Team MKC at MEDIQA-SYNUR 2026: Retrieval-Augmented Generation Based Nurse Observation Extraction

Kyomin Hwang[1], Nojun Kwak[1,2†]
[1]GSCST, Seoul National University, [2]AIIS, Seoul National University
{kyomin98, nojunk}@snu.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) have played a significant role in reducing human workload across various domains, a trend that is increasingly extending into the medical field. In this paper, we propose an automated pipeline designed to alleviate the burden on nurses by automatically extracting clinical observations from nurse dictations. To ensure accurate extraction, we introduce a method based on Retrieval-Augmented Generation (RAG). Our approach demonstrates effective performance, achieving an F1-score of 0.796 on the MEDIQA-SYNUR test dataset.

Keywords: Clinical Observation Parsing, Retrieval-Augmented Generation

1. Introduction

Recent advancements in Large Language Models (LLMs), exemplified by the GPT series (Brown et al., 2020), have accelerated a paradigm shift across a multitude of domains, demonstrating remarkable capabilities in natural language understanding and generation. This momentum has profoundly impacted the medical field (Kim et al., 2024a), leading to the development of specialized models such as MedGemma (Sellergren et al., 2025) and LLaVA-Med (Li et al., 2023a). These models have achieved state-of-the-art performance on various biomedical tasks by leveraging large-scale, domain-specific datasets. However, obtaining such high-quality annotated data remains a significant bottleneck due to privacy concerns and the high cost of expert annotation. In this context, the SYNUR (Corbeil et al., 2025) dataset has established a crucial benchmark, facilitating the automatic extraction of clinical observations from nursing dictations—a task essential for streamlining clinical documentation.

Building upon these developments, we propose a novel Retrieval-Augmented Generation (RAG) pipeline designed to accurately extract clinical observations from complex nurse dictations. Unlike conventional approaches that rely on resource-intensive fine-tuning, our automated pipeline integrates a synergistic dual-retrieval mechanism. First, we employ an ontology-based retrieval system to fetch the most relevant medical concepts based on the current utterance, ensuring terminological precision. Second, we utilize a memory bank to retrieve semantically similar dialogue segments and their corresponding gold-standard observations, enabling the model to learn extraction patterns via in-context learning. This dual approach equips the foundation LLM with both explicit medical context and structural guidance. Utilizing this training-free framework, we achieved an F1-score of 0.796 on the MEDIQA-SYNUR test set. These results demonstrate that our approach can effectively alleviate the nursing documentation burden by automating the interpretation of patient interactions with high reliability.

† Corresponding author.

2. Related Works

2.1. Foundation model in Medical Domain

Recent advancements in large-scale data acquisition and the exponential growth of computing power have led to the emergence of foundation models capable of performing diverse tasks within a unified framework (Jang et al., 2023; Brown et al., 2020). This paradigm has significantly influenced the medical domain, triggering the development of specialized foundation models. Early efforts focused on encoder-based models like ClinicalBERT (Huang et al., 2019) and BlueBERT (Peng et al., 2019), which are optimized for medical knowledge representation and retrieval. More recently, generative and multi-modal models such as LLaVA-Med (Li et al., 2023a) and MedGemma (Sellergren et al., 2025) have been introduced to handle complex biomedical reasoning. Currently, active research is dedicated to adapting these generalist foundation models to enhance their performance on specific clinical tasks (Li et al., 2023b; Toma et al., 2023). Aligning with this research process, this paper proposes a method for developing a Large Language Model (LLM) specifically optimized for analyzing medical observations within nurse dictations.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 66–72
16 May 2026.

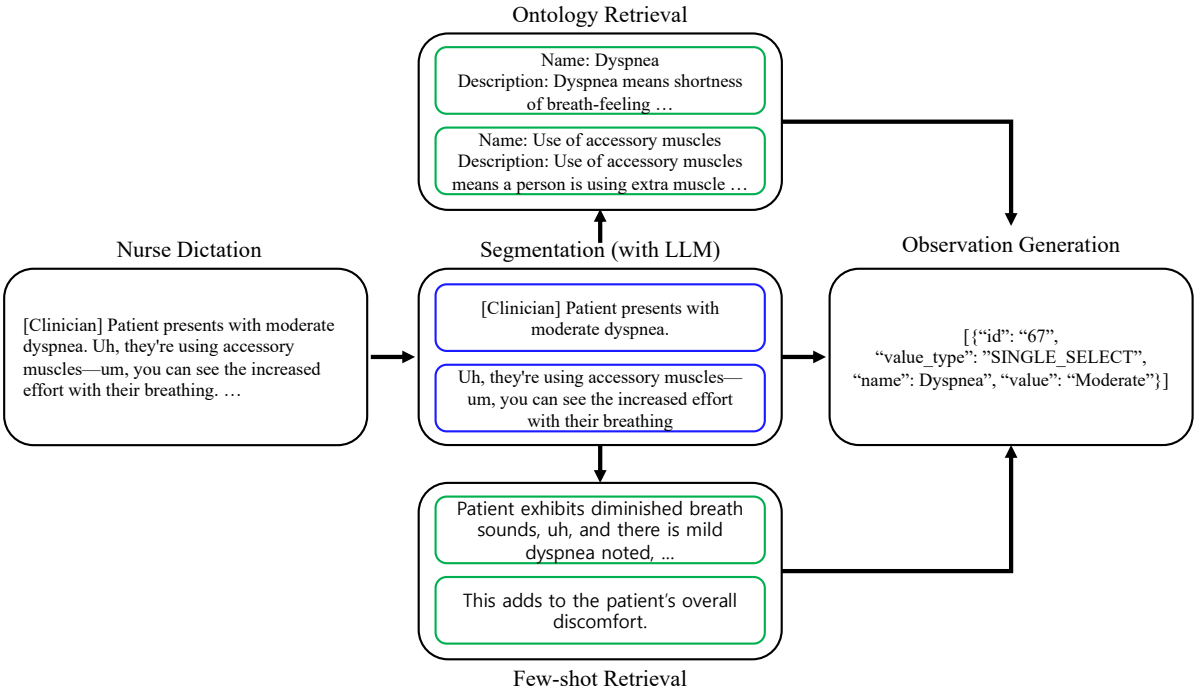

Figure 1: Full illustration of Retrieval-Augmented Generation (RAG) Based Nurse Observation Extraction Pipeline

2.2. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Kim et al., 2024b) has emerged as a prominent approach for tailoring foundation models to specific tasks. While foundation models possess vast general knowledge, they often struggle with domain-specific nuances and factual accuracy in critical fields like healthcare. RAG addresses this limitation by grounding the model's responses in retrieved evidence. Previous studies have demonstrated that utilizing RAG techniques can significantly enhance task-specific performance across various domains without the need for additional training of the foundation models (Chen et al., 2022; Kim et al., 2024b). Leveraging this capability to handle complex clinical terminologies and contexts, we propose a RAG-based automatic generation pipeline designed to specialize a foundation LLM for extracting medical observations from nurse dictations, all without requiring further model training.

3. Method

In this section, we present a pipeline designed to automatically extract clinical observations from nurse dictations. Figure 1 illustrates our medical observation pipeline.

3.1. Nurse Dictation Segmentation

Nurse dictations processed by LLMs are typically lengthy. Therefore, following the approach proposed by (Corbeil et al., 2025), we first segment the input dictation into distinct units containing clinical facts. The prompt used for this segmentation is shown in Figure 2.

Each nurse dictation is divided into multiple segments. For each segment, we generate observations by providing the LLM with two types of retrieved context: 1) relevant schemas retrieved from the ontology, and 2) semantically similar segments retrieved from the training dataset. The latter serves as few-shot examples, consisting of segment-observation pairs, to guide the LLM in generating the correct observation for the input segment.

3.2. Description for Medical Observation in Ontology

First, we recognized that directly inputting medical observations defined within an ontology (e.g., weight-bearing status) into an LLM could lead to performance degradation. Consequently, we implemented a process to append descriptions to each medical observation. Figure 3 illustrates the prompt used to generate these descriptions.

3.3. Retrieving Relevant Schemas from Ontology

To enhance the accuracy of observation extraction, we identify schemas relevant to the input segment from the ontology and provide them as input to the

Prompt for Nurse Dictation Segmentation

You are an expert clinical documentation assistant.

Given an input TRANSCRIPT of a nurse's observations about a patient, divide the TRANSCRIPT into contiguous SEGMENTS based on clinical facts.

A clinical fact refers to specific, verifiable information related to the health of a patient (symptoms, vitals, exam findings, interventions, response to treatment, labs/imaging results, functional status, safety events, etc.).

SEGMENTATION RULES:
- Output segments MUST preserve the original order.
- Segments MUST be contiguous chunks of the original transcript (do NOT reorder).
- Do NOT paraphrase; keep the segment text as close to verbatim as possible.
- Include all information from the transcript (no omissions).
- Split when the clinical topic/fact changes (new symptom, new vital sign, new intervention, new assessment finding, new timeframe, etc.).
- If multiple sentences describe one tightly-related clinical fact, keep them in the same segment.
- If the transcript contains speaker tags (e.g., [Nurse], [Patient]) keep them as they appear.

OUTPUT FORMAT:
Return ONLY a valid JSON object with exactly one key:
{
"segments": [
"<segment 1>",
"<segment 2>",
...
]
}

No additional keys. No extra text outside JSON.

TRANSCRIPT:

{NURSE DICTATION}

Figure 2: Prompt for nurse dictation segmentation

Description generation prompt

You are a helpful and professional medical assistant.
Your task is to explain the given medical assessment item (Name)
so that a patient or a layperson can easily understand it.
Provide the explanation using simple and friendly vocabulary.
Keep the explanation concise (1-2 sentences) and strictly avoid technical jargon.
Explain the meaning of '{term}'

Figure 3: Prompt for description generation

LLM. For retrieval, we employ a hybrid approach utilizing both BlueBert (Peng et al., 2019) and TF-IDF. We format each schema using its name, description and its value enumerations (options). The input format is as follows:

name. Description Options: Option1, Option2 (1)

We select the top 10 most relevant schemas to include in the LLM input.

3.4. Retrieving Few-Shot Examples from Training Dataset

In addition to schemas, we utilize few-shot examples to improve generation performance by retrieving segments from the training dataset that are similar to the input segment, along with their ground-truth observations.

To construct the retrieval pool, we first processed the training dataset using ChatGPT. We extracted individual segments and their corresponding observations to create a database of pairs. Figure 4 illustrates the prompt used for this extraction process,

Prompt for Segment Generation for Train dataset

You are a precise medical documentation assistant.
Your task is to segment the transcript strictly by **sentences** and map observations to them.

CLINICAL CONCEPT DEFINITIONS
Use the following definitions to understand the context of the observations. This will help you match the transcript text to the correct observation even if the wording is slightly different.
{ONTOLOGY LIST}

SEGMENTATION RULES:
1. **Unit:** The minimum segment unit is a full grammatical **sentence**. Do not split a single sentence into smaller phrases or clauses, even if it contains multiple facts.
2. **Completeness:** The entire transcript must be reconstructed in the "segments" list sequentially, including [Clinician].

MAPPING RULES:
1. **Multiple Matches:** Since a single sentence can contain multiple clinical facts, the "observations" list for a segment can contain multiple items.
2. **Relevance:** Include an observation in a segment ONLY if that specific sentence explicitly mentions or supports that observation (refer to the Definitions above).
3. **Empty List:** If a sentence is just conversational filler (e.g., "Hello," "Okay") and contains no specific medical orders/observations from the list, return an empty list `[]` for "observations".

CRITICAL CONSTRAINTS (MUST FOLLOW):
1. **NO OMISSIONS:** Every single object in the provided "Observations" input list **MUST** be mapped to at least one segment.
2. **Check Before Output:** Before generating the final JSON, verify that the count of unique observation IDs in your output matches the count of observation IDs in the input. If an observation is missing, find the most relevant sentence and map it there.

OUTPUT FORMAT:
Return a JSON object with a key "segments".
"segments": [
{{
"segment": "Full sentence text here.",
"observations": [{{ ...obs1... }}, {{ ...obs2... }}]
}},
...
]

Transcript:
{TRANSCRIPT}

Observations:
{OBSERVATIONS}

Figure 4: Prompt for segment generation for train dataset

Segment For Train Dataset
Segment: [Clinician] Patient presents with moderate dyspnea.
Observations: [{"id": "67", "value_type": "SINGLE_SELECT", "name": "Dyspnea", "value": "Moderate"}]
Segment: We're keeping an eye on all these symptoms to manage the patient's care effectively.
Observations: []

Figure 5: Segment Example for Train Dataset.

while Figure 5 displays examples of the resulting segment-observation pairs. From this processed dataset, we retrieve the top 15 examples to serve as few-shot context for the LLM. This retrieval step is performed using a hybrid method combining BlueBert and BM25.

3.5. Observation Generation

We retrieved N and K entries, respectively, from the ontology pool and the few-shot example pool generated in the previous section, instructing the LLM to extract medical observations relevant to the current utterance. Figure 6 presents the prompt used for this task.

4. Experiments

In this section, we present the evaluation results of our proposed approach and analyze the performance of various model configurations.

Prompt for Observation Generation

```
You are an expert clinical documentation assistant.
Your task is to extract clinical observations from the transcript segment and map them strictly to the provided SCHEMA.

### INSTRUCTIONS:
1. Review the TRANSCRIPT SEGMENT.
2. Extract ONLY information that matches the 'name' and definitions in the SCHEMA.
3. If no information matches the schema, return an empty list [].
4. For 'SINGLE_SELECT' or 'MULTI_SELECT', values MUST be from 'value_enum'.
5. Return ONLY a valid JSON list.

### RELEVANT SCHEMA:
{schema_str}

### REFERENCE EXAMPLES (JSONL format):
Below are similar examples from the training data. Follow the mapping logic shown in "observations".
{few_shot_text}

### CURRENT TRANSCRIPT SEGMENT:\n{transcript_segment}

Return ONLY the JSON list of observations.
```

Figure 6: Prompt for observation generation

Table 1: Performance comparison of the proposed methods. Best results are marked in **bold**.

Method	Precision	Recall	F1
+ Few-shot Ex	0.789	0.694	0.739
+ Few-shot Ex + Schema	**0.812**	**0.847**	**0.829**

4.1. Evaluation on Development Dataset

4.1.1. Overall Result

Table 1 summarizes the performance of different methods on the development dataset. In this experiment, we utilized GPT-5-mini as the backbone model. As shown in the table, incorporating schemas alongside few-shot examples yields superior performance in extracting automatic medical observations from nurse dictations, compared to providing few-shot examples alone.

4.1.2. Ablation on Retrieval Model

We conducted an ablation study on the retrieval models used to fetch schemas and few-shot examples. Table 3 summarizes the performance results. As shown in the table, BlueBERT achieves the best performance in terms of F1-score.

4.2. Evaluation on Test Dataset

In this section, we present the evaluation results on the MEDIQA-SYNUR test dataset, summarized in Table 2. Regarding model configuration, the combination of GPT-5.1 as the observation generator and GPT-5-mini as the segmentor yielded the best performance. We also observed a positive correlation between the number of few-shot examples and overall performance. Interestingly, while larger generator models generally improved performance (e.g., GPT-5.1 outperformed GPT-5-mini), this trend was not strictly monotonic. Specifically, GPT-5.2 exhibited performance degradation compared to GPT-5.1. Furthermore, the superior performance of the GPT-5-mini segmentor can be attributed to model consistency. Since the memory bank of segments was constructed using GPT-5-mini, employing the identical model for segmentation during inference ensures better alignment with the retrieved examples, leading to optimal results.

5. Discussion

Efficiency in Observation Extraction Pipeline In this study, we utilized in-context learning by retrieving medical terms from few-shot examples and an ontology. However, during the inference phase, when dictation is divided into segments, not all segments necessarily contain meaningful medical observations. Although this aspect was outside the scope of our current work, we propose that the pipeline's efficiency could be significantly improved by incorporating a preliminary classifier. This classifier would determine the presence of medical observations within a segment before prompting the LLM, thereby reducing unnecessary computational costs for irrelevant segments.

Table 2: Performance comparison with different configurations on the Test Dataset. †denotes the usage of refined prompt.

Base Model	Shots	LLM Backbone		Metrics		
		Generator	Segmenter	Precision	Recall	F1
BlueBert	3	GPT-5-mini	GPT-5-mini	0.739	0.810	0.773
	5	GPT-5-mini	GPT-5-mini	0.744	0.817	0.779
	10	GPT-5-mini	GPT-5-mini	0.744	0.818	0.779
	10	GPT-5.1	GPT-5-mini	0.785	0.805	0.795
	10	GPT-5.2	GPT-5-mini	0.766	0.825	0.794
	10	GPT-5.2	GPT-5.2	0.759	**0.827**	0.792
	15	GPT-5.1	GPT-5-mini	**0.786**	0.807	**0.796**

Table 3: Ablation on retrieval model.

Method	Precision	Recall	F1
ClinicalBert	0.811	0.845	0.828
BlueBert	0.807	0.860	0.833
PubMedBert	0.805	0.842	0.823

Dataset Analysis Our analysis indicates opportunities to improve data consistency. For example, values such as 37.5°C sometimes appear as "375", and unit annotations (*e.g.,* Temperature Unit) are not always consistent. Standardizing value formats and units would likely improve training efficiency and retrieval accuracy.

6. Conclusion

In this paper, we proposed a Retrieval-Augmented Generation (RAG)-based pipeline designed to automatically extract medical observations from nurse dictations. To achieve this, we constructed a memory bank leveraging both a medical observation ontology and previously annotated observation tags from existing dictations. By integrating these dataset into the LLM generation process, our approach guides the model to produce more accurate outputs. Consequently, our method demonstrated its effectiveness by achieving an F1-score of 0.796 on the MEDIQA-SYNUR test dataset.

7. Acknowledgements

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343, RS-2022-II220320, RS-2025-25442338).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

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*, pages 859–870, Suzhou (China). Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jiho Jang, Chaerin Kong, DongHyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):980–988.

Hyeonjin Kim, Min Kim, Jae Jang, KiYoon Yoo, and Nojun Kwak. 2024a. TEAM MIPAL at MEDIQA-

M3G 2024: Large VQA models for dermatological diagnosis. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 334–338, Mexico City, Mexico. Association for Computational Linguistics.

Taehoon Kim, Pyunghwan Ahn, Sangyun Kim, Sihaeng Lee, Mark Marsden, Alessandra Sala, Seung Hwan Kim, Bohyung Han, Kyoung Mu Lee, Honglak Lee, et al. 2024b. Nice: Cvpr 2023 challenge on zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365.

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.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.