Lakefront AI Ramblers at MEDIQA-SYNUR 2026: Hybrid Retrieval and LLM Verification for Open-Source Schema-Guided Clinical Information Extraction

Michael T. Saban[1,2], Arsalan Yaghoubi[1], Behnaz Eslami[1,2], Samie Tootooni[2], Dmitriy Dligach[1]

[1]Department of Computer Science, Loyola University Chicago; [2]Department of Health Informatics and Data Science, Loyola University Chicago
[1]1052 W. Loyola Ave, Chicago, IL 60660; [2]2160 S. First Avenue, Maywood, IL 60153
{msaban, ayaghoubi, beslami, mtootooni, ddligach}@luc.edu

Abstract

Schema-constrained clinical information extraction requires identifying text-supported observations and outputting exact schema identifiers and values. In the MEDIQA-SYNUR 2026 shared task, synthetic nursing dictations were mapped to structured JSON outputs aligned with a 193-concept clinical schema under strict exact-match evaluation.

We extended the baseline pipeline, which consists of transcript segmentation, schema retrieval, and LLM-based extraction, with hybrid schema retrieval, supervised fine-tuning (SFT) of open-source LLMs, and LLM-based verification. Our hybrid retrieval approach combined dense embeddings with sparse BM25 representations using a convex combination strategy, improving schema coverage to 0.994 recall@60 on the development set. We evaluated GPT-4o, GPT-4o-mini, Llama-3-8B-Instruct, and Llama-3.3-70B-Instruct, applying LoRA-based SFT to open-source models.

On the official test set, our best submitted configuration (Llama-3.3-70B-Instruct-SFT with union voting and GPT-4o-mini verification) achieved 0.711 F1. Post-competition experiments showed that Llama-3-8B-Instruct-SFT (train + dev) reached 0.723 F1 under the same post-processing pipeline. For reference, GPT-4o achieved 0.791 F1 and did not benefit from post-processing. Performance differences across development and test splits further highlight the sensitivity of post-processing strategies to variation across split distribution. Overall, integrating high-recall retrieval, SFT, and LLM verification substantially narrows the performance gap between open- and closed-source models for schema guided clinical extraction.

Keywords: Clinical NLP, Information extraction, Large Language Models (LLMs), Open source, Schema alignment, Hybrid retrieval, Supervised fine-tuning

1. Introduction

Clinical documentation, such as nursing dictations, contains detailed information about patient status, interventions, clinical reasoning, and more (Corbeil et al., 2025; Demsash et al., 2023). However, it is often recorded as unstructured free text. Extracting structured information from these narratives is necessary for tasks such as clinical decision support, patient monitoring, and research (Sezgin et al., 2023).

This work is conducted in the context of the MEDIQA-SYNUR @ ClinicalNLP 2026 shared task, which provides a curated dataset of synthetic nursing dictations paired with gold-standard structured observations corresponding to a predefined extraction schema. Systems are evaluated based on the accuracy of structured JSON outputs at the observation and field levels, using exact and partial matching criteria that require correct concept identifiers, value representations, and correct value types (Corbeil et al., 2025; Michalopoulos et al., 2026). Performance is reported in terms of precision, recall, and F1 score, with systems ranked by overall F1. We follow the official task setup and evaluation protocol without modification.

The original pipeline performs schema filtering using cosine similarity within a retrieval augmented generation framework (Corbeil et al., 2025), essentially relying on dense semantic similarity for selection. Dense representations have become a primary method in retrieval systems (Lin and Lin, 2023). However, prior work has shown that standard dense retrievers can fall behind sparse methods such as BM25 in matching exact lexical cues and rare terminologies, which motivates hybrid approaches that incorporate both semantic and lexical signals (Chen et al., 2022). Hybrid retrieval combining sparse lexical and dense semantic similarity has been studied in information retrieval, where convex combinations of scores consistently outperform either component individually (Bruch et al., 2023; Lin and Lin, 2023; Mandikal and Mooney, 2024). Motivated by this, we apply a convex combination of BM25 and dense embeddings to the schema filtering stage of the MEDIQA-SYNUR pipeline.

Beyond this, we extend the pipeline through supervised fine-tuning of open-source LLMs and evaluation of post-processing strategies, including voting and LLM-based verification. We show that hybrid retrieval substantially improves schema item recall during filtering, while LLM-based verification provides the largest

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 212–221

improvements in extraction F1. Together, these components significantly improve structured extraction performance and improve the competitiveness of open-source models in schema-guided clinical information extraction.

2. Methods

We build on the shared task's schema-guided extraction formula and baseline approach by incorporating hybrid schema retrieval and model- and post-processing-level enhancements (**Figure 1**). Each transcript is first segmented into subsets, then processed independently per segment using retrieval-augmented generation (RAG) to select a small subset of schema concepts. A large language model (LLM) produces schema-constrained JSON extractions, which are always followed by deterministic validation. Voting and LLM-based verification are optional experimental post-processing steps that were applied only in the configurations reported for those ablations.

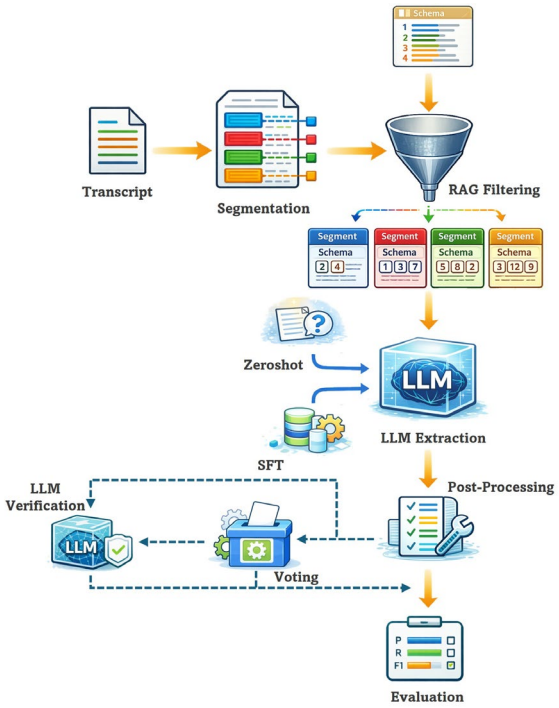

Figure 1. Overview of the schema-guided extraction pipeline. Solid lines indicate the default inference path, while dashed lines denote optional post-processing steps.

2.1 Transcript Segmentation

The MEDIQA-SYNUR transcripts are formatted with repeated "[Clinician]" tags before each span. In this dataset, every span is associated with a single clinician role. We apply a lightweight rule-based segmentation that splits transcripts at each "[Clinician]" tag. Each resulting span is treated as an independent segment for retrieval and extraction.

2.2 Schema-Guided Retrieval

The official schema contains 193 concepts, making it inefficient to include all concepts in-context for every segment. To address this, we use retrieval to select a subset of candidate schema items for each segment.

Dense embeddings are effective for capturing semantic similarity but may miss short lexical signals common in clinical terms (e.g. "Fall risk"). To address this, we adopt a hybrid retrieval strategy that combines dense and sparse representations (Bruch et al., 2023). For a given transcript segment x and schema item s, the final relevance score is calculated as a convex combination of normalized dense and sparse similarity scores:

$$\text{score}(x,s) = \alpha \cdot \hat{S}_{\text{dense}}(x,s) + (1-\alpha) \cdot \hat{S}_{\text{sparse}}(x,s)$$

where $\hat{S}$ denotes min-max normalized similarity scores and $\alpha \in [0,1]$ controls the relative weighting of semantic versus lexical similarity. For each segment, we select the top-N Schema items and provide them to the extraction model as constrained context.

2.3 LLM-Based Extraction

Given a transcript segment and its retrieved schema subset, we prompted an LLM to extract structured clinical observations in a strict JSON format. The prompt consisted of a system instruction defining the model's role as an electronic health record (EHR) flowsheet extraction expert and specifying detailed constraints on schema compliance. These constraints included extracting only explicitly stated or clearly implied observations, matching values exactly to schema-defined enumerations when applicable, returning numeric fields without units, formatting multi-select fields as arrays, and avoiding any generation of unsupported observations. The model was instructed to return a JSON object containing an array of observations with the required fields. This extraction step is model-agnostic and is evaluated with both API-based and open-source instruction-tuned LLMs.

2.4 Supervised Fine-Tuning

To better adapt open-source LLMs to schema-constrained extraction, we applied supervised fine-tuning (SFT) using low-rank adaptation (LoRA) (Hu et al., 2022). A challenge is that gold observations are provided at the transcript level, while segmentation is performed afterward for retrieval and extraction. The dataset does not provide alignment between individual observations and specific transcript segments, making segment-level supervision ambiguous. To avoid ambiguous segment-level supervision, we fine-tuned using transcript-level targets with a fixed retrieval context size (top-N) per transcript. At inference time, the same model is applied per segment.

2.5 Post-processing and Validation

Raw LLM outputs can contain formatting errors and unsupported (hallucinated) observations. We applied the following steps:

2.5.1 Schema Validation

The first stage involved deterministic validation and normalization. Model outputs were parsed and checked against the official schema, removing malformed JSON, discarding observations with invalid schema identifiers, and normalizing categorical values to match the allowed set of schema-defined options. For multi-select fields, single values were converted to lists when necessary, and invalid selections were removed.

2.5.2 Voting

The voting procedure was related to self-consistency decoding, which involves sampling multiple stochastic generations from a single model and aggregating them to reduce variance, although we applied it to schema-constrained information extraction rather than chain-of-thought reasoning (Wang et al., 2022).

Specifically, we ran extraction multiple times at a non-zero temperature and aggregated the resulting observation sets. We considered both majority voting, which retains observations only appearing in multiple runs, and union voting, which retains any observation produced across at least one run.

2.5.3 LLM Verification

To reduce unsupported or hallucinated extractions, we introduced a verification stage using a separate LLM. For each transcript segment, the verifier was provided with the original transcript and the set of extracted observations. The model was prompted to assess each observation and return a structured JSON mapping from observation ID to a binary decision ("KEEP" or "REMOVE"). Observations marked "REMOVE" were discarded from the final output. This stage operated independently from the extractor model and did not modify observations that were kept.

3. Experimental Setup

This section describes the datasets, configurations, and evaluation protocol used to assess the system.

3.1 Data

We evaluated our approach on the MEDIQA-SYNUR 2026 dataset, which contains synthetic nursing dictations annotated with gold structured observations derived from a 193-concept schema. The official split includes 122 training transcripts, 101 development transcripts, and 199 test transcripts.

3.2 Model Configurations

We evaluated four extractor LLMs for the extraction stage: (1) *GPT-4o*, (2) *GPT-4o-mini (Achiam et al., 2023)*, (3) *Llama-3-8B-Instruct*, and (4) *Llama-3.3-70B-Instruct (Grattafiori et al., 2024).* For the open-source models, we applied SFT using LoRA when specified. Fine-tuning was performed on the training set for development experiments. For final test submissions, fine-tuning was performed on the combined training and development sets to maximize exposure. Verification models were evaluated separately and could differ from the extraction model. All model comparisons were performed on the development set unless otherwise noted.

Supervised fine-tuning was performed using HuggingFace Transformers (Wolf et al., 2020) and PEFT with LoRA (Hu et al., 2022). The 70B models used 4-bit quantization using QLoRA (Dettmers et al., 2023) to fit in GPU memory, while the 8B models used standard LoRA without quantization. We evaluated multiple fine-tuning configurations across model sizes and training splits. Hyperparameter settings for model configurations are summarized in **Table 1**. Training was performed on a single NVIDIA H200 GPU.

Parameter	**8B**	**70B**
Epochs	5	3
LR	2e-5	1e-5
LoRA r	32	16
LoRA α	64	32
Max Len	4096	4096
Quant	bf16	4-bit

Table 1. SFT hyperparameters for 8B and 70B models. Identical configurations were used for train-only and train + dev runs within each model size.

3.3 Schema Retrieval Configurations

We compared five dense embedding models for schema retrieval: (1) *all-MiniLM-L6-v2*, (2) *all-MiniLM-L12-v2 (Wang et al., 2020)*, (3) *all-mpnet-base-v2 (Song et al., 2020)*, (4) *BAAI/bge-base-en-v1.5 (Xiao et al., 2024)*, and (5) OpenAI's *text-embedding-3-small*. Sparse retrieval was implemented using BM25 (Robertson and Zaragoza, 2009). We evaluated hybrid dense-sparse retrieval (Bruch et al., 2023) across alpha values. Based on development set recall, we selected *BAAI/bge-base-en-v1.5* with α=0.6. For each segment, the top N=60 schema items were retrieved and used as context for extraction.

3.4 Post-processing Configurations

All runs involved deterministic schema validation. Voting and verification were evaluated as optional

post-processing steps and were only applied in configurations explicitly labeled with those steps. For voting, we performed three independent extraction runs with a temperature of T=0.3. We evaluated both majority voting (threshold ≥ 2/3) and union voting (threshold ≥ 1/3). For the LLM verification step, we evaluated *GPT-4o-mini* and *Llama-3.3-70B-Instruct* as verifiers. Verification was applied after schema validation and optional voting.

3.5 Evaluation Metrics

We report precision, recall, and F1 using the official MEDIQA-SYNUR evaluation script. A predicted observation was considered correct if both the schema concept ID and the extracted value exactly matched the reference annotation. All development set experiments were evaluated locally, while test set results were obtained from the official leaderboard.

3.6 Infrastructure

Closed-source models were run via API on a local workstation. Open-source inference and fine-tuning were run on an HPC cluster equipped with NVIDIA H200 GPUs under a consistent software environment.

4. Results

We evaluated the system on the MEDIQA-SYNUR 2026 development set and report official leaderboard results on the test set where available. Unless otherwise noted, all ablation studies were conducted on the development set.

4.1 Schema Retrieval Performance

Table 2 reports schema retrieval recall@60 on the development set, measured as the fraction of gold schema items covered by the top-60 retrieved candidates (per segment). Across all dense retrievers, hybrid dense + sparse retrieval improved coverage relative to dense-only retrieval. The BM25 + BGE-base hybrid achieved the highest recall@60 (0.994), and this configuration was used for all subsequent experiments.

Retriever	Recall@60
MiniLM-L6	0.953
MiniLM-L12	0.960
MPNet	0.938
TE3-small	0.962
BGE-base	0.974
BM25	0.900
Hybrid (TE3-small+BM25)	0.985
Hybrid (BGE-base+BM25)	0.994

Table 2. Development set schema retrieval coverage (recall@60) for dense-only, sparse (BM25), and hybrid dense + sparse retrieval. Recall@60 measures the fraction of gold schema concepts covered by the top-60 retrieved candidates. Hybrid retrieval uses α=0.6.

4.2 Extraction Performance

Table 3 shows development set extraction performance under single-run inference. GPT-4o achieved the highest F1 (0.845), followed by GPT-4o-mini (0.803). Among the open-source models, Llama-3.3-70B-Instruct improved with SFT (0.765 vs. 0.742). The smaller Llama-3-8B-Instruct model benefited substantially from SFT (0.614 vs. 0.549).

Model	Precision	Recall	F1
GPT-4o	0.841	0.850	0.845
GPT-4o-mini	0.826	0.782	0.803
L3-70B-SFT	0.743	0.787	0.765
L3-70B	0.702	0.788	0.742
L3-8B-SFT	0.497	0.804	0.614
L3-8B	0.498	0.611	0.549

Table 3. Extraction performance on the development set under single-run inference. L3-70B and L3-8B denote Llama-3.3-70B-Instruct and Llama-3-8B-Instruct, respectively; SFT indicates supervised fine-tuning with LoRA.

4.3 Post-processing Effects

Table 4 summarizes the best-performing post-processing strategies for each model on the development set. Overall, LLM verification provided the most consistent improvements across models, typically improving precision while maintaining recall. Voting alone provided smaller, model-dependent gains, with majority voting yielding modest improvements and union voting increasing recall at a large cost of precision. Notably, verification increased Llama-3.3-70B-Instruct-SFT from 0.765 to 0.812 F1, and Llama-3-8B-Instruct-SFT from 0.614 to 0.776 F1. **Figure 2** visualizes precision-recall tradeoffs across all evaluated configurations. The combination of voting and LLM verification produced results similar to LLM verification by itself.

Model	GPT-4o-mini	L3-70B-SFT	L3-8B-SFT
Baseline	0.803	0.765	0.614
Best Voting	0.813	0.799	0.69
Best Verify	0.817	0.812	0.776
Union + Verify	0.82	0.8	0.723
Majority + Verify	0.818	0.811	0.77

Table 4. Summary of development set extraction performance (F1) for the single-run baseline and the best-performing voting, verification, and combined post-processing configurations. L3-70B and L3-8B denote Llama-3.3-70B-Instruct and Llama-3-8B-Instruct, respectively; SFT indicates supervised fine-tuning with LoRA.

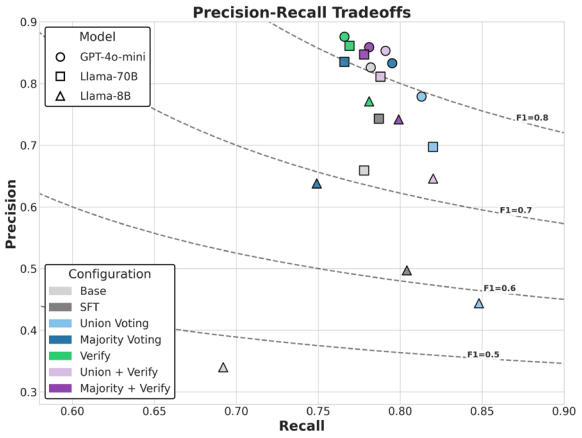

Figure 2. Precision-recall tradeoffs for all models and post-processing configurations on the development set. Shapes indicate model scale; colors indicate configuration; dashed curves show iso-F1 contours.

4.4 Test Set Results

Following the shared task, the organizers released gold labels for the test set, which enabled additional post-hoc evaluation beyond the official leaderboard submissions. **Table 5** reports the performance of our official submissions, all based on Llama-3.3-70B-Instruct-SFT (train + dev) with various post-processing strategies. Relative to the base SFT system (F1 0.680), both voting and verification improved performance, with union + GPT-4o-mini verification achieving the highest F1 among the submitted configurations (0.711).

Table 6 reports additional experiments conducted after the competition using the released test labels. These include base inference with GPT-4o, GPT-4o-mini, and Llama-3-8B-Instruct-SFT (train + dev), along with verification-only and voting + verification variants of the SFT model. These experiments were run once to characterize how post-processing transfers to the test distribution.

Across experiments, voting and verification remained beneficial overall, but relative performance differed across the development and test sets. In particular, the combined voting + verification increased performance more on the test set than verification-only.

Configuration	Precision	Recall	F1
L3-70B-SFT	0.723	0.643	0.680
+ Majority Vote	0.742	0.646	0.691
+ Union Vote	0.648	0.671	0.660
+ Union + GPT-4o-mini Verify	0.768	0.661	0.711
+ Majority + GPT-4o-mini Verify	0.802	0.637	0.710

Table 5. Official test set performance for configurations submitted to the MEDIQA-SYNUR competition. Results are reported for Llama-3.3-70B-Instruct-SFT (train + dev).

Configuration	Precision	Recall	F1
GPT-4o	0.768	0.816	0.791
GPT-4o-mini	0.797	0.586	0.675
L3-8B-SFT	0.733	0.656	0.693
L3-8B-SFT + L3-70B	0.764	0.652	0.704
L3-8B-SFT+ Union + GPT-4o-mini Verify	0.762	0.688	0.723

Table 6. Additional test set performance for configurations run after the MEDIQA-SYNUR competition.

5. Discussion

This study demonstrates that performance in schema-constrained clinical extraction is not exclusively dictated by raw language modeling capacity, but is also influenced by retrieval coverage, model decision supervision, and post-processing. Across our experiments, we found that schema selection defines the feasible context, fine-tuning improves overall output, and verification provides a more stable method of correcting LLM output than stochastic voting.

5.1 Task Difficulty in Schema-Constrained Clinical Extraction

The MEDIQA-SYNUR task highlights the difficulty of schema-constrained extraction because systems must do more than identify clinically relevant content in free text. They must also normalize each extracted observation to a schema by selecting the exact concept identifier and outputting a valid value representation (Fu et al., 2020). Under exact match scoring, errors in unsupported observations, wrong concept IDs, or invalid formatting are all penalized, which makes the evaluation very sensitive to hallucinations and variance in output. This is consistent with how exact-match criteria are usually harsher than overlap-based matching in clinical extraction evaluations (Hein et al., 2025).

The precision-recall tradeoffs across models and post-processing strategies (**Figure 2**) reflect this issue, where improvements in recall come at the cost of including unsupported observations, while aggressively filtering predicted observations increases precision but risks discarding valid clinical observations.

5.2 Effectiveness of Hybrid Schema Retrieval

Schema retrieval is critical in this task, as it determines the context available to the extractor LLM. The success of this step represents an upper bound on extraction performance, as concepts not retrieved in this step cannot be predicted, regardless of model capacity. While hybrid retrieval achieved near perfect recall@60, proving the benefit of combined dense and sparse representations (Bruch et al., 2023; Chen et al., 2022; Gao et al., 2021), extraction F1 was still low. This gap suggests that high schema coverage alone is insufficient, and that retrieval must balance coverage and precision to avoid overwhelming the LLM with irrelevant concepts.

5.3 Model Scaling and SFT Effects

Model scale is generally associated with improved performance in language modeling (Kaplan et al., 2020) and closed-source models such as *GPT-4o* provide a practical upper bound for this pipeline.

However, our findings indicate that scale is not the only determinant of performance under schema-constrained extraction. While *Llama-3.3-70B-Instruct* showed good out-of-box performance and further benefitted from SFT, the 8B model had very competitive results after fine-tuning. Interestingly, when both models were fine-tuned on the combined train + dev set, the 8B model achieved comparable or slightly higher test F1 than the 70B model. While this could suggest that effective fine-tuning can compensate for parameter count differences, we did not conduct hyperparameter optimization across model sizes. This would be supported by further evaluation of model scale under matched hyperparameter searches to isolate the effect of model size.

5.4 Verification vs Voting

Across development set experiments, LLM-based verification was the most consistent post-processing strategy, typically improving precision while maintaining recall. Voting resulted in more variable effects, where majority voting modestly improved performance for some models, but union voting traded precision for minor improvements in recall.

From a practical standpoint, verification is computationally more efficient. The task itself is much smaller compared to the full extraction prompts, so it requires fewer tokens and can be performed using a smaller or cheaper verifier model. Voting requires multiple full extraction passes, making it computationally expensive and less predictable.

Importantly, both strategies improved performance relative to single-run extraction. However, neither approach was completely superior across all settings. The benefit of verification and voting appears to depend on model size, data distribution, and task characteristics. Broader evaluation across datasets and extraction settings Is required to determine how consistently these patterns generalize.

5.5 Open-source competitiveness and deployment relevance

A key finding in this work is that open-source systems can approach or exceed the performance of some closed-source models when combining fine-tuning and verification. On the development set, *Llama-3.3-70B-Instruct* with SFT and verification surpassed the *GPT-4o-mini* baseline and approached *GPT-4o* performance. These results indicate that while proprietary models such as GPT-4o have strong performance without additional supervision, the performance gap between open and closed models narrows noticeably when retrieval, fine-tuning, and verification are optimized for open-source models. This is important when considering the advantages open-source provides in healthcare settings.

Open-source models allow for local deployment, better transparency, and reproducibility, which are all crucial in healthcare settings that are guided by regulatory and infrastructure constraints (Bernal and Mazo, 2022; Riedemann et al., 2024). Open-source alternatives are increasingly viable for schema-guided clinical extraction when combined with the appropriate retrieval and post-processing strategies.

5.6 Generalization Gap on the Test Set

Despite the strong development set performance, test set results had a roughly 0.1 F1 drop, which was primarily due to reduced recall, as precision remained relatively high. Several factors may contribute to this gap. First, differences in annotators across data may create distributional differences across data splits in aspects such as dictation style, concept frequency, or schema-value combinations. Second, the post-processing strategies in this pipeline inherently favored precision, so there were limited opportunities to recover missed observations. Third, the supervision mismatch introduced by transcript-level fine-tuning may have disproportionately affected segment-level evaluation. Gold labels were aligned at the transcript-level, while inference operates per-segment, which increased the likelihood that relevant schema items were excluded from a transcript's top-60 retrieval context even if they appeared in the gold labels.

The generalization gap was not limited to the overall decrease in F1, as the effectiveness of post-processing strategies also shifted. The ranking of verifier models shifted between development and test, and the magnitude of performance from voting + verification differed across these splits. Additionally, while the 70B model clearly outperformed the 8B model when trained on train-only data and evaluated on development, this reversed when both were trained on train + dev data and evaluated on test. Given these results, we can assume that these post-processing strategies and model scaling effects are sensitive to distributional differences and that configuration choices validated on development data may not transfer to unseen data. Post-processing should be viewed as generally beneficial but not universally optimal.

5.7 Future Work

Future work could focus on reducing retrieval noise through adaptive retrieval by dynamically truncating the candidate schema set per segment using score-based cutoffs. Given the high recall@60 achieved by hybrid retrieval, truncation could preserve coverage while providing a more exact context for extraction and reducing the chance of hallucinations. Segment-aligned supervision via heuristics or generated 'silver' segment labels could improve fine-tuning. Verification could be extended with calibrated confidence thresholds, ensemble verifiers, or "unknown" options rather than hard rejection. A broader exploration of newer instruction-tuned models and domain-specific verifiers could close the gap between open- and closed-source systems. Finally, external evaluation on real nurse dictations or other real-world clinical speech transcripts to assess generalizability beyond the synthetic shared-task setting is an important next step.

6. Conclusion

In this work, we investigated schema-constrained information extraction in the MEDIQA-SYNUR 2026 shared task, analyzing how hybrid schema retrieval, supervised fine-tuning, and post processing strategies interact under strict exact-match evaluation. Hybrid dense + sparse retrieval was a critical prerequisite, achieving near-perfect schema coverage, but it was not sufficient on its own. Downstream modeling choices and validation mechanisms play equally important roles in determining final extraction quality.

On the development set, LLM-based verification was the most reliable post-processing strategy, consistently improving precision while preserving overall recall. However, test-set analyses with released labels showed that the contributions of post-processing strategies can change across splits, suggesting sensitivity to distributional differences and post-processing configuration.

Overall, these results demonstrate that effective schema-constrained clinical extraction depends on tightly integrated retrieval, fine-tuning, and post-processing rather than model scale alone. Future work should focus on reducing retrieval noise, improving segment-aligned supervision, conducting systematic hyperparameter selection across model sizes, and calibrated verification strategies to improve system robustness under distribution shift and across related extraction tasks.

7. Limitations

There are several limitations worth noting. First, the dataset consists of synthetic nursing dictations provided as part of the shared task, so we were not able to evaluate on real-world nurse dictations. As a result, the generalizability of these findings to real clinical speech and documentation workflows is uncertain, as synthetic data may not completely capture the variability and ambiguity present in real-world clinical documentation. Although the schema and annotations were carefully constructed, inconsistencies in value formatting and concepts can still affect exact-match evaluation. Second, supervised fine-tuning relied on transcript-level targets due to the lack of segment-aligned annotations, which may limit the effectiveness of segment-level inference. Third, we did not perform any systematic hyperparameter optimization. These results should be interpreted as exploratory rather than as a fully tuned upper bound. Additional limitations include the fixed retrieval context size ($N=60$) and limited voting depth (three runs).

8. Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM012973 and by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number 5R01HL173037-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

9. Bibliographical References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bernal, J., & Mazo, C. (2022). Transparency of artificial intelligence in healthcare: insights from

professionals in computing and healthcare worldwide. *Applied Sciences*, *12*(20), 10228.

Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, *42*(1), 1-35.

Chen, X., Lakhotia, K., Oguz, B., Gupta, A., Lewis, P., Peshterliev, S., Mehdad, Y., Gupta, S., & Yih, W.-t. (2022). Salient phrase aware dense retrieval: can a dense retriever imitate a sparse one? In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 250-262.

Corbeil, J.-P., Abacha, A. B., Michalopoulos, G., Swazinna, P., Del-Agua, M., Tremblay, J., Daniel, A. J., Bader, C., Cho, K., & Krishnan, P. (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, pages 859-870.

Demsash, A. W., Kassie, S. Y., Dubale, A. T., Chereka, A. A., Ngusie, H. S., Hunde, M. K., Emanu, M. D., Shibabaw, A. A., & Walle, A. D. (2023). Health professionals' routine practice documentation and its associated factors in a resource-limited setting: a cross-sectional study. *BMJ health & care informatics*, *30*(1), e100699.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088-10115.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., & Wen, A. (2020). Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, *109*, 103526.

Gao, L., Dai, Z., & Callan, J. (2021). COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., & Vaughan, A. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hein, D., Christie, A., Holcomb, M., Xie, B., Jain, A., Vento, J., Rakheja, N., Shakur, A. H., Christley, S., & Cowell, L. G. (2025). Prompts to table: specification and iterative refinement for clinical information extraction with large language models. *medRxiv*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), pages.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lin, S.-C., & Lin, J. (2023). A dense representation framework for lexical and semantic matching. *ACM Transactions on Information Systems*, *41*(4), 1-29.

Mandikal, P., & Mooney, R. (2024). Sparse meets dense: A hybrid approach to enhance scientific document retrieval. *arXiv preprint arXiv:2401.04055*.

Michalopoulos, G., Corbeil, J.-P., Bader, C., Bodenstab, N., & Ben Abacha, A. (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, pages, Palma, Mallorca, Spain.

Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. *npj Digital Medicine*, *7*(1), 339.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333-389.

Sezgin, E., Hussain, S.-A., Rust, S., & Huang, Y. (2023). Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. *JMIR Formative Research*, *7*, e43014.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, *33*, 16857-16867.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38-45.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J.-Y. (2024). C-Pack: Packed resources for general chinese embeddings. In Proceedings of the 47th international ACM SIGIR conference on research and

development in information retrieval, pages 641-649.

Appendix A: Detailed Experimental Results

This appendix provides the full experimental results summarized in Table 4 of the main text. All results are reported on the development set unless otherwise noted.

A.1 Voting Strategies

Model	P	R	F1
GPT-4o-mini-M	0.836	0.786	0.810
GPT-4o-mini-U	0.824	0.796	0.810
L3-70B-SFT-M	0.750	0.791	0.770
L3-70B-SFT-U	0.648	0.805	0.718
L3-8B-SFT-M	0.546	0.816	0.654
L3-8B-SFT-U	0.449	0.837	0.585

Table A1. Effect of voting strategies on extraction performance (development set).

A.2 LLM Verification

Model	Verifier	P	R	F1
GPT-4o	GPT-4o	0.865	0.823	0.844
GPT-4o	GPT-4o-mini	0.865	0.827	0.846
GPT-4o	L3-70B	0.863	0.827	0.845
GPT-4o	L3-8B	0.859	0.825	0.841
GPT-4o-mini	GPT-4o-mini	0.870	0.763	0.813
GPT-4o-mini	L3-70B	0.876	0.766	0.817
GPT-4o-mini	L3-8B	0.868	0.756	0.808
L3-70B-SFT	GPT-4o-mini	0.835	0.766	0.799
L3-70B-SFT	L3-70B	0.861	0.769	0.812
L3-70B-SFT	L3-8B	0.854	0.752	0.800
L3-8B-SFT	GPT-4o-mini	0.698	0.786	0.739
L3-8B-SFT	L3-70B	0.771	0.781	0.776
L3-8B-SFT	L3-8B	0.712	0.762	0.736

Table A2. Extraction performance with and without LLM verification on the development set.

A.3 Combined Voting + Verification

Model	P	R	F1
GPT-4o-mini-U	0.853	0.791	0.820
GPT-4o-mini-M	0.859	0.781	0.818
L3-70B-SFT-U	0.811	0.788	0.800
L3-70B-SFT-M	0.847	0.778	0.811
L3-8B-SFT-U	0.646	0.820	0.723
L3-8B-SFT-M	0.742	0.799	0.770

Table A3. Pipeline performance for selected configurations on the development set. Llama-3.3-70B-Instruct was used as the verifier in all configurations.

Appendix B: LLM Prompts

This appendix provides the complete prompts used for LLM-based extraction and verification, as described in sections 2.3 and 2.5.2 of the main text. Template variables (scehma, transcript, observations) are replaced at runtime with the corresponding inputs.

B.1 Extraction Prompts

System Prompt

```
You are an expert at medical electronic health record (EHR) flowsheet analysis.

Your task is to extract clinical observations from nurse dictation transcripts and structure them according to a provided schema.

Guidelines:
1. Extract ONLY observations that are explicitly mentioned or clearly implied in the transcript
2. Match extracted values to the schema's value enum when applicable (use exact matches)
3. For NUMERIC types, extract only the numeric value (no units)
4. For MULTI SELECT types, return an array of applicable values
5. For STRING types, extract the relevant text as mentioned
6. For SINGLE SELECT types, choose the single best matching option from value enum
7. Do NOT fabricate or infer observations that are not in the transcript
8. If unsure, do NOT include the observation

Output Format:
Return a JSON object with key "observations" containing an array. Each observation must have:
- "id": The schema concept ID (string)
- "name": The schema concept name (string)
- "value type": The type from schema (string)
- "value": The extracted value (type depends on value type)

For MULTI SELECT, value should be an array of strings.
For NUMERIC, value should be a number.
For STRING and SINGLE SELECT, value should be a string.
```

User Prompt

```
SCHEMA:
{schema}

TRANSCRIPT:
{transcript}

Extract all clinical observations from the
TRANSCRIPT that match concepts in the SCHEMA.
Return a JSON object with key "observations"
containing an array of extracted
observations.

OUTPUT:
```

B.2 Verification Prompts

System Prompt

```
You are a clinical documentation
verification expert.

Your task is to verify whether extracted
clinical observations are actually supported
by the source nursing transcript.

For each observation, determine:
- KEEP: The observation is directly stated
or clearly implied in the transcript
- REMOVE: The observation is NOT supported,
was incorrectly inferred, or fabricated

Be conservative - only mark REMOVE if you are
confident the observation is not supported.
Small variations in wording are acceptable
if the meaning matches.
```

User Prompt

```
TRANSCRIPT:
{transcript}

EXTRACTED OBSERVATIONS TO VERIFY:
{observations}

For each observation ID, respond with either
"KEEP" or "REMOVE".
Return a JSON object mapping observation IDs
to decisions.

Example response format:
{"40": "KEEP", "56": "REMOVE", "89": "KEEP"}

Your verification:
```