SemAnTICA Lab at MediQA-SYNUR 2026: Route, Extract, and Verify – An LLM-gated Ensemble for Parsing Nurse Dictations

Sy Hwang[1], **Katherine S. Pitcher**[1], **Sue Hyon Kim**[1], **Yoonjae Lee**[1], **Hayoung K. Donnelly**[1], **Harsh Bandhey**[2], **Andrew J. King**[3], **Karen O'Connor**[1], **Ryan J. Urbanowicz**[2], **Danielle L. Mowery**[1]

[1]University of Pennsylvania, [2]Cedars Sinai Medical Center, [3]University of Pittsburgh
Philadelphia, PA, USA; Los Angeles, CA, USA; Pittsburgh, PA, USA
{sy.hwang, hayoung.donelly, karoc, dlmowery}@pennmedicine.upenn.edu,
{kpitch, kshyon, yoonjael}@nursing.upenn.edu, {ryan.urbanowicz, harsh.bandhey}@cshs.org,
andrew.king@pitt.edu

Abstract

We describe the **Sem**antic **An**alysis of **T**ext to **I**nform **C**linical **A**ction (SemAnTICA) Lab's system for the MediQA-SYNUR 2026 shared task on extracting structured clinical observations from nurse dictation transcripts. The task requires mapping observations from disfluent conversational text to a large, fixed ontology and producing strictly normalized outputs, where small amounts of concept over-selection severely degrade micro F1 score. Our approach evolved from a full-schema in-context baseline to a pipeline that explicitly separates concept selection from value extraction. We first preprocess transcripts, then generate transcript-specific concept candidates using hybrid sparse–dense retrieval. The candidates are then pruned with an evidence-based filter. For extraction, we adopt a system-level mixture-of-experts design with an online Large Language Model (LLM) router that selects a subset of domain-specialized experts per transcript. Each expert operates over a constrained schema partition to reduce spurious predictions. We enhance robustness with agreement-gated ensembling and targeted adjudication for ambiguous cases. Finally, we intersect complementary high-recall and high-precision runs to produce the best submission. Our system ranked first on the official test leaderboard with $F_1 = 0.814$, $P = 0.826$, $R = 0.801$.

Keywords: information extraction, ontology alignment, nursing documentation, large language models, verification

1. Introduction

In clinical care, patient documentation is both mission critical and burdensome, and nursing workflows are particularly documentation-intensive. To reduce the burden of documentation for nurses, voice recognition (VR) systems have been implemented. While the use of VR has been shown to decrease the time spent on documentation, nurses mainly document via the keyboard in electronic health record (EHR) flowsheets (Mayer et al., 2021). For VR to be more widely adopted, methods must be implemented to extract the relevant clinical data from nurse dictation transcriptions and translate this information into a format compatible with EHR workflow data. MediQA-SYNUR targets this challenge by converting informal nurse dictation transcripts into structured flowsheet-like observations. Two factors drive the difficulty of the automation of this task: (1) spoken language informality, noise and imprecision (Galatzan and Carrington, 2018) and (2) the large, heterogeneous output space. Systems must identify the relevant text, select the correct concept from a sizable ontology and then produce an appropriate value type. Depending on the concept, the value may come from an enumeration, take the form of a numeric measurement with units, or be unconstrained free text.

When attempting to extract and canonicalize outputs to an ontology across a large number and range of clinical concepts, several challenges emerge. At scale, concept selection is the dominant factor affecting performance, as even modest over-selection produces many false positives under strict matching. Free-text `STRING` values further increase difficulty: without a fixed value set, models must determine both extraction boundaries and surface form, while exact-match evaluation penalizes any paraphrasing. Additionally, we observed that some `STRING` references can vary stylistically across otherwise similar evidence, which compounds mismatch risk. These properties motivate a design that treats concept selection as a retrieval-and-verification problem and uses explicit uncertainty control during extraction.

Our approach follows three design principles: separate candidate generation from value extraction via retrieval and evidence checks; use online routing to invoke only relevant domain experts; and control variance with agreement-gated ensembling plus targeted adjudication for ambiguous cases. We describe the evolution of this system and the configurations used in our submissions for the shared task.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 27–37
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Schema-guided information extraction has long been a central strategy for normalizing clinical text into structured representations, particularly when downstream use requires consistent concept identifiers and typed values. MediQA-SYNUR sits squarely in this tradition, with the added challenge that the input is spoken dictation rather than edited clinical notes. The shared task is described in the official overview paper (Michalopoulos et al., 2026), and it builds on the SYNUR dataset introduced by Corbeil et al. (2025), which frames structured observation extraction from nurse dictations as a high impact but underexplored clinical NLP problem due to limited public data availability. As part of the broader MediQA series, the task also inherits a shared-task lineage emphasizing standardized evaluation and comparative system analysis across participants (Ben Abacha et al., 2019, 2021).

Prior work at the intersection of nursing workflows and clinical NLP has explored both speech-derived text and traditional nursing documentation. Song et al. (2022) examine NLP over auto-generated transcripts of patient–nurse communication in home health care, showing that conversational, Automatic Speech Recognition (ASR)-derived text can support downstream risk identification while highlighting noise sources that complicate reliable extraction. King et al. (2023) similarly integrate automatic transcription with NLP in a voice-based digital assistant designed to support Intensive Care Unit (ICU) rounds, illustrating the feasibility of speech-driven clinical decision support and the importance of robust normalization and evidence-grounded prompting. Complementing these speech-oriented efforts, Mitha et al. (2023) review a broader landscape of NLP applied to nursing notes, synthesizing common tasks and methodological patterns in nursing documentation analytics. Together, these studies motivate MediQA-SYNUR's focus on nurse-centered language and workflow realism, while underscoring a gap addressed by the shared task: ontology scale, schema-guided normalization of nursing observations from dictation transcripts under strict evaluation.

Recent clinical NLP systems increasingly employ large language models (LLMs) for extraction and normalization, motivated by their strong instruction following and few-shot generalization (Rodrigues and Teixeira Lopes, 2025). However, ontology scale constraints can shift the dominant failure mode away from value interpretation and toward concept over-selection. When many plausible schema elements are presented, models may default to exhaustive slot filling, producing spurious outputs that degrade precision under strict matching. This issue mirrors observations in other structured medical information extraction pipelines where specification, normalization conventions, and schema constraints strongly shape end performance, and it motivates architectures that separate candidate generation from value decoding.

Retrieval-augmented generation is often framed as a method for injecting external knowledge into generation (Lewis et al., 2020), yet retrieval can also play a purely structural role by selecting which schema elements should be considered for a given input. In schema-guided extraction, this retrieval-as-constraint view is particularly natural: candidate concept selection becomes an information retrieval problem over schema descriptors, synonyms, and notes, and subsequent extraction is performed only over a reduced candidate set.

Robustness methods for LLM-based systems frequently rely on ensembling and adjudication. Multi-model ensembles can reduce variance and provide a practical uncertainty signal through agreement, while selective verification concentrates compute on ambiguous cases. In parallel, "LLM-as-a-judge" approaches have grown into a widely studied paradigm for assessment and decision-making (Li et al., 2025).

Agentic and tool-using LLM systems emphasize input-conditional orchestration, where a controller selects actions or tools based on the current instance. Yao et al. (2023) formalize an interleaving of reasoning and acting, while Schick et al. (2023) demonstrate learning to decide when to call external tools. Related modular architectures such as the Modular Reasoning, Knowledge and Language (MRKL) system describes routing queries to specialized modules, highlighting the value of a router that dispatches work to domain-appropriate components (Karpas et al., 2022). In our setting, the closest analogue is an online LLM router that selects which domain-specialized extraction experts to invoke per transcript. We describe this as a system-level mixture-of-experts (MoE) design with LLM-based gating, conceptually related to conditional computation and routing ideas popularized in sparse MoE models (Lepikhin et al., 2021; Fedus et al., 2022), while differing in that our experts are separate LLM calls rather than trainable subnetworks in a single model.

3. Data and Task

The MediQA-SYNUR task provides annotated synthetic nurse dictation transcripts and a fixed ontology of observation concepts. Each output record links a clinical observation mention in the transcript to a schema concept and a typed value. The schema spans multiple clinical domains and includes four value types: categorical `SIN-`

GLE_SELECT (choose exactly one option), categorical MULTI_SELECT (choose one or more options), NUMERIC measurements, and free-text STRING.

Figure 1 summarizes two corpus-level properties that shape our system design. (a) The distribution of annotated observations by value type shows that categorical and numeric fields constitute the majority of labels, while unconstrained STRING fields are a small fraction. (b) The distribution of the number of annotated observations per transcript in the training and development splits shows a similar central tendency (≈ 13 observations per sample). Together, these trends indicate that the effective output space per transcript is small relative to the full ontology, and that most supervision is concentrated in non-STRING types. This motivates curation of schema concepts specific to each transcript, and conservative handling of unconstrained free-text values under exact-match scoring.

The transcripts are synthetic but resemble clinical speech, containing hedges and fillers, abbreviated phrasing and implicit references. Because the evaluation requires exact matching of normalized outputs, the task rewards conservative, evidence-backed extraction. In practice, this means that a system must balance recall against the risk of over-selecting concepts that are only weakly implied.

4. Methods

4.1. Overview

Our system is a multi-stage pipeline that gradually narrows uncertainty. We (1) normalize and clean transcripts, then (2) generate a transcript-specific set of candidate concepts via retrieval. Next, we either (3) filter candidates using an evidence gate that removes concepts without direct textual support, or (4) use an online router to select which domain experts to invoke for the transcript. Each expert extracts observations from a constrained subset of the schema. When we run multiple models within an expert, (5) we retain only extractions supported by cross-model agreement and send remaining ambiguous cases to an adjudicator. Finally, (6) we canonicalize outputs and apply a consensus-based intersection of complementary runs.

Throughout, we view the system as taking on two distinct decisions: identifying which concepts are present in the transcript, and then extracting their values. In our error analysis, misses were driven primarily by the first step, concept selection, rather than by value extraction.

4.2. Transcript pre-processing

We pre-process transcripts to reduce superficial variability that confuses both retrieval and generation. Specifically, we repair encoding artifacts when present (e.g. "\u00e2\u20ac\u201d"), removing speaker tags (e.g. "[Clinician]") and conversational markup (e.g. "Okay, let's see here"), and lightly filtering disfluencies that do not carry clinical meaning (e.g. "uh" or "um"). This pre-processing improves lexical and semantic matching between transcript phrases and schema descriptors.

4.3. Schema enrichment

Early error analyses showed that surface-form mismatch was a major source of both missed concepts and unstable retrieval. To improve coverage, we enriched clinical concepts from the schema with concise synonym strings and concept-specific normalization guidance that defines how values should be represented. This information is injected at extraction time as a compact constraint, reducing variation in generated values. For example, when a transcript indicates that an assessment was performed, the notes can specify a preferred normalized value rather than allowing the model to invent phrasing.

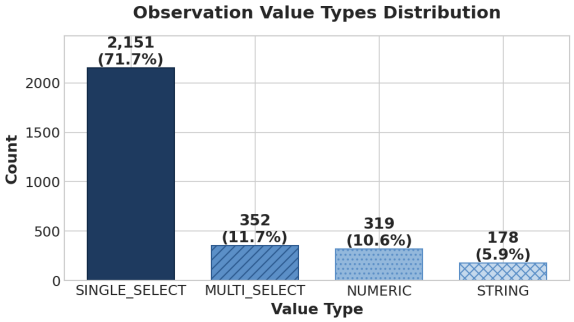
4.4. Candidate concept generation via hybrid retrieval

Presenting all schema concepts to a single extractor encourages over-selection. Instead, we generate a transcript-specific candidate set using retrieval over schema descriptors (concept name, synonyms, and notes). We use sparse lexical matching via term frequency-inverse document frequency (TF-IDF) to capture exact phrase overlap and dense embeddings to capture semantic similarity via cosine distance. A hybrid fusion strategy balances the two, allowing synonymy to be recovered without losing the precision of lexical anchors.

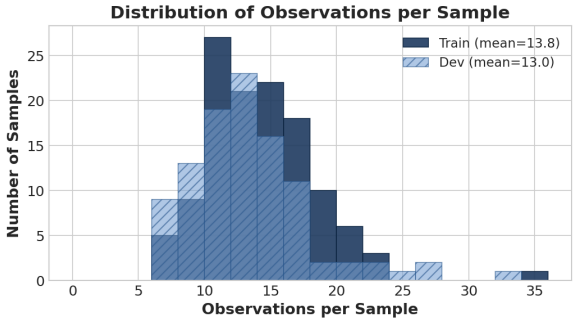
This stage is best understood as candidate generation in an information retrieval pipeline. It is intentionally permissive, but it should shrink the effective output space from the full ontology to something closer to the expected number of observations per transcript.

4.5. Evidence-based candidate filtering

Hybrid retrieval improves recall but can still propose candidates that are merely thematically related to the transcript. To reduce false positives, we apply a lightweight evidence check. For each candidate concept, a model must point to explicit supporting text or make a tightly constrained inference that is defensible from the transcript. Candidates that cannot be justified are removed. This step shifts the system from semantic association toward evidence-grounded selection, which is essential under strict evaluation.



(a) Distribution of value types.



(b) Distribution of the number of observations per sample.

Figure 1: Data characteristics of MEDIQA-SYNUR training and development sets: (a) value-type distribution; (b) observations per transcript.

4.6. Few-shot example retrieval

Many remaining errors reflect normalization conventions, including categorical choice, measurement formatting, and output structure. To stabilize these choices, we retrieve a small set of similar training transcripts and include their labeled outputs as demonstrations. Few-shot examples are selected using the same representations used for schema retrieval, aligning the model's context with the transcript's phrasing.

4.7. Value extraction with online LLM routing and domain experts

4.7.1. Reduced-schema global extraction

In the reduced-schema variant, we expose a single extractor, or a committee of extractor models, to the transcript's candidate concepts distilled from the retrieval process, rather than the full ontology. This proves effective when candidate reduction is sharp, but can still leave the model facing a large and heterogeneous schema when many concepts remain plausible. We therefore introduce online routing as a complementary strategy for managing schema overload. Instead of one global pass over a broad reduced schema, we select a small set of domain-partitioned experts per transcript to limit full ontology exposure.

4.7.2. Domain-specialized experts with constrained output spaces

To mitigate prompt overload and curb concept over-selection, we decompose the full ontology into domain-specialized extraction experts (e.g. cardiovascular or mobility), each responsible for a disjoint subset of schema concepts (approximately ≤ 20 concepts per expert). Each expert is implemented as an LLM call with a domain-specific prompt containing only its assigned concepts, which constrains the output space and reduces the likelihood of spurious extractions. This design functions as a system-level mixture-of-experts, where "experts" are domain-specialized LLM extractors and constraint is enforced by the schema partition.

Figure 2 illustrates our system-level mixture-of-experts design. Given a nursing dictation transcript, an LLM router selects a subset of clinical domains (optionally subject to a routing budget K). Each selected domain expert operates over a disjoint schema partition and extracts candidate observations constrained to its domain-specific schema and normalization rules. For robustness, each domain expert can be instantiated as a small committee of heterogeneous LLMs; extracted observations are aggregated using an agreement gate over $(\texttt{concept\_id}, \texttt{value})$ pairs (accept if at least 2/3 models agree). Accepted observations are merged across domains to form the final structured output.

We define domains aligned with clinical practice patterns (e.g., vital signs, respiratory, neuro exam, GI, GU, skin/wound, mobility/safety, behavioral risk, pain, and general patient information). Domain decomposition substantially reduces per-call schema size and improves precision by narrowing attention to semantically related concepts.

4.7.3. Online LLM routing

Rather than invoking all experts for every transcript, we use an online LLM router to select which domain-specialized experts to run for each input. The router reads the normalized transcript and returns a subset of relevant domains, optionally with confidence scores or an expert budget. Routing is input-conditional: a transcript centered on shortness of breath and oxygen therapy may activate respiratory and cardiopulmonary experts, whereas a wound-care update may activate skin/wound and mobility/safety experts. This reduces unnecessary expert calls and directly targets the main failure mode observed in early submissions: inflated numbers of predicted observations caused by ontology-

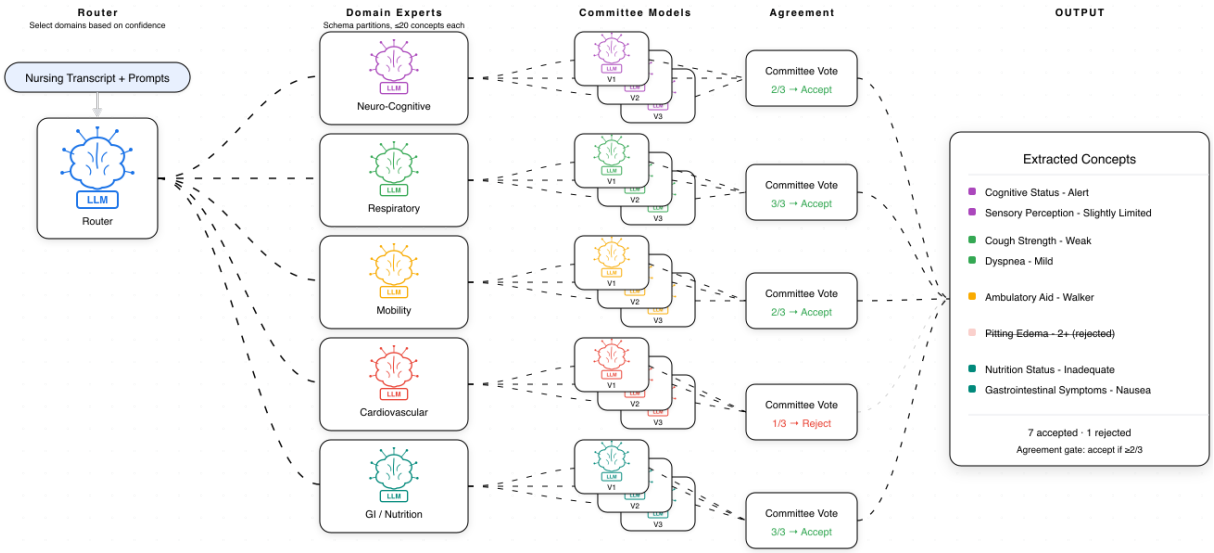


Figure 2: Ontology partitioning into domain-specialized experts (Section 4.7.2).

scale over-selection.

In implementation, the router is prompted to be conservative. It is instructed to select only domains supported by explicit transcript evidence and to avoid precautionary or "just in case" routing. When router confidence is low, or when a transcript appears genuinely multi-domain, we allow a larger expert budget and rely on downstream agreement gating and validation to control false positives. The effective subset of concept space considered by the system is therefore input-dependent rather than fixed.

This routing step gives the system a mixture-of-experts style architecture at the system level: the router acts as an LLM-based gating mechanism that determines which specialized extractors are applied to a given transcript. We use "mixture-of-experts" in this operational sense only; we do not train an end-to-end learned gating network.

4.8. Agreement-gated ensembles and targeted adjudication

Within each routed domain expert, we optionally run a committee of heterogeneous LLMs and aggregate extractions using agreement-gated voting over `(concept_id, value)` pairs. Items failing the agreement threshold, or items with small margins between competing outputs, are routed to targeted adjudication: a separate verifier model re-checks transcript evidence and accepts or rejects disputed items. Because adjudication is triggered only for ambiguous cases, it provides a favorable cost–accuracy trade-off relative to always-on verification. Importantly, this design complements online routing: routing limits which experts run, and agreement/adjudication limits which extractions survive.

4.9. `STRING` handling and competition-time decision

`STRING` concepts behaved differently from categorical and numeric concepts. Without enumerated value sets, the model must determine both the boundary of the extracted text and its canonical phrasing. Under exact-match evaluation, even semantically correct paraphrases score as errors. In addition, clinician review of training and development set reference standards suggested that a significant portion of `STRING` references can be inconsistent across similar transcript evidence. To avoid optimizing against noisy supervision, and given the disproportionate false positive risk introduced by unconstrained values, we adopted a pragmatic policy to exclude `STRING` concepts from output in configurations where doing so improved overall micro F1. This decision is not a claim that `STRING` extraction is unimportant; rather, it reflects the difficulty of unconstrained value generation under strict scoring in noisy transcripts.

4.10. Validation and canonicalization

We apply deterministic post-processing to enforce schema constraints, ensure type consistency, normalize casing and units where applicable, and remove invalid or malformed outputs. This stage also protects against model "creativity" by rejecting outputs that do not match the schema's expected structure.

4.11. Late fusion rule for the final run

In the final phase, we treated system combination as a decision-level fusion rule rather than a run-specific heuristic. We maintained two complementary extractors: a high-recall configuration that

favors broader candidate generation and routing, and a high-precision configuration that uses conservative routing with agreement gating. Because ontology scale evaluation heavily penalizes spurious concepts, we used cross-system agreement as a confidence signal at the concept level. Concretely, we accepted a concept only when it was predicted by both systems, and we selected the corresponding value from the configuration with the lower value-format error rate (the high-precision system). This rule can be viewed as a special case of thresholded multi-system voting and was chosen based on development-set calibration to control false positives without materially reducing recall.

4.12. Implementation Details

This subsection summarizes the principal implementation choices needed to understand and approximately reproduce the system configurations evaluated in this study. Unless stated otherwise, hyperparameters were tuned on the training/development splits and fixed for inference.

Schema candidate retrieval. Each schema concept is serialized as a text string comprising the concept name and auxiliary descriptors (e.g., synonyms, concise notes). We employ a hybrid retrieval strategy: sparse retrieval utilizes a TF–IDF vectorizer over unigrams and bigrams with L_2 normalization, while dense retrieval computes cosine similarity between transcript and concept embeddings using OpenAI's `text-embedding-3-small`. Given the dataset scale, we employ brute-force cosine scoring. The sparse and dense signals are combined via linear score fusion after min–max normalization:

$$s(c \mid x) = \alpha \cdot \tilde{s}_{\text{tfidf}}(c \mid x) + (1-\alpha) \cdot \tilde{s}_{\text{emb}}(c \mid x), \quad (1)$$

where $\alpha \in [0, 1]$ is a run-specific hyperparameter (see Table 2). We retain the top-k candidates (typically 40–60) for downstream processing.

Few-shot selection. Demonstrations are retrieved from the training set using the same hybrid TF–IDF/embedding strategy. We select k_{shot} examples (typically 3–10) and format them as transcript–label pairs using the official output schema to encourage structural consistency.

Evidence gating. To mitigate false positives from permissive retrieval, we introduce an evidence gate. This module verifies candidate concepts against the transcript, rejecting those lacking explicit support. A candidate is retained only if the model identifies a supporting span and the value type aligns with the evidence.

Online routing and expert invocation. Domain-specialized experts are implemented as distinct extraction prompts, each restricted to a disjoint schema partition (≈ 20 concepts per domain). A router model (GPT-5) processes the transcript and domain descriptions to output: (i) a subset of relevant domains, and (ii) a ranked list of concept candidates with confidence scores. We enforce a global budget by selecting the top-K candidates based on router confidence. These selections determine which expert prompts are invoked and constrain the schema exposed to each expert.

Models and decoding. For extraction, we utilize temperature $T = 0.2$. For reasoning-capable models, we set `reasoning_effort="high"`. All components (router, expert, and judge) are constrained to emit a fixed JSON schema of (*concept_id*, *value_type*, *value*) tuples to ensure deterministic parsing. Generation limits are calibrated to cover expected observation counts while penalizing verbosity. Unless otherwise specified in Table 2, the same family of LLMs used for extraction was also used for evidence-gated candidate verification, while GPT-5 served as the router and, in most committee configurations, the adjudication model.

Committee aggregation and adjudication. When operating as a committee, outputs are aggregated by voting. A concept-value pair (c, v) is accepted if its support fraction exceeds a threshold τ_{agree}:

$$\frac{1}{n}\sum_{i=1}^{n}[(c, v) \in O_i] \geq \tau_{\text{agree}}, \quad (2)$$

where O_i is the output set from voter i. Conflicts where competing values for a single concept have a support margin difference below δ are flagged and optionally routed to a judge model for evidence-based resolution.

Post-processing and canonicalization. We filter invalid concept identifiers and malformed JSON. Categorical values are mapped to canonical schema labels, while numeric fields undergo normalization (e.g., unit standardization, decimal formatting). Final outputs are deduplicated and serialized into the submission format.

Decision-level fusion. The final submission utilizes decision-level fusion across two independently configured systems. A concept is extracted only if predicted by both systems (cross-system agreement). The value is selected from the system demonstrating lower value-format error on the development set, serving as a conservative consensus filter.

Prompt and schema refinement. We iteratively refined prompts using the development set, guided by structured review from three clinically trained nurses on our team. Review focused on recurring failure modes such as prediction consistency, value normalization, and missed mentions. In parallel, the nurse reviewers expanded the schema with additional surface forms such as synonyms and common abbreviations for high frequency concepts to improve candidate recall during retrieval and reduce brittle exact-match errors. All refinements were conducted using only the shared task development data. Clinician input was limited to qualitative review and schema expansion. No additional labeled examples beyond the released development annotations were created. We did not use and did not have access to the test set for prompt or schema updates.

5. Results

5.1. Official leaderboard performance

Our best run (Run 526481 on Codabench) achieved $F_1 = 0.814$, $P = 0.826$, $R = 0.801$, ranking first on the official test leaderboard. A comparison of the ablation and fusion variants we evaluated is provided in Appendix A.

5.2. What improved performance late in development

The strongest improvements came not from better prompting in isolation but from controlling the number of plausible concepts a model is asked to consider at once. Our initial test submissions exhibited high recall but weaker precision, consistent with inflated predicted observation counts. Introducing online expert routing reduced irrelevant concept exposure, and domain specialization made it easier for extractors to stay faithful to transcript evidence. Agreement gating further dampened variance, while targeted adjudication prevented ambiguous edge cases from dominating errors. Results for each configuration from our test submissions and their corresponding performance on the development set are provided in Table 1.

As a shared task system, our goal was to assemble a reliable high-precision extraction pipeline rather than to isolate each module under a fully factorial experimental design. The reported comparisons therefore support the utility of the overall architecture and several major design choices, but do not fully disentangle all higher-order interactions among routing, agreement gating, adjudication, and late fusion.

5.3. Qualitative behavior of the final system

The final system exhibited a more conservative extraction profile: it was less likely to output observations that were merely plausible in context, and more likely to output observations with explicit textual support. This behavior is reflected in the mean number of predicted observations per transcript, $\bar{x}_{\text{obs}}$, which was released to participants after the submission phase closed. The gold test set label density had $\bar{x}_{\text{obs}} = 13.61$, closely matching our final system's predicted $\bar{x}_{\text{obs}}$. Concept-level intersection filtering provided an additional safety layer by removing low-confidence, single-run concepts while preserving the core set of clinically salient extractions.

6. Discussion

6.1. Ontology scale makes concept selection the bottleneck

A central lesson of this shared task is that ontology scale changes the nature of extraction. When the number of candidate concepts is large relative to the number of true observations per transcript, false positives become the dominant failure mode. Even if a model can accurately extract values conditional on the correct concept, overall performance degrades if it cannot reliably decide which concepts are actually present. Candidate generation, evidence filtering, and routing are therefore not ancillary utilities; they are the primary mechanisms that turn a general-purpose generator into a disciplined, ontology scale extractor.

6.2. Online routing is a practical form of structured inference

Online routing offers a middle ground between monolithic prompting and full learned MoE. It uses an LLM's semantic understanding to decide where to spend attention and compute, but it preserves strict output constraints by delegating extraction to experts with narrow schema partitions. In practice, routing behaves like structured inference: it encourages the system to commit to a small set of hypotheses and to ignore the rest.

6.3. Agreement as uncertainty, adjudication as targeted verification

Ensembling is often used for robustness, but its most useful role here was as a proxy for uncertainty. Disagreement reliably marked concepts that were weakly evidenced, ambiguously phrased, or prone to normalization variation. By adjudicating only those cases, we achieved much of the benefit

Table 1: Development and test set performance across system configurations. Rows are sorted by test F1 (descending).

	Development Set			Test Set			
run	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1** ↓	$\bar{x}_{\text{obs}}$
A	0.860	0.870	0.865	**0.826**	0.801	**0.814**	13.26
B	**0.868**	0.863	*0.865*	0.802	0.813	*0.807*	13.86
C	*0.862*	*0.870*	**0.866**	0.800	0.813	0.807	13.89
D	0.842	0.867	0.854	0.767	0.845	0.804	15.06
E	0.856	0.843	0.850	*0.802*	0.778	0.790	13.26
F	0.823	**0.882**	0.851	0.707	**0.853**	0.773	16.51
G	0.843	0.819	0.831	0.738	0.806	0.770	13.89
H	0.821	0.833	0.827	0.654	*0.848*	0.738	17.74
I	0.753	0.797	0.774	0.719	0.630	0.671	11.98
J	0.772	0.724	0.747	0.706	0.637	0.670	12.34

of verification without the cost of verifying every extraction.

6.4. `STRING` remains an open challenge under strict scoring

The difficulty of free-text extraction under exact-match evaluation is not unique to this task, but the combination of spoken dictation and inconsistent normalization can make it especially severe. Our decision to exclude `STRING` in some configurations reflects a reality: unconstrained outputs create a large surface for mismatches, and a small number of systematic `STRING` false positives can dominate micro F1. Future shared tasks may benefit from evaluation protocols that reward semantic correctness or span fidelity for free-text fields.

6.5. Annotation protocol shift across splits

During error analysis, our clinical reviewers noted apparent inconsistencies in the normalization of some observations in the development set reference standard, most prominently for unconstrained `STRING` concepts where similar transcript evidence could correspond to different reference values. In follow-up communication, the organizers clarified that the training and development splits each underwent a single round of annotation, with multiple trained annotators labeling different portions of the data, whereas the test split received an additional round of annotation by different annotators followed by reconciliation to resolve disagreements and improve consistency. This design is a reasonable allocation of annotation effort for shared tasks and likely increases the reliability of the held-out evaluation. At the same time, differing annotation workflows across splits can introduce an annotation protocol shift, where systems tuned on noisier or more heterogeneous training and development labels may generalize differently under a more consistent test reference standard. This effect is especially salient for free-text normalization, where stylistic variation and boundary ambiguity amplify disagreement. Future iterations could strengthen interpretability by applying comparable adjudication procedures across splits when feasible, or by reporting split-specific inter-annotator agreement and reconciliation policies to clarify expected test set behavior.

7. Conclusion

We presented SemAnTICA Lab's MediQA-SYNUR 2026 submission, a pipeline that treats ontology scale extraction as a problem of disciplined concept selection and controlled generation. By separating candidate generation from value extraction, adding evidence-based filtering, and introducing an online LLM router to gate domain-specialized experts, we improved precision without sacrificing recall. Agreement-gated ensembling and targeted adjudication further stabilized outputs, and a consensus-based intersection of complementary runs produced our best leaderboard result. More broadly, the shared task underscores that schema scale and value type heterogeneity remain central obstacles for dictation to structure extraction. Our results suggest that these challenges are best addressed not by a single monolithic model call, but by disciplined selection and validation of candidates through retrieval, expert routing, and consensus systems. We expect future work to build on this principle by improving evidence calibration and value normalization while retaining the same system-level guardrails.

8. Limitations

Our system improves precision through staged control, but this comes with important tradeoffs. Retrieval, routing, committee inference, and optional

adjudication increase latency and cost relative to simpler baselines, and our best configurations rely on multiple external LLMs whose availability, pricing, and institutional acceptability may change over time. Because performance emerges from the interaction of several modules, the contribution of any single component is only partially isolated by our ablations. In addition, the system was tuned for a benchmark with synthetic nurse dictations and strict exact-match scoring, so its behavior may differ in real clinical documentation settings, especially for unconstrained free-text fields. Finally, our decision to exclude STRING in some configurations reflects benchmark-specific optimization rather than a general solution to free-text extraction.

9. Ethics Statement

The task is motivated by reducing documentation burden, but automated extraction systems can omit, distort, or overstate clinically relevant information. Any real deployment should include human oversight, careful auditing, and calibration for high risk concepts. Our work reports aggregate performance and common failure patterns and does not attempt to infer sensitive attributes beyond the task's intended outputs from this synthetic dataset.

10. Bibliographical References

Asma Ben Abacha, Yassine Mrabet, 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, Online. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

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

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Benjamin J. Galatzan and Jane M. Carrington. 2018. Exploring the state of the science of the nursing hand-off communication. *Computers, Informatics, Nursing: CIN*, 36(10):484–493.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. 2022. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *CoRR*, abs/2205.00445.

Andrew J. King, Derek C. Angus, Gregory F. Cooper, Danielle L. Mowery, Jennifer B. Seaman, Kelly M. Potter, Leigh A. Bukowski, Ali Al-Khafaji, Scott R. Gunn, and Jeremy M. Kahn. 2023. A voice-based digital assistant for intelligent prompting of evidence-based practices during icu rounds. *Journal of Biomedical Informatics*, 146:104483.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

LeAnn Mayer, Dongjuan Xu, Nancy Edwards, and Gordon Bokhart. 2021. A comparison of voice recognition program and traditional keyboard charting for nurse documentation. *Computers, Informatics, Nursing: CIN*, 40(2):90–94.

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.

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. *CIN: Computers, Informatics, Nursing*, 41(6):377–384.

Tiago Rodrigues and Carla Teixeira Lopes. 2025. Harnessing large language models for clinical information extraction: A systematic literature review. *ACM Trans. Comput. Healthcare*, 6(4).

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the 37th Conference on Neural Information Processing Systems*.

Jiyoun Song, Maryam Zolnoori, Danielle Scharp, Sasha Vergez, Margaret V. McDonald, Sridevi Sridharan, Zoran Kostic, and Maxim Topaz. 2022. Is auto-generated transcript of patient-nurse communication ready to use for identifying the risk for hospitalizations or emergency department visits in home health care? a natural language processing pilot study. In *AMIA Annual Symposium Proceedings*, pages 992–1001. Published in 2023 as AMIA Annu Symp Proc. eCollection 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

A. Experiment Configurations

Table 2 details the full configuration for each experimental run.

Table 2: Experiment Configurations

run	fusion	full	STR	part	tfidf_s	emb_s	tfidf_f	emb_f	top_k	fs_k	agree	margin	router	committee	judge
A	Run B+D	No	No	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
B	None	No	No	Yes	NA	NA	0.3	0.7	60	5	0.6	0.15	NA	• gpt-5 • claude-sonnet-4.5 • gpt-o3 • gpt-oss-120b • gpt-4.1	gpt-5
C	Run B+F	No	No	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	None	No	No	No	0.3	0.7	0.3	0.7	50	5	0.6	0.1	30	• gpt-5 • claude-sonnet-4.5 • gpt-4.1	gpt-5
E	None	No	No	Yes	NA	NA	0.3	0.7	40	6	0.7	0.15	NA	• gpt-5 • claude-sonnet-4.5 • gemma-3-12b • gpt-o3 • gpt-oss-20b • gpt-4.1	NA
F	None	No	No	No	0.4	0.6	0.4	0.6	50	5	0.6	0.2	30	• gpt-5 • gpt-oss-120b • claude-sonnet-4.5	gpt-5
G	None	No	No	No	0.5	0.5	0.5	0.5	60	10	0.7	0.3	60	• gpt-5 • gemma-3-12b • gpt-o3 • gpt-oss-20b • gpt-4.1	gpt-5
H	None	Yes	No	No	NA	NA	NA	NA	NA	0	NA	NA	NA	• gpt-5	NA
I	None	No	Yes	No	0	1	0	1	50	3	NA	NA	30	• gpt-5	NA
J	None	No	Yes	No	1	0	1	0	50	3	NA	NA	20	• gpt-5	NA

Column definitions

Run-level: `run` (run ID); `fusion` (decision-level fusion rule).

Prompt / eligibility: `full` (full ontology vs. candidate subset); `STR` (STRING concepts inclusion); `part` (schema partitioning with domain experts).

Schema retrieval: `tfidf_s`/`emb_s` (sparse vs. dense weights); `top_k` (top schema candidates retained).

Few-shot retrieval: `tfidf_f`/`emb_f` (sparse vs. dense weights); `fs_k` (retrieved demonstrations inserted).

Voting / control: `agree` (minimum agreement to accept a (*concept_id*, *value*) pair); `margin` (minimum separation between top and runner-up support needed to go to adjudication); `router` (per-transcript budget after routing).

Models: `committee` (LLM models used for extraction and voting); `judge` (adjudication model for conflicts).