.5923 (precision = 0.5292, recall = 0.6725).

As shown in Figure 7, these results reflect the final system state as it existed before the competition deadline, and they represent the most effective balance of precision and recall identified through our iterative exploration.

6. Discussion

The experiments conducted prior to the competition deadline highlight the core strengths and limitations of the proposed extraction pipeline. Overall, the system demonstrated strong recall, driven largely by effective transcript segmentation and the combined influence of the precision agent and suppression table. Segmentation proved especially valuable for isolating clinically dense sections of text, allowing the model to capture relevant observations consistently and with fewer omissions. However, the pipeline also exhibited persistent precision challenges due to the generative nature of large language models. Despite explicit instructions to avoid inference, the model sometimes produced clinically plausible but unsupported observations, a form of generative drift. This tendency to "over-interpret" clinical cues resulted in hallucinated predictions such as inferred nutrition status or IV fluid details that were not present in the transcript. The example analysis further illustrated how the model could accurately capture ground-truth observations while simultaneously introducing additional, schema-shaped predictions without clear evidence. To counteract these issues, the precision agent and suppression table played a crucial role. The precision agent filtered weakly supported predictions, while the suppression table blocked patterns of recurrent hallucinations identified during error analysis. Ablation results showed that removing either component substantially increased false positives, confirming their necessity for maintaining competitive precision and overall F1 performance. In summary, the pipeline's success stemmed from its ability to balance high recall with post-processing mechanisms that restrained the generative biases of LLMs. Nonetheless, the observed drift underscores ongoing challenges in evidence-grounded extraction and highlights areas for further refinement.

7. Conclusion

In this work, we presented a multi-agent LLM-based system for structured nursing observation extraction from clinical transcripts, addressing the challenges of evidence grounding and value normalization. By decomposing the extraction process into specialized agents and iteratively refining the pipeline through systematic error analysis, our system achieved a strong balance of precision and recall, with an F1 score of 0.6930 on the development set and 0.5923 on the test set. Despite these results, limitations remain, particularly the reliance on hand-crafted suppression tables and transcript-specific processing, which can reduce robustness across domains. Future work will explore fine-tuning the underlying LLM on synthetic or multi-institutional clinical transcripts, improving model-level controllability, and incorporating human-in-the-loop validation to create a more reliable and generalizable system. Overall, our findings demonstrate that a multi-agent LLM approach can effectively structure free-text clinical documentation, providing a foundation for more automated and evidence-grounded clinical data extraction.

8. Limitations

Despite the strong performance of the proposed system, several limitations must be acknowledged. First, although prompting strategies and suppression mechanisms were designed to strictly constrain the model to evidence-based extraction, the underlying LLM still exhibited generative tendencies. This occasionally led to clinically plausible but unsupported predictions, lowering precision, particularly for schema fields that were semantically broad or frequently referenced in the transcripts. Notably, the suppression table, crafted based on the development set, was also used during testing; this resulted in reduced performance, highlighting the sensitivity of static, hand-crafted controls. Second, the system relies heavily on transcript segmentation, precision filtering, and the manually derived suppression table. While effective, these components require substantial engineering effort and are sensitive to variations in transcript style or domain-specific phrasing.

9. Bibliographical References

Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th workshop on biomedical language processing*, pages 74–85.

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 18th bioNLP workshop and shared task*, pages 370–379.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 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*, pages 503–513.

Hammaad Adam, Junjing Lin, Jianchang Lin, Hillary Keenan, Ashia Wilson, and Marzyeh Ghassemi. 2025. Clinical information extraction with large language models: A case study on organ procurement.

Ragnhildur I Bjarnadottir, Walter O Bockting, Sunmoo Yoon, and Dawn W Dowding. Computers, informatics, nursing nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study-manuscript draft-manuscript number: Cin-d-17-00098r1 full title: Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study powered by editorial manager® and produxion manager® from aries systems corporation. Technical report.

Ragnhildur I. Bjarnadottir and Robert J. Lucero. 2018. What can we learn about fall risk factors from ehr nursing notes? a text mining study. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 6:21.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, et al. 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*, pages 859–870.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025b. 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*, pages 11–16.

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. 2025c. 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.

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.

Adi V. Gundlapalli, Guy Divita, Andrew Redd, Marjorie E. Carter, Danette Ko, Michael Rubin, Matthew Samore, Judith Strymish, Sarah Krein,

Kalpana Gupta, Anne Sales, and Barbara W. Trautner. 2017. Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. *Journal of Biomedical Informatics*, 71:S39–S45.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review.

Kexin Huang, Tamryn F. Gray, Santiago Romero-Brufau, James A. Tulsky, and Charlotta Lindvall. 2021. Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study. *Journal of the American Medical Informatics Association*, 28:1660–1666.

Sookyung Hyun, Stephen B. Johnson, and Suzanne Bakken. 2009. Exploring the ability of natural language processing to extract data from nursing narratives. *CIN - Computers Informatics Nursing*, 27:215–223.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

Shazia Mitha, Jessica Schwartz, Mollie Hobensack, Kenrick Cato, Kyungmi Woo, Arlene Smaldone, and Maxim Topaz. 2023. Natural language processing of nursing notes: An integrative review.

Nadia Saeed. 2024. Medifact at mediqa-m3g 2024: Medical question answering in dermatology with multimodal learning. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 339–345.

Danielle Scharp, Mollie Hobensack, Anahita Davoudi, and Maxim Topaz. 2024. Natural language processing applied to clinical documentation in post-acute care settings: A scoping review.

Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, and Yuxiang Li. 2024. An entity extraction pipeline for medical text records using large language models: Analytical study. *Journal of Medical Internet Research*, 26.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 7.

A. Additional Dataset Analysis

This study uses the dataset released as part of the MEDIQA-SYNUR shared task, hosted on the Codabench evaluation platform. The task focuses on extracting structured nursing observations from unstructured clinical transcripts under strict schema and evidence constraints. Each record consists of a free-text nursing transcript paired with gold-standard structured annotations. The dataset is divided into training, development, and test splits. The training split contains 122 patient records, while the development split contains 101 patient records. The test split contains 199 patient records. Analysis of concept distributions across the training and development splits shows consistent usage patterns, suggesting that the data is reasonably balanced at the split level despite intrinsic variability across clinical observations. The transcripts follow a conversational clinical style and frequently contain hesitations, discourse markers, and loosely structured descriptions, characteristics also noted in the original dataset publication. This introduces realistic challenges for automated extraction systems, including ambiguous phrasing, overlapping concepts, and implicit contextual cues.

Figures 8 and 9 illustrate the top 30 most frequent nursing observation categories in the training and development sets, respectively. Across both splits, commonly annotated observations include cognitive status, mobility, oxygen delivery device, oxygen saturation, urinary symptoms, and skin turgor. In contrast, many observation categories occur infrequently, reflecting the broad range of clinical scenarios and variability in nursing documentation captured by the dataset. The dataset exhibits a pronounced long-tailed distribution, in which a small subset of observation categories accounts for a

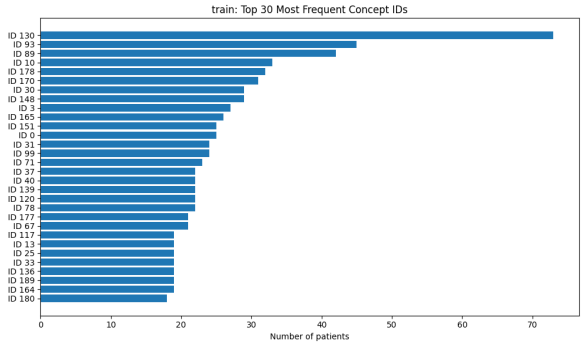

Figure 8: Top 30 most frequent observation concept IDs in the training set. The distribution highlights a long-tailed pattern, where a small number of observation categories occur frequently across patients, while many schema concepts appear sparsely.

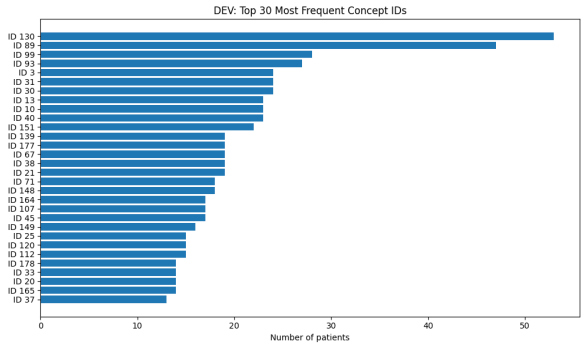

Figure 9: Top 30 most frequent observation concept IDs in the development set. The distribution highlights a long-tailed pattern, where a small number of observation categories occur frequently across patients, while many schema concepts appear sparsely.

large proportion of annotations, while a substantial number of categories appear only sparsely. This distribution poses two key challenges for automatic extraction systems. First, achieving high recall for rare observation categories is difficult due to limited training signals. Second, maintaining high precision for frequently occurring but semantically broad categories is challenging, as such categories often encompass overlapping or closely related clinical concepts. In addition, semantic overlap among related observations, such as cognitive status, orientation, and delirium symptoms, further increases task complexity and contributes to potential ambiguity during extraction. Importantly, all gold-standard annotations are strictly grounded in the transcript text. The task requires systems to extract only observations that are explicitly stated and to provide verbatim evidence spans from the transcript, without relying on inference or clinical interpretation beyond the documented text.