# Overview of the MEDIQA-SYNUR 2026 Shared Task on Observation Extraction from Nurse Dictations

**George Michalopoulos, Jean-Philippe Corbeil, Cari Bader**
**Nathan Bodenstab and Asma Ben Abacha**
Microsoft Healthcare & Life Sciences
{georgemi, jcorbeil, caribader, nbodenstab, abenabacha}@microsoft.com

**Abstract**

Hospital nurses spend a significant portion of their shifts performing manual data entry tasks. An automatic solution for extracting medical information from nurse dictations into large spreadsheet ontology (flowsheet) could reduce the documentation burden of nurses and alleviate nurse burnout. We introduce the MEDIQA-SYNUR shared task, the first challenge on extracting and normalizing clinical observations from nurse dictations and mapping them to a large ontology of clinical concepts. 13 teams participated in the challenge and experimented with a broad range of approaches. In this paper, we describe the MEDIQA-SYNUR task, the datasets, and the participant's results and solutions.

**Keywords:** nursing shared task, observation extraction, ontology mapping

## 1. Introduction

Nurses currently manually enter information into the electronic health record (EHR) using flowsheets, which are similar to spreadsheets. Each flowsheet can have multiple tabs and hundreds or even thousands of rows, with each row representing a specific data point such as temperature, heart rate, etc. Working with flowsheets is a time-consuming process, since the nurse must first navigate to the appropriate tab in the flowsheet, identify the row or groups of rows corresponding to the assessment that is being documented, then manually enter the information cell by cell. Due to the complexity of the flowsheet, nurses often cannot document during rounds and instead delay this task until breaks, leading to delayed documentation and potential accuracy issues.

Surveys show growing interest in NLP for both nursing notes (Mitha et al., 2023) and wider nursing tasks (Panchal and Thakur, 2024). Yet since nurses still chart mainly in electronic health record (EHR) flowsheets, any automation must mesh with that structure.

LLMs can already translate clinical narratives into structured variables with little or no task-specific training (Ling et al.; Dagdelen et al., 2024). Yet their edge over fine-tuned encoder models remains debated (Gutiérrez et al., 2022; Ling et al., 2023). Consequently, no existing system simultaneously (i) ingests nurse dictations, (ii) conditions on the local flowsheet context, and (iii) outputs EHR-ready observations.

The MEDIQA-SYNUR shared task tackles the problem of extracting and normalizing clinical observations from nurse dictations and mapping them to a large ontology of clinical concepts (flowsheet), as described in Figure 1.

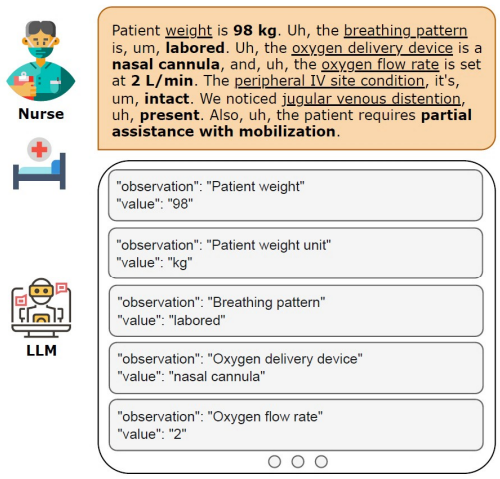


Figure 1: The observation extraction from nurse dictations task takes a nurse dictations and a local flowsheet and extracts EHR-ready observations

In this paper, we present the task and the datasets in Sections 2 and 3. In Section 4, we present the evaluation metric used in the shared task. Section 5, describes the approaches of the participating teams, and in Section 6 we discuss the final insights from the official challenge results.

## 2. Task Description

The MEDIQA-SYNUR 2026 shared task[1] addresses the problem of extracting clinical observations from nurse dictations.

[1] https://sites.google.com/view/mediqa2026/mediqa-synur



*Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026*, pages 19–26
16 May 2026. 

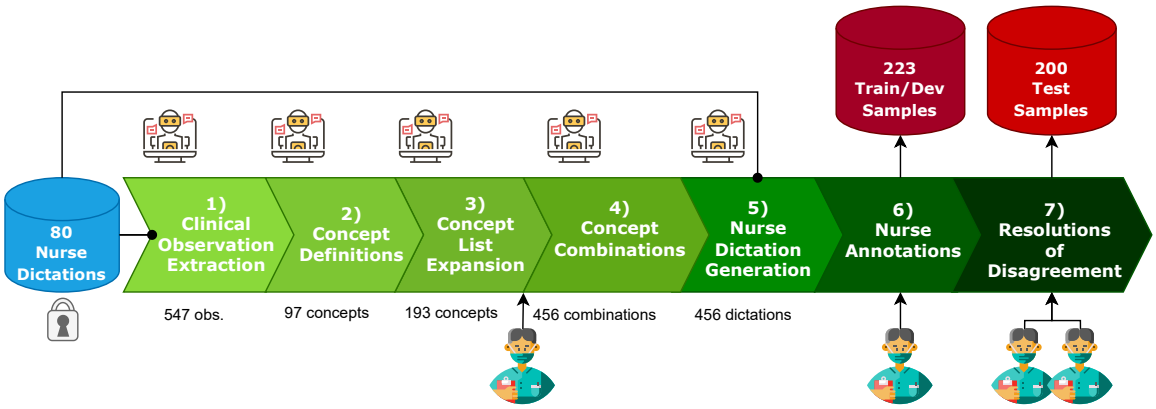


Figure 2: SYNUR Dataset Creation Pipeline for generating realistic, synthetic nurse dictations with expert nurse annotations, comprising six steps: starting from 80 human-verified, fabricated dictations to final expert annotations. The number below each step represents the final amount of output elements generated at this specific step. The output of one step is the input of the next step, except for step 5 for which we also include 5 randomly sampled examples from the seed dictations. At steps 6 & 7, the annotators also have access to the schema of step 3 along the synthetic dictation to produce its gold annotations.

Specifically, in the task data, we provide a hospital flowsheet schema that contains a list of all the rows (keys) in the flowsheet, the data type associated with each row (string, picklist, etc.), and, for rows associated with a picklist, a list of all valid values. It should be noted that the flowsheet schema can include hundreds to thousands of rows but it does not include any data specific to patients.

In addition, we provide a set of clinical transcripts where each transcript contains a realistic synthetic nurse dictation that mentions or describes a set of medical observations.

This task focuses on automatically populating the flowsheet by extracting all relevant medical information that are defined in the flowsheet and exist in each dictation.

## 3. Data Creation

In previous work (Corbeil et al., 2025b), we introduced SYNUR, a synthetic nursing dataset, which uses 80 proprietary fabricated seed dictations that were reviewed by practicing nurses for realism. Figure 2 summarizes the seven-stage pipeline, which alternates between the domain experts and the *GPT-4o-0806* language model.

1. **Observation mining**: The LLM iteratively extracted observation phrases (e.g., “dark yellow”) and linked them to clinical concepts (e.g., “urine colour”) from the 80 notes, yielding 547 unique observations.

2. **Concept consolidation**: We distilled these observations into 97 clinical concepts and assigned each a data type among boolean, integer, string, single-choice, or multi-choice.

3. **Ontology expansion**: Leveraging the LLM’s medical knowledge, we asked it to add relevant missing concepts with example sentences, expanding the ontology from 97 extracted concepts to 193 concepts, which an experienced nurse then validated and corrected.

4. **Scenario generation**: We prompted the LLM with the final concept set to craft coherent patient-case rationale and compatible observation combinations.

5. **Dictation synthesis**: Each scenario along with 5 randomly sampled seed dictations are passed to the LLM to generate a realistic nurse dictation that includes natural speech disfluencies and on-the-fly corrections.

6. **Train & development set annotation**: Expert nurse annotators verified the synthetic transcripts, and produced reference annotations using the ontology build at step 3 for a total of 223 dictations for training and development purposes that contain 3000 observations.

7. **Test set annotation**: We then produce a test set of 200 transcripts with an enhanced quality achieved by a disagreement-resolution step with the expert annotators.

For the MEDIQA-SYNUR 2026 shared task, we used the above mentioned 223 dictations as the training and development data—i.e., a 55:45 split, respectively. We also used the above mentioned 200 dictations as the test set that all teams were evaluated on.

| Team | Affiliation | Paper | Code |
|---|---|---|---|
| Semantica Lab | University of Pennsylvania & Cedars Sinai Medical Center & University of Pittsburgh - USA | (Hwang et al., 2026) | 1 |
| MasonNLP | George Mason University - USA | (Karim and Özlem Uzune, 2026) | 2 |
| Smart_solutions | M42 - UAE | (Munjal, 2026) | 3 |
| HSE NLP | Higher School of Economics - Russia | (Valiev, 2026) | 4 |
| MKC | Seoul National University - South Korea | (Hwang and Kwak, 2026) | 5 |
| Gladiators | AbbVie - USA | (Satyanarayana et al., 2026) | 6 |
| MIDAS_SYNUR | MIDAS Lab, IIIT - India | (Sriram and Sahoo, 2026) | 7 |
| NSN | University of Pittsburgh & Washington University - USA | (Aryoyudanta et al., 2026) | 8 |
| Lakefront AI Ramblers | Loyola University Chicago - USA | (Saban et al., 2026) | 9 |
| AnotherOne | International Institute of Information Technology - India | (Thomas and Krishnamurthy, 2026) | 10 |
| LTRC-MEDICOM | International Institute of Information Technology - India | (Deepak et al., 2026) | 11 |
| SQUCS | Sultan Qaboos University - Oman | (JeebAllah et al., 2026) | 12 |
| LTRC-IIIT | International Institute of Information - India | (Vaish and Sharma, 2026) | 13 |

1 https://github.com/syhwng/SemanticaLab-MediQA-SYNUR_submissions
2 https://github.com/AHMRezaul/MEDIQA-SYNUR-2026
3 https://github.com/PrateekMunjal/MEDIQA26_SYNUR_submission_Smart_solutions
4 https://github.com/Rebell-Leader/MEDIQA-SYNUR-HSE_NLP
5 https://github.com/KyominHwang/ClinicalNLP-2026-MKC
6 https://github.com/singhak-abbvie/MEDIQA-SYNUR-Gladiators
7 https://github.com/Swekrish303/MEDIQA-SYNUR-Team-MIDAS_SYNUR
8 https://github.com/yudanta/nsn-mediqa-synur
9 https://github.com/arsalanyaghoubi/Lakefront_AI_Ramblers-MEDIQA_SYNUR_2026
10 https://github.com/jr-john/mediqa-synur-2026
11 https://github.com/sushvinmarimuthu/ltrc-medicom-mediqa-synur-clinicalnlp-2026
12 https://github.com/RihamJeeballah/mediqa-synur-2026
13 https://github.com/av-dx/mediqa-synur-ltrc-iiit

Table 1: MEDIQA-SYNUR 2026: Participating teams, papers and codes.

# 4. Evaluation

## 4.1. Evaluation Metrics

For the task of extracting and normalizing clinical observations from nurse dictations and mapping them to a large flowsheet ontology of clinical concepts, we rely on the standard $F1$ metric for the evaluation of the extracted observations. In addition, we also present both precision ($P$) and recall ($R$) as secondary metrics for analysis purposes.

It should be noted that for each extracted multi-choice observation, we only mark it as true positive if it contains all the values that exist in the corresponding reference observation's list value. We provide a standardized evaluation script available on GitHub[2].

[2] https://github.com/abachaa/MEDIQA-SYNUR-2026

## 4.2. Code Verification

For additional validation, we required the submission of the code in addition to the models' outputs/runs, in order to verify the reproducibility of the reported runs. The participants shared their private codes with the organizers on GitHub following provided guidelines.

## 4.3. Baseline System

We built a GPT-4.1 based baseline system, with deterministic outputs (temperature=0). Our approach contains three sub-tasks (Corbeil et al., 2025b):

1. **Segmentation**: The segmentation step splits the streaming nurse transcript into medically coherent, continuous and non-overlapping segments via an LLM.

2. **RAG**: We minimize the size of the schema by filtering potential rows to the top $N$ candidates given the current segment based on cosine similarity.

3. **Extraction**: Finally, we use the LLM to extract the relevant text from the transcript segment and canonicalize it based on the (now reduced) schema.

## 5. Official Results

### 5.1. Participating Teams

Thirteen finalist teams, out of 66 registered teams, submitted their code and runs for the MEDIQA-SYNUR 2026 shared task[3]. Table 1 presents the papers and the code for each team that participated. We limited the number of submitted runs to 15 runs per team.

### 5.2. Leaderboard Ranking

Table 2 presents the official results of the MEDIQA-SYNUR competition. Overall, the results show a relatively narrow performance gap among the top-ranked systems, with most approaches converging toward similar F1 scores despite substantial differences in model size and pipeline design.

Although excluded from the official ranking, the GPT-4.1 baseline (Corbeil et al., 2025b) achieved the highest overall performance with an F1 score of 0.837, outperforming the top ranked system by 2.3 absolute points. The baseline also achieved the highest precision and the second-highest recall, indicating that the proposed segmentation and RAG-based pipeline remains a very strong approach for the task.

Among the official submissions, the first-ranked system from *Semantica Lab* (Hwang et al., 2026) achieved an F1 score of 0.814 using a committee of frontier LLMs combined with an LLM-as-a-Judge verification strategy. This confirms the effectiveness of ensemble and verification-based pipelines, which appear frequently among the top-performing systems.

A notable observation is that systems ranked between second and fifth positions achieved very similar F1 scores (0.796–0.804), despite relying on different models and architectures. Many of these approaches were based on GPT-5 models or very large open-weight models such as Kimi-K2 MoE and Qwen3 235B, suggesting that performance differences in this range are driven more by pipeline design (e.g., retrieval, verification, or post-processing) than by the base model alone.

Different optimization strategies are also visible in the precision–recall trade-offs. For example, the *AnotherOne* system achieved the highest recall (0.835) but relatively low precision, while other top systems prioritized precision through verification or consensus mechanisms. This indicates that systems relying on verification stages tend to favor precision, whereas single-pass extraction systems often favor recall.

Another important result is the performance of smaller models. The *NSN* team successfully fine-tuned a Qwen3 2B model using SFT and GRPO, achieving an F1 score of 0.740 without complex pipeline components. This result is notable because it demonstrates that fine-tuned small language models can remain competitive with much larger models when properly trained for the task.

Finally, the lower-ranked systems (positions 11–13) relied primarily on prompting Llama-based models (8B and 70B), which resulted in lower overall performance. This suggests that prompting-only approaches without retrieval, verification, or fine-tuning are insufficient for this task compared to multi-stage pipelines or fine-tuned models.

Overall, the leaderboard highlights three main successful strategies: (i) multi-stage pipelines with retrieval and verification, (ii) ensembles or multi-model systems, and (iii) fine-tuning smaller models specifically for the task. These findings provide useful insights for future work on clinical observation extraction from nurse dictations.

### 5.3. Approaches

#### 5.3.1. Semantica Lab

The *Semantica Lab* team (Hwang et al., 2026) introduced a 6-step mutli-stage pipeline that (i) normalize transcripts, (ii) generate a transcript-specific set of candidate concepts, (iii) filter candidates using an evidence gate or use an online router to select which domain experts to invoke for the transcript (iv) retain only extractions supported by cross-model agreement (v) send remaining ambiguous cases to an adjudicator. (vi) and canonicalize outputs and apply a consensus-based intersection of complementary runs.

#### 5.3.2. MasonNLP

The *MasonNLP* team (Karim and Özlem Uzune, 2026) presented a modular retrieval-augmented extraction pipeline that contains the following components: (i) training-set exemplar retrieval, (ii) schema conditioned prompting (full schema or pruned candidate schema), (iii) deterministic schema-based postprocessing (iv) and a second-pass audit to refine outputs.

#### 5.3.3. Smart_solutions

The *Smart_solutions* team (Munjal, 2026) adopted a multi-stage pipeline of agents. Specifically, they

[3] https://www.codabench.org/competitions/12113/

| Rank | Team | Model(s) | Pipeline Components | | | | | Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Desc. | Seg. | RAG | #LR | Ver. | P | R | F1 |
| - | `Baseline` | GPT-4.1 | | ✓ | D | 2 | R | **0.843** | 0.831 | **0.837** |
| 1 | Semantica Lab | GPTs & Claudes | | | | 6 | J | 0.826 | 0.801 | 0.814 |
| 2 | MasonNLP | GPT-5.2 | | | D | 2 | V | 0.786 | *0.822* | *0.804* |
| 3 | Smart_solutions | Kimi-K2 | ✓ | | S | 5 | | *0.805* | 0.799 | 0.802 |
| 4 | HSE NLP | Qwen-235B & GPT-5 | | | | 3 | J/X | 0.781 | 0.819 | 0.800 |
| 5 | MKC | GPT-5-mini & GPT-5.1 | ✓ | ✓ | H | 1 | | 0.786 | 0.807 | 0.796 |
| 6 | Gladiators | Claude Opus 4.5 | | | | 1 | R | 0.731 | *0.822* | 0.774 |
| 7 | MIDAS_SYNUR | GPT-5.2 | | | | 1 | | 0.779 | 0.767 | 0.773 |
| 8 | NSN | Qwen3 2B $_{\text{SFT/GRPO}}$ | | | | 1 | | 0.701 | 0.785 | 0.740 |
| 9 | Lakefront AI Ramblers | GPT-4o-mini | | ✓ | H | 2 | V | 0.768 | 0.662 | 0.711 |
| 10 | AnotherOne | GPT-OSS 120B | | | | 4 | R | 0.607 | **0.835** | 0.703 |
| 11 | LTRC-MEDICOM | Gemma2 & Llama3 8B | | | | 2 | R | 0.619 | 0.721 | 0.666 |
| 12 | SQUCS | Llama3.3 70B | | | | 3 | R | 0.529 | 0.673 | 0.592 |
| 13 | LTRC-IIIT | Llama3.1 8B | | | | 2 | | 0.598 | 0.551 | 0.574 |

Table 2: Leaderboard results of MEDIQA-SYNUR (metrics: **P**recision, **R**ecall, **F1**). **Bold**, underlined and *italic* values are first, second and third scores per column, respectively. Several acronyms were used to differentiate steps across pipelines (excluding the mandatory extraction step): *Desc.* stands for description of concepts in the ontology; *Seg.* is for the transcript segmentation stage; *RAG* has the sparse (S), dense (D) and hybrid (H) systems; *#LR* stands for numbers of language-model roles; *Ver.* is for verification such as llm verification (V), llm-as-a-judge (J), rules (R), reparation (X), etc.

decomposed the main nursing task into 5 different subtasks: (i) observation extraction, (ii) ontology candidate retrieval, (iii) relevance scoring (iv) deterministic value assignment (v) post-processing and evaluation.

#### 5.3.4. HSE NLP

The *HSE NLP* team (Valiev, 2026) introduced a three-stage pipeline that includes: (i) complementary generators that maximize candidate coverage (ii) a dedicated adjudicator that operates as a verifier rather than a generator, (iii) a token-efficient repair step that restores strict schema compliance.

#### 5.3.5. MKC

The *MKC* team (Hwang and Kwak, 2026) proposed a RAG-based pipeline that first segment the input dictation into distinct units containing clinical facts. Afterwards, they constructed a memory bank for leveraging both a medical observation ontology and previously annotated observation tags from existing dictations. Finally, they integrated these dataset into a LLM generation process, in order to guide the model to produce more accurate outputs.

#### 5.3.6. Gladiators

The *Gladiators* team (Satyanarayana et al., 2026) presented a hybrid approach that integrates large language models with domain-specific validation rules. Their system used a variety of comprehensive prompt engineering strategies and specialized filters and correction mechanisms. In addition, they included a supplementary rule-based component which includes a large set of regex patterns in order to captures high-confidence observations.

#### 5.3.7. MIDAS_SYNUR

The *MIDAS_SYNUR* team (Sriram and Sahoo, 2026) showcased a prompting-based system that adopts a single-prompt, field-rich few-shot strategy, to jointly generate all schema fields in one structured output. Specifically, the few-shot demonstrations are curated and grouped by value type, in order to promote consistency across heterogeneous value distributions.

#### 5.3.8. NSN

The *NSN* team (Aryoyudanta et al., 2026) presented a fine-tuned decoder only LLM for clinical observation extraction trained on the MEDIQA-SYNUR shared task dataset using an SFT followed

by GRPO-RL pipeline. In addition to the shared-task dataset, the team also incorporated an augmented CoT dataset during SFT, which improved model performance after GRPO-RL training.

#### 5.3.9. Lakefront AI Ramblers

The *Lakefront AI Ramblers* team (Saban et al., 2026) introduced a hybrid schema retrieval with model/post-processing-level enhancements that includes the following steps: (i) segmentation of each transcript (ii) processing each segment by using RAG to select a small subset of schema concepts (iii) LLM-generated schema-constrained JSON extractions (iv) validation with optional post-processing steps (voting and LLM-based verification).

#### 5.3.10. AnotherOne

The *AnotherOne* team (Thomas and Krishnamurthy, 2026) adopted a modular four-stage LLM pipeline that includes: (i) knowledge-enhanced concept detection using medical domain clustering (ii) evidence-grounded value extraction (iii) schema-constrained value normalization, and (iv) deterministic post-processing with fuzzy matching and unit pairing.

#### 5.3.11. LTRC-Medicom

The *LTRC-Medicom* team (Deepak et al., 2026) presented a pipeline that consists of four distinct phases: (i) Semantic Clustering (ii) Systematic prompt engineering that includes a 9-step chain-of-thought system prompt (iii) Supervised Fine-Tuning using QLoRA (Quantized Low-Rank Adaptation) and (iv) transcript verification.

#### 5.3.12. SQUCS

The *SQUCS* team (JeebAllah et al., 2026) presented a multi-agent LLM-based system that decomposes the extraction process into specialized agents responsible for (i) schema-guided extraction which includes an additional transcript segmentation mechanism (ii) rule-based validation (iii) and precision-oriented filtering verification. Finally, they also include a rule-based post-processing step informed by systematic error analysis on the development set.

#### 5.3.13. LTRC-IIIT

The *LTRC-IIIT* team (Vaish and Sharma, 2026) introduced a pipeline that consists of two distinct stages: (i) a retrieval module designed to filter relevant flowsheet rows from the full set of schema rows (ii) a generation module that extracts values using a local instruction LLM.

## 6. Conclusion

The MEDIQA-SYNUR 2026 shared task was tackled by a wide variety of approaches from the participating teams. The runs submitted by the participating teams explored different retrieval-augmented generation, fine-tuning and prompting methods which we believe will provide new insights for future research directions in the task of observation extraction from nurse dictations.

The best performance was achieved from closed-weight model LLMs (e.g. GPT and Claude frontier models) but also teams that used very large open-weight model (e.g. Kimi-K2 MoE and Qwen3 235B) achieved similar results—in line with previous works (Dada et al., 2025; Corbeil et al., 2025a). A noticeable submission finetuned Qwen3 2B with SFT and GRPO achieving competitive results for such a small language model that was fine-tuned on a very small training set, paving the way to apply SLMs for the nurse observation extraction task.

Considering the diverse pipeline designs, we noticed a few successful strategies: adding descriptions to ontological concepts, segmenting the dictations, using RAG with dense embeddings, including several LLMs (i.e., agents or ensemble), and post-processing with LLM-as-a-Judge.

Incorporating all of these insights seems promising for the future of systems tackling the nurse observation extraction task. We hope that this shared task will encourage further efforts towards automatic observation extraction from nurse dictations to reduce the workload for medical professionals.

## Limitations

The paper does not cover all possible methods and models for extracting clinical observations from nurse dictations. The SYNUR dataset is also limited in terms of size and types of clinical observations. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.


## Acknowledgments

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their support for the shared task. We also thank our annotation team, Molly Benton, Dominique Beck, Laurie Bonci, Ramona Brown, Nakisha Sanders, and Stefanie Barrera, for annotating the new SYNUR test set, as well as the participating teams who contributed to the success of the shared task through their interesting approaches and strong engagement.

## 7. Bibliographical References


Bayu Aryoyudanta, Maria Yuliana, Mikie Rachman, and I Made Agus Setiawan. 2026. Night shift nerds at mediqa-synur 2026: Pushing small large language model capability for clinical observation extraction and normalization from nurse dictation using rlvr. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025a. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025b. 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.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Jens Kleesiek, et al. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM$^2$)*, pages 46–59.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Pasumarthy Deepak, Sushvin Marimuthu, and Parameswari Krishnamurthy. 2026. Ltrc-medicom at mediqa-synur 2026: Schema-guided clinical information extraction with hybrid clustering-sft-verification. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Kyomin Hwang and Nojun Kwak. 2026. Team mkc at mediqa-synur 2026: Retrieval-augmented generation based nurse observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Sy Hwang, Katherine S. Pitcher, Sue Hyon Kim, Yoonjae Lee, Hayoung K. Donelly, Harsh Bandhey, Andrew J. King, Karen O'Connor, Ryan J. Urbanowicz, and Danielle L. Mowery. 2026. Semantica lab at mediqa-synur 2026: Route, extract and verify – an llm-gated ensemble for parsing nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Riham JeebAllah, Adhari AlZaabi, and Abdulrahman Khalifa AAlAbdulsalam. 2026. Squcs at mediqa-synur 2026: A multi-agent open source llm system for nursing observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

A H M Rezaul Karim and Özlem Uzune. 2026. Masonnlp at mediqa-synur 2026: Retrieval-augmented large language models for schema-constrained clinical information extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Mika Oishi, Takao Osaki, Katsushi Matsuda, et al. Improving open information extraction with large language models: A study on demonstration uncertainty. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

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.

Prateek Munjal. 2026. Smart_solutions at mediqa-synur 2026: A multi-stage llm pipeline for nursing

observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

S. Panchal and P. Thakur. 2024. Harnessing the power of natural language processing in nursing services. *International Journal of Advanced Research*, 12(5):154–156.

Michael T. Saban, Arsalan Yaghoubi, Behnaz Eslami, Samie Tootooni, and Dmitriy Dligach. 2026. Lakefront ai ramblers at mediqa-synur 2026: Hybrid retrieval and llm verification for open-source schema-guided clinical information extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Siva Satyanarayana, Raju Pusapati, and Ankit Singh. 2026. Gladiator at mediqa-synur 2026: Contextual clinical extraction: Integrating foundation models with domain-specific validation rules. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Swetha Krishna Sriram and Akshitaa Sahoo. 2026. Midas_synur at mediqa-synur 2026: A prompting study for clinical observation extraction from nurse dictation transcriptions. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Jerrin John Thomas and Parameswari Krishnamurthy. 2026. Anotherone at mediqa-synur 2026: Detect, extract, normalize - knowledge-grounded llm pipeline for clinical observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Aashwin Vaish and Dipti Misra Sharma. 2026. Ltrciiit at mediqa-synur 2026: Benchmarking a fully local, training-free rag pipeline. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Airat A. Valiev. 2026. Hse nlp team at mediqa-synur 2026: Consensus adjudication ensemble (ace): Balancing precision and recall for schema-bystander clinical extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.