MasonNLP at MEDIQA-SYNUR 2026: Retrieval-Augmented Large Language Models for Schema-Constrained Clinical Information Extraction

A H M Rezaul Karim, Özlem Uzuner
George Mason University, VA, USA
akarim9@gmu.edu, ouzuner@gmu.edu

Abstract

Conversational nurse-patient transcripts contain actionable observations, but converting these transcripts into structured representations at scale remains challenging. Documentation burden is substantial, with prior studies showing clinicians spend large portions of their workday on documentation and related desk work rather than direct patient care. MEDIQA-SYNUR focuses on observation extraction from conversational nurse-patient transcripts, requiring systems to normalize these narratives into a predefined schema with value-type constraints. We propose a modular retrieval-augmented generation (RAG) pipeline that uses the training set as an exemplar corpus, combines schema-constrained prompting (full schema vs. pruned candidate schema), deterministic schema-based postprocessing, and a second-pass audit, with two LLM backbones: *Llama-4-Scout-17B-16E-Instruct* and *GPT-5.2* with corresponding embedding models for RAG. Our best configuration uses *GPT-5.2* with full schema, RAG, and a second-pass auditing, achieving 80.36% F_1 score. Overall, our results show that RAG consistently improves performance, while the optimal degree of schema constraint depends on the model, and second-pass auditing yields modest additional gains by correcting residual schema-adherence errors.

Keywords: Retrieval Augmented Generation (RAG), Clinical Information Extraction, LLM

1. Introduction

Clinical documentation is essential for continuity of care and quality reporting, yet it remains a major source of clinician workload and professional dissatisfaction (Muhiyaddin et al., 2022). A study in ambulatory practice has shown that physicians spend nearly half of their clinic day on electronic health record (EHR) and desk work (49.2%) and substantially less time in direct clinical face time (27.0%), with additional after-hours work of approximately 1-2 hours per night devoted mostly to EHR tasks (Sinsky et al., 2016). Similar EHR log-based analyses report that primary care clinicians spend 145.9 minutes/day actively using the EHR (Rotenstein et al., 2022). These high demands for documentation motivate methods that can automatically capture and convert clinical information into structured representations with minimal manual entry. As a step in this direction, particularly for workflows such as nursing assessments, clinical documentation is frequently transcribed (Mayer et al., 2022). Conversational nurse-patient transcripts are examples of such documentation.

Information extraction (IE) can convert transcripts into structured representations (Balasubramanian et al., 2025; Hu et al., 2026). In many clinical IE settings, the goal is to identify spans, assign concept labels, or extract relations from clinical text, often within relatively limited label spaces and without requiring the model to generate a fully normalized structured output. Strict requirements on output, such as conforming to a large predefined schema with consistent formatting and validity constraints (Karim and Uzuner, 2025; Builtjes et al., 2025) pose additional challenges for clinical IE.

MEDIQA-SYNUR task presents an observation extraction task that encompasses these challenges, with a specific focus on conversational nurse-patient transcripts. Given a transcript, MEDIQA-SYNUR requires automated systems to extract clinically salient observations normalized to a predefined schema with explicit value-type requirements (e.g., numeric vs. categorical values and enumerated option validity) (Michalopoulos et al., 2026). The *schema* for the task captures the complete set of target observation concepts together with their value types and any allowable categorical value sets. The SYNUR dataset accompanying this task provides an open-source synthetic corpus annotated by expert nurses with a large set of structured observations, enabling controlled evaluation of observation extraction from conversational nurse-patient transcripts (Corbeil et al., 2025).

Recent advances in instruction-following large language models (LLMs) have enabled prompt-based clinical IE for converting clinical narratives into structured representations (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). Compared with task-specific supervised systems, LLMs offer a more flexible framework, and prior work has shown that, with carefully designed prompts, they can support few-shot extraction across clinical IE tasks (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). However, reliability and schema adherence remain persistent concerns, particularly

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 153–162

as output spaces grow. LLMs tend to hallucinate concept names, violate type constraints, and produce inconsistent formatting when schema complexity increases (Karim and Uzuner, 2025; Builtjes et al., 2025). Evidence also suggests that open-weight and closed-weight models exhibit meaningfully different instruction-following behaviors and robustness profiles under constraint-heavy prompts (Builtjes et al., 2025). In this work, we explore open-weight and closed-weight LLMs in observation extraction by constraining their output through a schema that varies in its complexity. We refer to this as schema-constrained observation extraction.

We hypothesize that retrieval-augmented generation (RAG) (Lewis et al., 2020) can improve performance in observation extraction by conditioning output on retrieved exemplars (Shlyk et al., 2024; Lopez et al., 2025; Zhan et al., 2025; Liu et al., 2025). To test this hypothesis, we study two RAG (Lewis et al., 2020) approaches that use the training set as the retrieval corpus. For *Llama-4-Scout-17B-16E-Instruct* (Meta, 2025), we find that constraining the output space using a pruned candidate schema improves results. In contrast, with *GPT-5.2* (OpenAI, 2025), RAG yields the best performance with full schema, and a second-pass auditing provides additional gains.[1]

Our study makes the following contributions:

- The question of the interaction of RAG with the schema is in Clinical IE. Our work directly addresses this gap by studying how RAG interacts with different schema constraints across LLM backbones of differing parameter scales, providing a systematic analysis of this interaction in observation extraction.
- We show that a pruned candidate schema is beneficial for a smaller open-weight model but counterproductive for a larger model, suggesting that the formulation of the output constraints, namely how the fixed task schema is presented to the model during generation, should be model-aware, i.e., adapted to the behavior and capabilities of the underlying model.
- We present second-pass auditing that provides modest gains primarily by correcting schema-adherence and normalization errors, but does not substitute for RAG and schema constraint, indicating it is best used as a final refinement stage rather than a core component.

Overall, our findings provide practical evidence that schema-constrained observation extraction should be model-aware. Our results clarify when a pruned candidate schema is helpful versus when a full schema is preferable, and further show how RAG and second-pass auditing can be combined to improve robustness.

[1] Implementation details can be found here: https://github.com/AHMRezaul/MEDIQA-SYNUR-2026

2. Related Work

Clinical information extraction (IE) has been driven by tasks that standardize evaluation for concept extraction, assertion detection, and relations (Uzuner et al., 2011; Henry et al., 2020; Fu et al., 2020; Navarro et al., 2023). Benchmarks such as i2b2/VA and n2c2 established protocols for span extraction and normalization, but largely focus on relatively small label spaces (Uzuner et al., 2011; Henry et al., 2020; Mahajan et al., 2023). In contrast, MEDIQA-SYNUR focuses on a large, heterogeneous output space (193 typed observation concepts), which must be extracted under strict type and enumeration constraints, motivating methods that explicitly condition generation on schema.

Early studies of conversational nurse-patient transcripts coupled speech recognition with IE to produce structured handover documents, highlighting both feasibility and the difficulty of extracting structured representations from transcripts (Johnson et al., 2014a,b; Dawson et al., 2014). Subsequent shared evaluations released transcribed handover datasets and structured annotations, often using synthetic data to mitigate privacy barriers (Suominen et al., 2015b,a, 2016). Building on this line, Corbeil et al. (2025) introduced SYNUR to support systematic evaluation of structured output from conversational nurse-patient transcripts, leaving open how to enforce schema adherence reliably at scale.

Within LLM-based clinical IE, most prior work has emphasized direct prompting or few-shot extraction, often in settings where outputs are interpreted more flexibly and are not tightly constrained by a large predefined schema (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). Work that explicitly examines schema adherence has shown that structured extraction becomes substantially more difficult as output spaces grow and validity constraints become stricter (Karim and Uzuner, 2025; Builtjes et al., 2025). Although retrieval-augmented generation has begun to show promise in clinical extraction by grounding predictions in relevant examples or evidence (Lewis et al., 2020; Shlyk et al., 2024; Lopez et al., 2025; Zhan et al., 2025; Liu et al., 2025), its interaction with schema-constrained generation remains underexplored. Likewise, limited prior work has compared how open-weight and closed-weight LLMs behave under such constraint-heavy extraction settings. Our work addresses these gaps in the context of schema-constrained observation extraction.

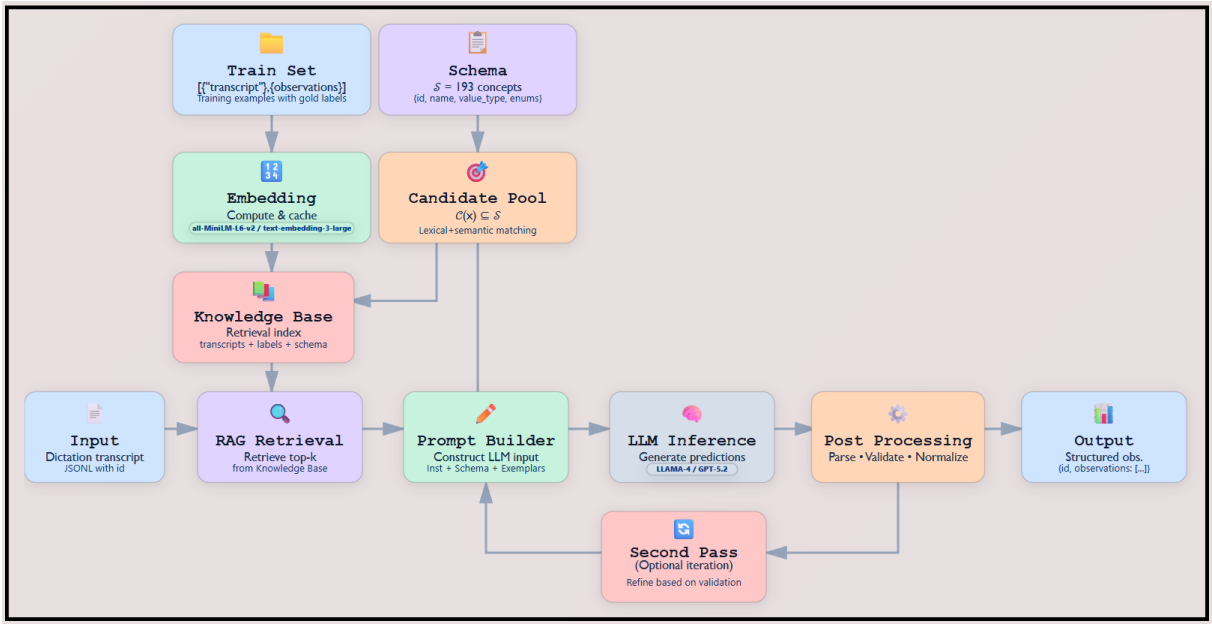

Figure 1: Retrieval-augmented, schema-constrained pipeline that retrieves training exemplars, conditions the LLM on schema (full or pruned candidate), and post-processes outputs with second-pass auditing.

3. Task Description

3.1. Problem Formulation

MEDIQA-SYNUR defines the task as an *observation extraction* from conversational nurse-patient transcripts. Given a transcript x, the goal is to identify clinically salient observations and normalize them to a predefined schema.

The schema is a set of $M = 193$ observation concepts $\mathcal{S} = \{c_1, \ldots, c_M\}$. Each concept c_m has an identifier id_m, a name N_m, a value type $\tau_m \in \{\textsc{single\_select}, \textsc{multi\_select}, \textsc{numeric}, \textsc{string}\}$, and, for categorical types, an allowed value set $\mathcal{V}_m$.

For an input x, a system outputs a list of extracted observation instances $\hat{O} = [o_1, \ldots, o_n]$, where each instance is an object from the schema $o_i = \{\hat{\mathrm{id}}_i, \hat{N}_i, \hat{\tau}_i, \hat{v}_i\}$. The evaluation measures the correctness of the predicted observations under this schema.

3.2. Dataset

The SYNUR dataset (SYnthetic NURsing) (Corbeil et al., 2025) is a JSONL file split into `train`, `dev`, and `test`. Table 1 contains the dataset statistics of each split. Each instance contains a unique case identifier `id` and a free-text `transcript`. It also contains ground truth `observations`, represented as a list of objects from the schema with values that may be categorical (single- or multi-select), numeric, or free-text, reflecting the heterogeneity of routine nursing documentation.

Split	No. of Inst.	Total Obs.	Unique Obs.
Train	122	1685	166
Dev	101	1315	170
Test	199	2552	164

Table 1: Dataset statistics per split with the number of instances, total number of observations, and the number of unique observations in each split.

Across all splits, conversational nurse-patient transcripts are moderately long, averaging 192 words, with lengths ranging from 59 to 343 words. In the train and development sets, each instance contains 13.45 observations on average, ranging from 6 to 34 observations per case, with no duplicated concept IDs within an instance. Label frequency follows a long-tailed distribution, while a small set of observations occur very frequently (e.g., *Cognitive status*, *Mobility*, *Oxygen saturation*, *Nausea*), which are normally the most common observations made by nurses. Many concepts are sparse, with 55 concepts appearing at most five times and 23 concepts appearing only once or twice in the labeled data (e.g., *Drain output*, *Prosthetic use*, *Foreign object removal*).

The provided schema defines 193 unique observation concepts with explicit value types and, for categorical concepts, enumerated allowable values. The schema is dominated by categorical fields (130 SINGLE_SELECT and 12 MULTI_SELECT), with additional NUMERIC (20) and STRING (31) concepts. Categorical concepts have relatively small label sets on average, but include larger option lists for clinically

richer fields such as *Bowel movement description* (12 options) and *Pain severity* (11 options). The schema also includes 15 explicit "unit" concepts (e.g., oxygen saturation unit, respiration unit), representing measurements as value–unit pairs. In the labeled data, 178 out of the 193 schema concepts appear at least once, which supports broad coverage while preserving a realistic distribution.

Overall, this dataset is well-suited for evaluating schema-constrained extraction because it combines (i) conversational nurse-patient transcripts of varying length, (ii) a large and heterogeneous schema with strict type and enumeration requirements, including unit pairing, and (iii) a naturally imbalanced concept distribution that makes rare observations harder to capture while maintaining strong schema-adherence.

4. Methodology

4.1. Overview

Our approach is implemented as a modular workflow in which each component can be enabled or disabled, allowing controlled comparisons across design choices and LLM backbones. Figure 1 provides a complete illustration of our system. Given an input transcript x and the schema $\mathcal{S}$, the system constructs a prompt that conditions the LLM on (i) task instructions enforcing strict structured output, (ii) schema (full or pruned schema), and (iii) retrieved in-context exemplars when RAG is enabled. The LLM is instructed to output only a JSON list of (id, v) pairs (concept identifier and value), which we then deterministically expand to the required submission format by looking up each concept's canonical `name` and `value_type` from the schema; this avoids errors from hallucinated names or types and guarantees metadata adheres to the schema. Concretely, each final predicted observation is represented as `id, name, value_type, value`, where `name` and `value_type` are taken from $\mathcal{S}$ and `value` is generated by the LLM and validated against $\mathcal{S}$. We evaluate the same pipeline configurations with two LLM backbones: *Llama-4-Scout-17B-16E-Instruct* (open-weights) (Meta, 2025) and *GPT-5.2* (closed-weights) (OpenAI, 2025), and report the comparative effects of RAG, schema-constraints (full or pruned schema), and second-pass auditing in the ablation study.

4.2. Retrieval Corpus and Knowledge Base

We use the training split as the sole retrieval corpus and source of in-context demonstrations. Each training instance is parsed from JSONL as a tuple (x_i, y_i) containing a transcript x_i and a gold label observation list y_i. We convert the gold label observations into a standardized representation $y_i^{\text{std}} = [(id_j, v_j)]_{j=1}^{K_i}$, where id_j is the concept identifier and v_j is the corresponding value. We also derive a summary text for each training instance by mapping their gold label concept IDs to their concept names and concatenating them with a delimiter (e.g., "$\text{name}_1 \mid \text{name}_2 \mid \cdots$"). This yields two complementary textual views per training example: the original transcript text and a concept name summary text derived from the schema.

For semantic retrieval, we embed training instances and queries using the same encoder within each pipeline: Sentence-Transformers `all-MiniLM-L6-v2` (Reimers et al., 2021) for the Llama-4 and OpenAI `text-embedding-3-large` (OpenAI, 2024) for the GPT-5.2. We precompute and cache embeddings for each transcript view (and, when retrieval is enabled, for the concept name view as well), along with y_i^{std} for use as few-shot examples. At inference time, the query transcript is embedded under the same encoder and compared against the cached training embeddings to retrieve candidate exemplars.

4.3. Exemplar Retrieval for RAG

When retrieval is enabled, we use an exemplar retrieval strategy that leverages both the narrative content of the transcript and the structured observation concepts associated with each training instance. Each training example is represented by two textual views: (i) the original transcript and (ii) a schema-derived concept-name summary constructed by mapping its gold label concept IDs to their standardized schema names. For a query transcript, we first retrieve a candidate pool using cosine similarity over transcript embeddings, and then re-rank the pool using a weighted score:

$$\text{score(candidate pool)} = w_t.s_t + w_c.s_c + w_l.s_l$$

where s_t is cosine similarity between transcript embeddings, s_c is cosine similarity between concept-name summary embeddings, and s_l is a lexical overlap score computed as Jaccard similarity over token sets (Schütze et al., 2008). We set $(w_t, w_c, w_l) = (0.70, 0.25, 0.05)$ based on experiments on the dev set, as this combination yielded the best overall retrieval quality and downstream extraction performance. The top-k exemplars under this score are inserted into the prompt as few-shot demonstrations, each consisting of the exemplar transcript paired with its standard gold label output as a JSON list of (id, v) pairs. We vary the number of retrieved exemplars with $k \in \{3, 5, 10, 20, 30, 50\}$ and analyze its effect.

4.4. Schema Constraints

We study two strategies for conditioning the LLM on the task schema $\mathcal{S}$, which enumerates all observation concepts and their value constraints. In *full schema* prompting, we provide the complete schema to the model, including each concept's identifier, standard name, value type, and (when categorical) its allowable value set. In *pruned candidate schema* prompting, we instead construct a per-instance candidate concept set $\mathcal{C}(x) \subseteq \mathcal{S}$ intended to cover concepts relevant to transcript x while limiting the output space. The prompt then includes only a compact candidate table for $\mathcal{C}(x)$ (identifier, name, value type, and truncated allowable values for categorical concepts).

To construct $\mathcal{C}(x)$, we score concepts by lexical match between transcript tokens and concept-name tokens (and, for categorical concepts, limited overlap with enumerated tokens), retaining the highest-scoring lexical candidates. We augment this with a semantic schema match by retrieving nearest concept names from an embedding index over all names in the schema and keeping matches above a fixed similarity threshold. We then expand $\mathcal{C}(x)$ using retrieval expansion by adding concept IDs observed in the retrieved exemplars' gold label outputs, and we additionally inject a small set of common observation patterns (e.g., vital signs and frequently occurring nursing assessment fields). Finally, we size the candidate set to a target budget (with minimum and maximum bounds) by trimming low-scoring candidates while preserving a small set of common concepts, or padding with additional lexical/semantic candidates when the set is too small.

4.5. Prompt Construction and LLM Inference

For each input transcript, we construct a prompt designed to elicit strictly structured outputs that adhere to the schema. Figure 2 illustrates the prompt, which begins with system-level instructions that enforce a JSON-only response and explicitly prohibit additional text, code fences, or explanations. We require a single JSON list of objects, each containing exactly an observation identifier and its value (i.e., (id, v)), and we instruct the model to omit any observation not supported by evidence in the transcript and to avoid duplicate identifiers.

We then provide either a full schema or a pruned candidate schema and add few-shot examples when RAG is enabled: up to k retrieved exemplars formatted as a transcript followed by its gold label output. Finally, we append the query transcript and a concise instruction to produce the JSON output. We apply the same prompt template across LLM backbones, differing only in the inference interface

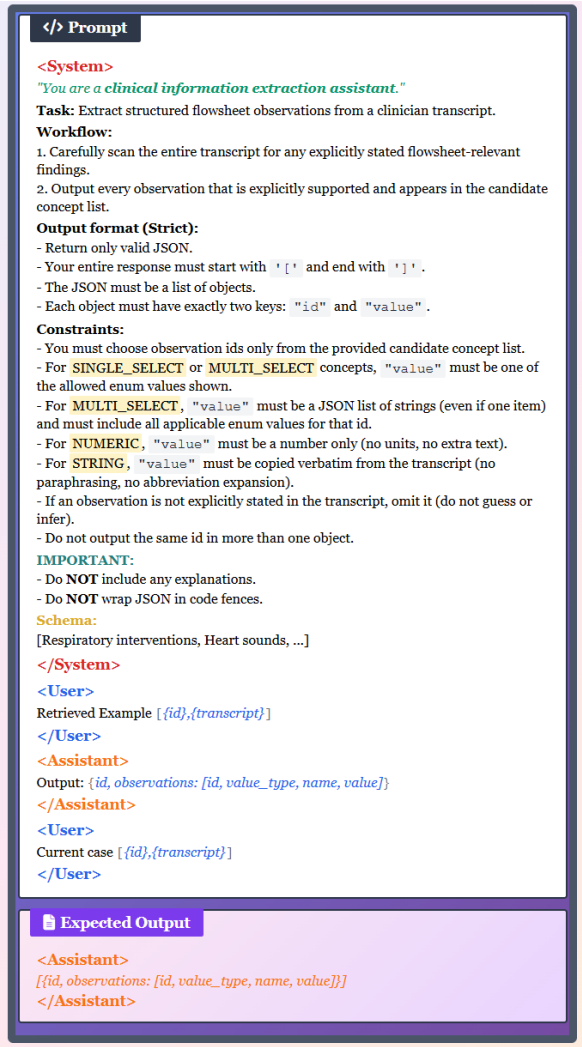

Figure 2: The structured prompt with retrieved exemplars, schema, and the expected output.

(Transformers for *Llama-4-Scout-17B-16E-Instruct* and API-based inference for *GPT-5.2*).

4.6. Postprocessing and Schema Validation

We apply postprocessing to make model outputs robust. First, we deterministically clean and parse the generated text into a JSON list of (id, v) pairs, removing common formatting errors when necessary. We then validate predictions against the schema $\mathcal{S}$ by dropping unknown concept IDs and normalizing values according to each concept's `value_type`. For categorical concepts, outputs are constrained to allowable enumerated values; for multi-select concepts, values are converted to lists and deduplicated; and for numeric concepts, numeric strings are converted to numbers when possible. Duplicate IDs are removed by keeping the first valid occurrence. Finally, we expand each validated (id, v) pair into the required sub-

LLM Backbone	Setting	Precision	Recall	F_1 score
Llama-4	Prompt-only + full schema	72.92	52.84	61.28
Llama-4	Prompt-only + pruned candidate schema	74.64	64.02	68.92
Llama-4	Prompt + RAG + full schema	77.24	68.24	72.46
Llama-4	Prompt + RAG + pruned candidate schema	78.58	68.50	73.20
GPT-5.2	Prompt-only + full schema	81.97	73.17	77.32
GPT-5.2	Prompt-only + pruned candidate schema	69.80	83.57	76.07
GPT-5.2	Prompt + RAG + full schema	78.59	82.03	80.27
GPT-5.2	Prompt + RAG + pruned candidate schema	81.71	73.54	77.41
GPT-5.2	Prompt + RAG + full schema + 2nd pass	78.62	82.18	**80.36**

Table 2: Performance (%) across Llama 4 and GPT-5.2 with schema/RAG configurations. Values are rounded to two decimals. Best F_1 is bolded.

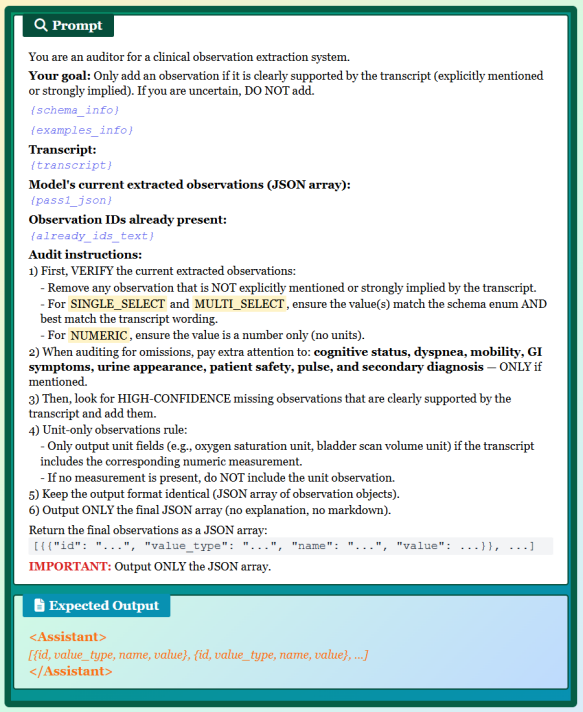

Figure 3: The structured prompt for the second pass audit with retrieved exemplars, schema, first pass solution, and the expected output.

mission format by retrieving the standard `name` and `value_type` from the schema, producing `id`, `name`, `value_type`, `value` entries.

4.7. Second-Pass Auditing

Figure 3 shows the prompt for the second pass auditing step. We evaluate this step as a refinement stage over the first pass, which is the initial system output produced by the LLM from a single pass over the transcript and schema. The auditor is prompted with the original transcript, the same schema used in the first pass (full schema or pruned candidate schema), and the first-pass prediction, and is instructed to remove unsupported items, correct schema adherence issues, and add only clearly supported missing observations. The audited output is then passed through the same schema validation and deterministic expansion step as the first pass.

4.8. Evaluation Strategy

We compute precision, recall, and F_1 over extracted observations by matching predicted and reference items by concept `id` and `value`. For MULTI_SELECT concepts, the evaluation script expands each selected value into a separate item prior to matching, effectively scoring multi-select outputs at the individual choice level. Categorical and string values are compared as strings, and numeric values are compared after conversion to floating point.

5. Results and Discussion

5.1. Overview

Table 2 reports the performance of our retrieval-augmented, schema-constrained extraction pipeline across LLM backbones and design variants. The strongest configuration uses *GPT-5.2* with OpenAI `text-embedding-3-large` for exemplar retrieval, combined with full schema prompting and $top\text{-}k = 30$ retrieved exemplars, yielding the best overall F_1. We observe that increasing the number of retrieved exemplars beyond this point produces only marginal changes, suggesting diminishing returns from additional exemplars. Finally, adding a second-pass auditing stage yields a small additional improvement, indicating that most performance gains are driven by the primary system, with second-pass primarily addressing residual normalization and schema-adherence issues.

5.2. Ablation Study

We conduct an ablation analysis to isolate the contributions of RAG, schema, and second-pass auditing. We start with a prompt-only baseline using *Llama-4-Scout-17B-16E-Instruct* and the full schema, which illustrates the difficulty of extracting and normalizing concepts when the model must reason over a large output space under strict type and enumeration constraints. In this case, errors are dominated by missed concepts and schema-inconsistent outputs, reflecting both limited grounding from the input and the challenge of selecting among many semantically adjacent concepts.

Adding retrieval augmentation improves performance for both LLMs. Concretely, retrieved examples reduce false negatives by providing localized evidence for what constitutes an extractable observation and how it should be normalized under the schema. They also reduce formatting and validity errors by reinforcing the expected JSON structure and value through examples. This effect is especially visible for categorical concepts, where the examples encourage selecting values from the allowed set rather than generating free-form paraphrases.

We then study different schema strategies. For open-weight LLM, pruned candidate schema prompting produces the largest improvement, yielding the best results for *Llama-4* when combined with retrieval. This behavior is consistent with a constrained-decision-space effect: pruning removes many implausible concepts, making it easier for a smaller model to adhere to the schema and to focus extraction on the most relevant portion of the schema. This change improves recall while keeping precision stable or slightly improved, indicating that the model misses fewer relevant observations without a corresponding increase in over-extraction.

For the larger close-weights model, we observe a different interaction between retrieval and schema-constraint. Candidate pruning tends to increase recall but decrease precision, consistent with a checklist-like bias in which the reduced candidate set implicitly encourages broader extraction even when supporting evidence is weak. Under retrieval augmentation, the best *GPT-5.2* performance is instead obtained with full schema prompting, suggesting that a larger model can exploit entire schema context to better disambiguate related concepts while using retrieved exemplars primarily for grounding and normalization. Finally, adding a second-pass auditing stage on top of the best *GPT-5.2* configuration yields a small additional gain. The magnitude of this improvement indicates that most performance is already captured by the combination of retrieval and schema-constraints, with the second-pass primarily correcting residual schema-adherence and normalization errors rather than changing extraction coverage.

Taken together, the ablation results support a clear finding: retrieval augmentation is broadly beneficial, but the optimal schema constraint is model-aware. Pruned candidate schema prompting is critical for controlling the effective output space for the smaller open-weight model, whereas the larger closed model benefits more from full schema context when paired with RAG. The best system performance is because of the combination of (i) exemplar grounding that reduces omissions and stabilizes normalization, and (ii) a full schema that enables consistent disambiguation under strict type and enumeration requirements.

Error type	Count
False positives	**608**
Spurious concept (id not in gold)	498
Wrong value (id in gold)	110
False negatives	**485**
Missed concept (id not predicted)	382
Wrong value (id predicted)	103

Table 3: Breakdown of false positives (FP) and false negatives (FN) into spurious concept errors versus value mismatches.

6. Error Analysis

We analyze errors by comparing predicted observations against the reference output (with MULTI_SELECT values treated as one item per selected choice). Across the test set (199 cases), the system produces 2236 true positives, 608 false positives, and 485 false negatives, as shown in Table 3, indicating that most residual errors arise from either predicting extra items or missing gold label items rather than near-miss value disagreements. Consistent with this, 81.9% of false positives (498/608) correspond to *spurious predictions* whose concept IDs are not present anywhere in the ground truth set for that case, while the remaining 18.1% (110/608) are cases where the correct concept is predicted but with an incorrect value. On the false-negative side, 78.8% (382/485) are *omitted concepts* (the concept ID is not predicted at all), and 21.2% (103/485) are attributable to value mismatches. We also find that errors increase primarily with *clinical content density*: the number of errors correlates more strongly with the number of gold label items ($\rho = 0.45$) and the number of predicted items ($\rho = 0.53$) than with transcript length ($\rho = 0.24$), suggesting that dense dictations amplify both omission risk and over-extraction.

A notable fraction of errors concentrates in a small set of systematic patterns. First, we observe a persistent class of representation-level mis-

Top FP concepts				Top FN concepts			
id	**Name**	**Type**	**FP**	**id**	**Name**	**Type**	**FN**
3	Respiratory interventions	MULTI_SELECT	65	3	Respiratory interventions	MULTI_SELECT	73
31	Secondary diagnosis	STRING	61	148	Gastrointestinal symptoms	MULTI_SELECT	29
0	Broset violence checklist	SINGLE_SELECT	41	0	Broset violence checklist	SINGLE_SELECT	23
162	Patient identification	STRING	26	6	Weightbearing status	MULTI_SELECT	20
6	Weightbearing status	MULTI_SELECT	19	26	Delirium symptoms	STRING	17
26	Delirium symptoms	STRING	17	7	Oral mucosa status	SINGLE_SELECT	16
7	Oral mucosa status	SINGLE_SELECT	17	179	Temperature unit	SINGLE_SELECT	15
117	Patient safety	SINGLE_SELECT	15	89	Mobility	SINGLE_SELECT	15
130	Cognitive status	SINGLE_SELECT	13	107	Fall risk identification	SINGLE_SELECT	14
67	Dyspnea	SINGLE_SELECT	12	130	Cognitive status	SINGLE_SELECT	14

Table 4: Most frequent concept-level errors. False Positive (FP) counts reflect extra predicted items; False Negative (FN) counts reflect missing gold-labeled items (after expanding MULTI_SELECT choices).

matches involving a small subset of low-number concept identifiers that appear with leading zeros in the reference output (e.g., `03` vs. `3`). These mismatches produce both false positives and false negatives even when the extracted content is clinically correct, and they account for a substantial share of total errors (147/485 false negatives and 159/608 false positives). A related phenomenon occurs for temperature units, where the schema's categorical tokens do not match the reference encoding, yielding systematic value mismatches. These issues motivate future work on *robust standardization* for heterogeneous schema and annotations, including normalization of identifier formats and encoding-aware mapping for categorical tokens, as a general requirement when deploying schema-constrained extraction across data sources with inconsistent conventions.

Second, the largest precision losses arise from a small number of *over-predicted* concepts that are plausible given the transcript but rarely used in the references. For example, Table 4 shows that `Secondary diagnosis` appears only 3 times in the gold labels but is predicted 61 times, and `Patient identification` appears only 3 times in the gold labels but is predicted 26 times. This pattern suggests that the model often identifies clinically relevant information but maps it into schema slots that are sparsely annotated or used under stricter, dataset-specific criteria. Addressing this reliably is not a simple rule-based fix; it requires learning concept-specific decision policies and calibration that reflect when a mention should be normalized into a structured field versus left unexpressed.

Finally, residual recall errors frequently reflect *implicit or distributed evidence* that is harder to consolidate into a single normalized observation, especially for checklist-style multi-select fields (e.g., `Gastrointestinal symptoms`) and summary concepts (e.g., `Mobility`). In addition, free-text fields exhibit high brittleness under strict value matching: `Pain description` is present 12 times in gold labels and is frequently predicted, yet none of these predictions match exactly at the string level, indicating that minor paraphrasing or formatting differences can dominate the residual error mass. Improving these cases likely requires methods that better ground string-valued slots in transcript spans (e.g., structured span-copying or alignment) and that model schema overlap explicitly so that evidence expressed in supporting fields is more consistently propagated into the corresponding summary concepts.

7. Conclusion

Schema-constrained extraction from conversational nurse-patient transcripts is an important step toward reducing documentation burden by transforming clinical documentation into structured, schema-adherent observations. Our work presents a modular retrieval-augmented extraction pipeline that combines training-set exemplar retrieval, schema-constrained prompting (full schema or pruned candidate schema), deterministic schema-based postprocessing, and a second-pass audit to refine outputs. Our results support three key findings: pruned candidate schema is beneficial for smaller open-weight models but can be counterproductive for larger models that instead benefit from richer full schema context; retrieval-augmented generation consistently improves performance, with gains strongly influenced by how schema information is presented; and second-pass auditing provides modest additional improvements by correcting residual schema-adherence and normalization issues. Overall, these findings underscore that effective schema-constrained observation extraction is achieved by jointly selecting retrieval and schema-constraint strategies that match the capabilities of the underlying model.

8. References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeya Balaji Balasubramanian, Daniel Adams, Ioannis Roxanis, Amy Berrington de Gonzalez, Penny Coulson, Jonas S Almeida, and Montserrat García-Closas. 2025. Leveraging large language models for structured information extraction from pathology reports. *Journal of Pathology Informatics*, page 100521.

Luc Builtjes, Joeran Bosma, Mathias Prokop, Bram van Ginneken, and Alessa Hering. 2025. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA open*, 8(5):ooaf109.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

Linda Dawson, Maree Johnson, Hanna Suominen, Jim Basilakis, Paula Sanchez, Dominique Estival, Barbara Kelly, and Leif Hanlen. 2014. A usability framework for speech recognition technologies in clinical handover: A pre-implementation study. *Journal of medical systems*, 38(6):56.

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, et al. 2020. Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Cathy Shyr, Qingyu Chen, et al. 2026. Information extraction from clinical notes: are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, page ocaf213.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014a. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1):94.

Maree Johnson, Paula Sanchez, Hanna Suominen, Jim Basilakis, Linda Dawson, Barbara Kelly, and Leif Hanlen. 2014b. Comparing nursing handover and documentation: forming one set of patient information. *International Nursing Review*, 61(1):73–81.

A H M Rezaul Karim and Ozlem Uzuner. 2025. MasonNLP at MEDIQA-OE 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 57–67, Virtual. 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, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Siru Liu, Allison B McCoy, and Adam Wright. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432.

LeAnn Mayer, Dongjuan Xu, Nancy Edwards, and Gordon Bokhart. 2022. A comparison of voice

recognition program and traditional keyboard charting for nurse documentation. *CIN: Computers, Informatics, Nursing*, 40(2):90–94.

AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 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.

Raghad Muhiyaddin, Asma Elfadl, Ebtehag Mohamed, Zubair Shah, Tanvir Alam, Alaa Abd-Alrazaq, and Mowafa Househ. 2022. Electronic health records and physician burnout: a scoping review. *Informatics and Technology in Clinical Care and Public Health*, pages 481–484.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

OpenAI. 2024. text-embedding-3-large Model. OpenAI API documentation.

OpenAI. 2025. GPT-5.2 Model. OpenAI API documentation.

Nils Reimers, Iryna Gurevych, and Sentence-Transformers contributors. 2021. sentence-transformers/all-MiniLM-L6-v2. Hugging Face model card.

Tiago Rodrigues and Carla Teixeira Lopes. 2025. Harnessing large language models for clinical information extraction: A systematic literature review. *ACM Transactions on Computing for Healthcare*.

Lisa S Rotenstein, A Jay Holmgren, Michael J Healey, Daniel M Horn, David Y Ting, Stuart Lipsitz, Hojjat Salmasian, Richard Gitomer, and David W Bates. 2022. Association between electronic health record time and quality of care metrics in primary care. *JAMA Network Open*, 5(10):e2237086–e2237086.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. Real: A retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 380–389.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760.

Hanna Suominen, Maree Johnson, Liyuan Zhou, Paula Sanchez, Raul Sirel, Jim Basilakis, Leif Hanlen, Dominique Estival, Linda Dawson, and Barbara Kelly. 2015a. Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *Journal of the American Medical Informatics Association*, 22(e1):e48–e66.

Hanna Suominen, Liyuan Zhou, Lorraine Goeuriot, and Liadh Kelly. 2016. Task 1 of the clef ehealth evaluation lab 2016: Handover information extraction.

Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. 2015b. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2):e4321.

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

Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. Ramie: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, 32(3):545–554.