Smart_solutions at MEDIQA-SYNUR 2026: A Multi-Stage LLM Pipeline for Nursing Observation Extraction

Prateek Munjal
M42
Abu Dhabi, UAE
pmunjal@m42.ae

Abstract

Extracting clinical observations from nursing dictations addresses an important problem of addressing burden in clinical documentation. In this work, we describe our approach submitted to MEDIQA-SYNUR 2026, which achieved third place among participating teams with a balanced precision and recall of 0.80 on the unseen test set. Our approach, instead of finetuning LLMs, is to adopt a multi-stage pipeline of agents: Observation Agent, Ontology Matching Agent, Relevance Scoring Agent, Evidence Assignment Agent, and Formatting Agent. First, the Observation Agent extracts clinical observations and corresponding evidence from the nurse transcript. These observations are then processed by the Ontology Matching Agent, which maps them to a restricted set of candidate ontology fields via TF-IDF–based retrieval, and subsequently evaluated by the Relevance Scoring Agent, which assigns continuous support scores (1–5) to each candidate field. Finally, field value assignments are performed by the Evidence-Based Agent, which extracts values strictly from nurse transcripts and clinical observations (Observation Agent outputs) to populate each ontology field. These outputs are then formatted by the Formatting Agent to ensure correct submission structure with the necessary metadata. Our agentic system results suggest that combination of agents with prompt engineering can narrow the gap between general and specialized clinical NLP models, making it an immediately deployable alternative to traditional fine-tuning.

Keywords: clinical information extraction, EHR structuring, LLMs, agents, clinical NLP, nursing transcripts

1. Introduction

Clinical nursing documentation (Willard-Grace et al., 2019; Friganović et al., 2019) is a crucial problem which helps in capturing continuous patient treatment and clinical decision-making. The nurses record patient observations frequently during the patient's stay in the hospital. These observations are typically mental status, functional ability, vital signs, and risk assessments, which are converted into structured EHR records. This process in practice is time-consuming, leading to a documentation burden for nurses(Padilla Fortunatti and Palmeiro-Silva, 2017; McHugh et al., 2011). It is important to note that automated extraction of nursing observations presents unique challenges compared to other clinical information extraction tasks. Nursing dictations are often conversational, fragmented, and context-dependent, while the target output is a large, fine-grained ontology of structured fields with strict value constraints. High-stakes domains like healthcare make the errors in this setting costly, highlighting the need for ideal system to prioritize both precision and recall metrics (Hurd et al., 2024; Altermatt et al., 2025).

The MEDIQA-SYNUR shared task (George Michalopoulos, 2026) addresses this problem by benchmarking systems for extracting observation from nurse dictations/transcripts, requiring the participants to map free-text nurse dictations to 192 structured EHR fields. These fields span binary, categorical, and free-text values. Unlike prior medical order or entity extraction tasks, MEDIQA-SYNUR task emphasizes on grounding and balanced precision–recall performance by measuring F1. Recent works (Ntinopoulos et al., 2025) in clinical NLP has increasingly explored LLMs for information extraction, often relying on single-stage, end-to-end prompting or domain specific fine-tuning of generalist models. While the finetuning approaches (Liu et al., 2025a; Majdik et al., 2024; Singhal et al., 2025; Christophe et al., 2024; Tu et al., 2024) demonstrate report strong performances gains, in this study we deliberately adopt to use the generalist LLM (Yang et al., 2025; Liu et al., 2024; Team et al., 2025; Grattafiori et al., 2024; Achiam et al., 2023) without any task specific training. We make this choice to benchmark immediately deployable baseline for resource-constrained clinical settings where fine-tuning is either infeasible or too costly.

In this paper, we present our approach of agentic system submitted to MEDIQA-SYNUR 2026, which rather than treating observation extraction as a single inference task, we decompose it into a multi-stage pipeline consisting of 5 agents: namely;(i) Observation Agent, (ii) Ontology Matching Agent, (iii) Relevance Scoring Agent, (iv) Evidence Assignment Agent, and (v) Formatting Agent. Our system achieves third place among participating teams on the MEDIQA-SYNUR test set, with balanced precision and recall of 0.80. The high F1 score

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 83–91

of 0.80 demonstrates that Agentic AI (Shimgekar et al., 2025) can support in automating extracting observations from nurse dictations, and help with documentation burden experienced by nurses.

2. Related Work

Clinical information extraction has traditionally been approached through supervised learning using domain-specific encoders and sequence labeling models. Early work (Zhang and Chen, 2022; Hu et al., 2024) focused on named entity recognition and relation extraction over structured clinical notes, with models such as BioBERT (Lee et al., 2020) and its variants demonstrating strong performance on biomedical benchmarks (Luo et al., 2022). While effective for well-defined entity types, these approaches require substantial annotated data and struggle to generalize to long, conversational inputs such as nurse dictations or patient doctor conversations (Corbeil et al., 2025a).

Subsequent work explored encoder–decoder architectures for joint extraction and generation, enabling models to produce structured outputs directly from clinical text (Kim et al., 2025; Mehta, 2025). Although these models improve flexibility, they often exhibit narrative drift and schema inconsistency (Asgari et al., 2025; Zeba et al., 2025) when applied to large, fine-grained ontologies. In nursing documentation tasks, where hundreds of fields may be sparsely expressed, such models tend to overgenerate defaults or infer unstated normal findings.

More recently, large language models (LLMs) have been applied to clinical extraction tasks through fine-tuning or prompt-based inference (Kim et al., 2025; Karim and Uzuner, 2025; Sellergren et al., 2025). Fine-tuned medical LLMs show promise in capturing domain-specific terminology, but their deployment is constrained by computational cost and limited adaptability to evolving schemas. Prompt-based approaches avoid retraining but commonly rely on single-stage, end-to-end generation, which can lead to hallucinated fields, weak evidence grounding, and poor abstention behavior in safety-critical settings. (Zeba et al., 2025) Several shared tasks in the MEDIQA series (George Michalopoulos, 2026; Corbeil et al., 2025a) have highlighted these challenges, including medical order extraction and clinical question answering. Successful systems increasingly emphasize structured outputs, explicit evidence attribution, and rule-based post-processing (Agrawal et al., 2022; Yan et al., 2026). However, most existing approaches still conflate semantic relevance determination with factual assignment, forcing models to decide what is relevant and what is true in a single step. Our work explicitly decoupling relevance judgment from value assignment. We employ LLMs as controlled reasoning components for high-recall observation extraction and continuous relevance scoring (refer Figure 1), and we note that this design aligns (Liu et al., 2025b; Adam et al., 2024) that treat LLMs as modular decision-makers rather than end-to-end generators, and it is particularly well suited to ontology-grounded nursing documentation tasks with low tolerance for inference or hallucination.

3. System Overview

Our approach is designed as a multi-agent extraction framework that converts free-text nurse dictations into structured EHR observations. Instead of relying on a single end-to-end generation, our approach orchestrates a sequence of specialized agents, each responsible for a well-defined subtask within the extraction process.

Figure 1 illustrates the overall architecture. Given a nurse dictation/transcript as input, the system proceeds through five stages: (1) observation extraction, (2) ontology candidate retrieval, (3) relevance scoring, (4) deterministic value assignment, and (5) post-processing and evaluation. All the agents use Kimi-K2 (Team et al., 2025) as LLM with different system prompts under zero shot settings.

3.1. Observation Agent

In the first stage, the LLM Agent processes the full nurse transcript to extract candidate clinical observations along with supporting evidence spans. This stage is intentionally permissive (maximizing recall) as its free form extractions tries to extrall all potential relevant observation without thinking of what ontology fields can be extracted. In the first stage, the Observation Agent (an LLM Agent) processes the full nurse transcript to extract candidate clinical observations along with their supporting evidence spans. This stage is intentionally permissive (maximizing recall), allowing free-form extraction of all potentially relevant clinical observations without considering any value specific constraints. At this point, the agent focuses solely on identifying explicit clinical facts expressed in the transcript, and not focus on value assignment, normalization, or any reasoning that would require converting or interpreting the extracted observations.

System &/ User Prompts

System prompt:
You are a clinical information extractor.

From the transcript, extract all explicit patient observations, symptoms, clinical findings, interventions, and clinician interpretations.

Rules:

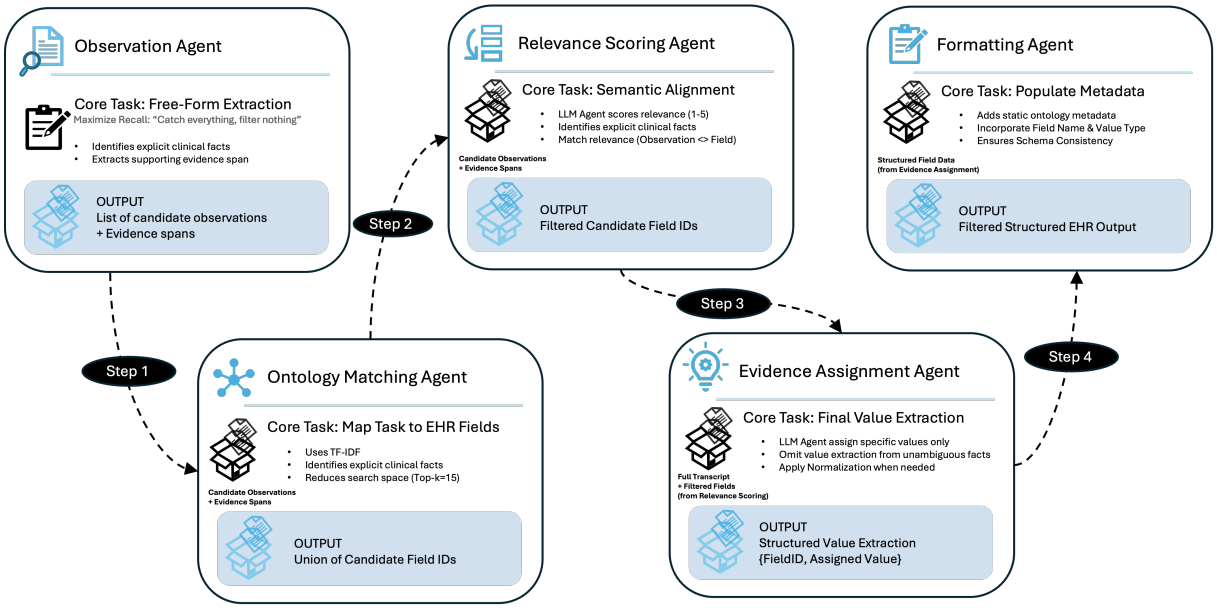

Figure 1: Overview of the LLM-driven pipeline (agent-inspired decomposition) workflow for MEDIQA-SYNUR shared task. The pipeline starts from (1) Observation Agent which extracts clinical concepts, then (2) Ontology Matching Agent reduces search space using TF-IDF retrieval, and (3) Relevance Scoring Agent filters candidate keys (score ≥ 4). Finally, outputs are generated via (4) Evidence Assignment Agent using Medprompt-inspired multi-step reasoning and finally, (5) Formatting Agent, which formats and postprocesses the outputs in the final JSON schema requested in shared task.

- Use short, precise phrases
- Stick strictly to what is stated or directly implied
- Do NOT map to any predefined schema
- Do NOT infer values that are not supported
- Output as a JSON array of atomic observations in format {"concept":...., "evidence":....}

User prompt:

**Transcript: {transcript}

Let's think step-by-step for finding Observations:

3.2. Ontology Matching Agent

The Ontology Matching Agent is responsible for mapping extracted clinical observations to a restricted set of candidate EHR fields. Given the large schema fields (192 fields), directly including all the fields in LLM calls would be inefficient and prone to hallucination/over-generation (Asgari et al., 2025). Following the work of (Zhan et al., 2021), we employ a lightweight TF-IDF–based retrieval mechanism to reduce the search space while preserving recall. For each field, we construct a textual representation consisting of the field name and when available its enumerated values. These representations form a TF-IDF document corpus using unigram and bigram features with ℓ_2 normalization. For each extracted observation, we create a query by concatenating the observation concept and its supporting evidence span. The query is projected into the same TF-IDF space, and cosine similarity is used to find the top-k fields ($k = 15$) with positive similarity scores as candidate fields. The union of retrieved field IDs across observations forms the candidate set of fields. We note that this stage significantly reduced the search space while maintaining high recall.

3.3. Relevance Scoring Agent

The Relevance Scoring Agent (an LLM Agent) evaluates the semantic alignment between extracted observations and candidate ontology fields by assigning a continuous relevance score (1 to 5; where 1 means no support and 5 means explicit transcript support) to each candidate field. We introduce this stage because LLMs, when directly asked to extract structured outputs from transcripts, often struggle to abstain and may over-generate for false/unsupported fields which eventually hurt the precision metric.

System &/ User Prompts

System prompt:
You are a clinical information extraction system.

Your task is to SCORE the relevance of candidate EHR fields.

You will be given:
1) The original clinical transcript.
2) A list of extracted clinical observations.
3) A list of candidate EHR fields with field ids.

For EACH candidate field, assign a relevance score from 1 to 5 based on how well it is supported by the transcript.

Scoring scale:
- 5 → Explicit, direct evidence in the transcript
- 4 → Strongly implied by transcript wording
- 3 → Weak, partial, or indirect evidence
- 2 → Mentioned but likely not relevant
- 1 → No meaningful support or irrelevant

Important rules:
- Prefer transcript wording over extracted observations.
- Negated findings SHOULD still receive a high score if the field is clinically relevant (negation will be handled in value extraction).
- Historical or uncertain information MAY still be scored if explicitly mentioned.
- When unsure, prefer assigning a LOWER score rather than excluding the field.
- Do NOT assign values.
- Do NOT explain your reasoning.

Output format:
- Return a valid JSON array of objects, one per candidate field:

{
"field_id": "<string>",
"score": <integer between 1 and 5>
}

Ensure every candidate field appears exactly once in the output.

Output ONLY the JSON array.

User prompt:

CLINICAL TRANSCRIPT:
{transcript}

EXTRACTED OBSERVATIONS:
{observations from step 3.2}

**CANDIDATE EHR FIELDS:
{candidate_fields}

Let's think step-by-step.

After scoring, only fields with threshold ($>= 4$) scores are retained and passed to the Evidence Assignment Agent. We note that this filtering is critical in reducing false positives, while maintaining recall for strongly supported clinical findings.

3.4. Evidence Assignment Agent

The Evidence Assignment Agent (an LLM Agent) performs the final value extraction given the full clinical transcript and the filtered candidate fields from the Relevance Scoring Agent. In this the agent is instructed to assign values only when explicit, unambiguous transcript evidence exists. Further following the recent work (Nori et al., 2023; Mehta, 2025), we enforce a mandatory multi-step decision process (spanning explicit evidence identification, controlled field value assignemtn, redundancy resolution etc). To further prevent hallucination and over-generation, the agent is explicitly instructed from inferring implicit clinical meaning or producing partially supported outputs. For categorical fields (e.g., SINGLE_SELECT or MULTI_SELECT), values must be chosen strictly from the provided enumerations. For STRING fields, short phrases are extracted from the transcript without paraphrasing.

System &/ User Prompts

System prompt:
You are a deterministic clinical information extraction system. Your sole function is to map extracted clinical observations to structured EHR fields with maximum precision and zero inference. You must operate as a fixed state machine. Follow the decision process exactly. Skipping or reordering steps is not allowed.

TASK
You will be given: 1) A list of extracted observations, each with a concept and supporting evidence. 2) A list of candidate EHR fields, each with a field id, value type, and allowed values.
Your task is to assign values to EHR fields strictly based on explicit evidence.

MANDATORY DECISION PROCESS

Step 1: Context Scan
Read all observations and candidate fields completely before assigning any value.

Step 2: Definitive Support Test
For each candidate field, check whether there exists a clear, direct, and unambiguous observation that explicitly supports that field.

EXTRACT ONLY IF:
- The observation concept or clinical evidence directly matches the field intent. - The support is explicit and literal.

IGNORE IF:
- The evidence is implied, inferred, summarized, or clinically reasoned.
- The language is tentative, speculative, explanatory, or interpretive.

- The support is partial, indirect, or distributed across multiple observations.
- The field represents a normal, negative, baseline, or default state not explicitly stated.

Step 3: Field Assignment
Assign a value ONLY if the Definitive Support Test is passed. If a field is NOT directly supported by any single observation (either by concept or by clinical evidence), you MUST OMIT it entirely. Absence of evidence is NOT evidence.

Step 4: Redundancy & Conflict Resolution
If multiple candidate fields could encode the same clinical fact, select ONLY the most specific field. Do NOT output multiple fields representing the same information.

Step 5: Abstention Bias
When evidence strength is uncertain, OMIT the field. Omission is always preferred over weak or partial assignment.

VALUE ASSIGNMENT RULES
- Use ONLY the provided observations and candidate fields.
- Do NOT answer partially supported fields.

For SINGLE_SELECT & MULTI_SELECT fields:
- Choose values STRICTLY from the provided value_enum.

For STRING fields:
- Extract a short phrase directly and verbatim from the observation evidence.
- Do NOT combine or paraphrase multiple observations.

STRICT PROHIBITIONS
- Do NOT invent fields, values, or interpretations.
- Do NOT infer clinical meaning beyond what is explicitly stated. - Do NOT infer normal, negative, improved, worsened, or default states.
- Do NOT combine multiple observations to justify a single field.
- Do NOT output field names, value types, enums, confidence, or explanations.

MANDATORY FINAL CHECK (INTERNAL)
Before returning the output, verify:
- Every field is explicitly supported by evidence.
- No inferred or default states are present.
- No redundant or overlapping fields exist.
- Output strictly follows the required JSON format.

Correct any violation before returning the final result.

OUTPUT FORMAT
Return a valid JSON array.
Each item must be exactly:
{ "id": "<field_id>", "value": <value or list of values> }

Output ONLY the JSON array.

User prompt:

You are filling structured EHR fields from extracted observations.

**Observations: {observations}

**Candidate_Fields: {Candidate_fields}

Instructions for each candidate field:

- Assign a value ONLY if supported by observations.
- SINGLE_SELECT / MULTI_SELECT: choose ONLY from value_enum.
- STRING: extract a short phrase from evidence.
- If no evidence exists, OMIT the field.
- Do NOT invent new fields or values.
- Output JSON array:
{ "id": <field_id>, "value": <value or list> }

3.5. Formatting Agent

The Formatting Agent performs post-processing to transform the Evidence Assignment Agent outputs into the final structured format required for submission. The previous stage generates only the minimal structured information (field `id` and assigned `value`), while ontology metadata such as field name and value type (e.g., STRING, SINGLE_SELECT, MULTI_SELECT) remain fixed and are not regenerated by the LLM. Since there exists a one-to-one mapping between each field ID and its associated metadata, we add them to final predictions via post-processing rather than requiring the Evidence Assignment Agent to reproduce static schema information. This design reduces generation errors and enforces schema consistency.

4. Experiments

4.1. Dataset Description

We evaluate our approach on the MEDIQA-SYNUR dataset (Corbeil et al., 2025b), which consists of free-text nurse dictations paired structured EHR observations. The task requires mapping clinical concepts from nurse dictations/transcript to a large ontology (approx. 192) distinct fields, spanning binary, categorical, and free-text value types. The test split provided by the organizers is used for final evaluation and leaderboard submission. We note that no additional external data or annotations are used.

Rank	Participant	Precision	Recall	F1
1	syhwng	0.83	0.80	0.81
2	Rezaul	0.79	0.82	0.80
3	***Smart_solutions***	**0.80**	**0.80**	**0.80**
4	avaliev	0.78	0.82	0.80
5	kyominhwang	0.79	0.81	0.80
6	singhankit16	0.73	0.82	0.77
7	krish_303	0.78	0.77	0.77
8	yudanta	0.70	0.79	0.74
9	msabanluc	0.77	0.66	0.71
10	abrygo	0.61	0.83	0.70
11	Sushvin Marimuthu	0.62	0.72	0.67
12	akabdul	0.53	0.67	0.59
13	av_dx	0.60	0.55	0.57
–	**Task Organizers**	**0.84**	**0.83**	**0.84**

Table 1: Official MEDIQA-SYNUR test set results. Rankings are shown for participating teams only.

4.2. Implementation Details

All the LLM-based agents make use of the widely cited `Kimi-K2-Instruct` (Team et al., 2025) model, which we served using vLLM(Kwon et al., 2023) on Nvidia H200 GPUs. Our approach operates in a zero-shot setting without any task-specific fine-tuning being performed. The Ontology Matching Agent performs retrieval in Stage 2 using TF-IDF–based similarity over field names and descriptions, with the top-k candidates ($k = 15$) retained per observation. We used batch processing where possible to improve throughput, while ensuring that each transcript was processed independently to preserve consistency across stages. We note that the current implementation prioritizes correctness and reliability over latency, and future work will explore optimizations and more efficient model variants to improve deployment feasibility.

4.3. Evaluation Metrics

Official MEDIQA-SYNUR (George Michalopoulos, 2026) evaluated submissions by reporting precision, recall, and F1 scores over extracted EHR fields. The Precision metric measures the proportion of predicted fields that are correct, while recall measures the proportion of gold fields that are successfully extracted, and F1 provides their harmonic mean. Given the safety-critical nature of nursing documentation, balanced precision and recall are emphasized, as both false positives and false negatives can lead to incorrect clinical records.

5. Results

In this section we report the official results of MEDIQA-SYNUR shared task in Tab 1. Our approach (Smart_solutions) ranked at third place among participating teams on the test set, with a precision of 0.80, recall of 0.80, and F1 score of 0.80.

The top participating system achieved an F1 score of 0.81, indicating that our zero-shot multi-agent pipeline performs competitively with only a marginal difference. We note that in the contrast to other approaches that favor recall at the cost of increased false positives, our approach maintains precision without sacrificing coverage, which is particularly important in clinical documentation settings. The tight clustering of scores among the top-performing systems highlights the sensitivity of the task to system design choices. We note that, despite operating in a fully zero-shot setting without any task-specific training, our pipeline achieves competitive performance. This indicates that carefully designed systems built on generalist models can serve as strong, immediately deployable baselines, particularly in settings where fine-tuning is not feasible.

Model	#Parameters	Precision	Recall	F1
GPT-OSS	120B	0.80	0.72	0.76
DeepSeek-V3.1	675B	0.77	0.66	0.71
Kimi-K2 Instruct	**1T**	**0.80**	**0.80**	**0.80**

Table 2: Zero-shot performance comparison across three widely cited open-weight LLMs under identical experimental settings.

To evaluate the impact of model choice under identical pipeline settings, we benchmark three widely cited open-weight LLMs in a fully zero-shot setting. As shown in Table 2, Kimi-K2 Instruct (1T parameters) achieves the strongest overall performance, with balanced precision and recall of 0.80. In comparison, GPT-OSS-120B (Agarwal et al., 2025) achieves competitive precision (0.80) but lower recall (0.72), resulting in an F1 score of 0.76. DeepSeek-V3.1 (Liu et al., 2024) exhibits reduced recall (0.66), leading to an F1 score of 0.71.

These results suggest that larger-scale instruction-tuned models provide improved robustness in structured clinical extraction tasks under zero-shot conditions. Notably, all models were evaluated without task-specific fine-tuning, reinforcing that performance differences arise from model capacity and instruction alignment rather than supervised adaptation.

6. Ablations

To assess the contribution of different stages in our proposed pipeline, we perform a lightweight ablation analysis by selectively removing or modifying components. The goal is to understand how different stages impacts the balance between precision and recall.

The results in Table 3 highlights the importance of decomposing the extraction process into structured stages. The single end-to-end baseline, which jointly performs observation extraction and value assignment, achieves a lower F1 score of 0.70,

Variant	Precision	Recall	F1
Single end to end call	0.73	0.68	0.70
Our Pipeline	**0.80**	**0.80**	**0.80**
w/o Relevance Scoring Agent	0.72	0.79	0.75
w/o Evidence Assignment Constraints	0.76	0.79	0.77

Table 3: Removing relevance scoring leads to increased false positives, while removing evidence constraints results in unsupported assignments.

indicating that conflating relevance determination with factual assignment leads to both missed fields and unsupported predictions.

In contrast, our full pipeline achieves balanced precision and recall (F1 = 0.80), demonstrating the effectiveness of separating these decisions. Removing the Relevance Scoring Agent leads to a significant drop in precision (0.72) while slightly increasing recall (0.82), reflecting the inclusion of unsupported fields due to the absence of explicit filtering, which supports our decision of using relevance scoring agent in final pipeline. Similarly, removing evidence-based constraints in the assignment stage results in reduced precision (0.76), as the model produces weakly grounded or partially supported outputs. Overall, these results provide empirical support for our design choice of multi agent pipelines proposed for this task.

7. Conclusion

We presented a multi-stage LLM based agentic pipeline for nurse observation extraction, as submitted to the MEDIQA-SYNUR 2026 shared task(George Michalopoulos, 2026). By decomposing the task into observation extraction, ontology-level relevance scoring, and abstention-biased value assignment, our system achieves competitive performance while minimizing hallucination and implicit inference.

Our approach placed third (refer Table 1) among participating teams on the MEDIQA-SYNUR test set, with balanced precision and recall of 0.80, despite operating in a fully zero-shot setting without task-specific training. These results demonstrate that careful system design and controlled use of LLMs can effectively address the challenges of ontology-grounded clinical documentation.

The key insight of this work is that relevance judgment and factual assignment should be treated as separate problems in clinical information extraction. Treating LLM based agents as modular decision components, rather than end-to-end generators, enables safer and more interpretable extraction pipelines. We also note that while our pipeline yields strong performance, its reliance on Kimi-K2-Instruct model ($\sim$1T parameters) adds a substantial computational overhead and may limit its practical deployment, especially in low resource settings. Future work will explore addressing such issues, and the extension of this framework to other clinical documentation tasks.

Ethics Statement and Limitations

This work focuses on automated extraction of clinical observations from nursing transcripts in a shared task setting. While the proposed pipeline aims to reduce documentation burden, it is not intended for direct clinical deployment without human oversight. All experiments were conducted on publicly available shared-task data, and no identifiable patient information was accessed outside the task guidelines. Any real-world deployment would require strict adherence to healthcare data privacy regulations and comprehensive clinical validation. Despite achieving competitive performance, several limitations remain. The pipeline relies on a large instruction-tuned model (Kimi-K2), which introduces substantial computational overhead and may limit deployment in resource-constrained environments. The multi-stage design, while improving control and interpretability, increases latency and system complexity. Additionally, the zero-shot setup may be sensitive to domain shifts and variations in clinical language. Finally, the system has not been evaluated in real-world clinical workflows, and its robustness to noisy or incomplete transcripts remains an open question.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Huda Adam, Jialing Lin, Jason Lin, Hannah Keenan, Andrew Wilson, and Marzyeh Ghassemi. 2024. Clinical information extraction with large language models: A case study on organ procurement. In *AMIA Annual Symposium Proceedings*, pages 115–123.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Fernando R Altermatt, Andres Neyem, Nicolás I Sumonte, Ignacio Villagrán, Marcelo Mendoza,

Hector J Lacassie, and Alejandro E Delfino. 2025. Evaluating gpt-4o in high-stakes medical assessments: performance and error analysis on a chilean anesthesiology exam. *BMC Medical Education*, 25(1):1499.

Ehsaneddin Asgari, Natalia Montaña-Brown, Marie Dubois, Sally Khalil, James Balloch, Jonathan A. Yeung, and Daniela Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

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*, Suzhou (China). Association for Computational Linguistics.

Adriano Friganović, Polona Selič, Biljana Ilić, and Biserka Sedić. 2019. Stress and burnout syndrome and their associations with coping and job satisfaction in critical care nurses: a literature review. *Psychiatria Danubina*, 31(Suppl 1):21–31.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiacheng Hu, Runyuan Bao, Yang Lin, Hanchao Zhang, and Yanlin Xiang. 2024. Accurate medical named entity recognition through specialized nlp models. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 578–582. IEEE.

Thelma C Hurd, Fay Cobb Payton, Darryl B Hood, et al. 2024. Targeting machine learning and artificial intelligence algorithms in health care to reduce bias and improve population health. *The Milbank Quarterly*, 102(3):577.

AHM Karim and Ozlem Uzuner. 2025. Assessing large language models for structured medical order extraction. *arXiv preprint arXiv:2510.10475*.

Min Soo Kim, Paul Chung, Nima Aghaeepour, and Namkug Kim. 2025. Information extraction from clinical texts with generative pre-trained transformer models. *International Journal of Medical Sciences*, 22(5):1015–1028.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hui Liu, Ziyi Chen, Peilin Li, Yuan-Zhi Liu, Xiangtao Liu, Ronald X Xu, and Mingzhai Sun. 2025a. Resource-efficient instruction tuning of large language models for biomedical named entity recognition. *Journal of Biomedical Informatics*, page 104896.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng Lai, Thomas Gracie, Ashwin Kishtagari, et al. 2025b. Human level information extraction from

clinical reports with finetuned language models. *Scientific Reports*, 15(1):45239.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Zoltan P Majdik, S Scott Graham, Jade C Shiva Edward, Sabrina N Rodriguez, Martha S Karnes, Jared T Jensen, Joshua B Barbour, and Justin F Rousseau. 2024. Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study. *Jmir ai*, 3:e52095.

Matthew D. McHugh, Ann Kutney-Lee, Jeannie P. Cimiotti, Douglas M. Sloane, and Linda H. Aiken. 2011. Nurses' widespread job dissatisfaction, burnout, and frustration with health benefits signal problems for patient care. *Health Affairs*, 30(2):202–210.

Parth Mehta. 2025. Pnlp at mediqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 75–83.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Vasileios Ntinopoulos, Hector Rodriguez Cetina Biefer, Igor Tudorache, Nestoras Papadopoulos, Dragan Odavic, Petar Risteski, Achim Haeussler, and Omer Dzemali. 2025. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ health & care informatics*, 32(1):e101139.

Cecilia Padilla Fortunatti and Yaritza K. Palmeiro-Silva. 2017. Effort-reward imbalance and burnout among icu nursing staff: A cross-sectional study. *Nursing Research*, 66(5):410–416.

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*.

Soorya Ram Shimgekar, Shayan Vassef, Abhay Goyal, Navin Kumar, and Koustuv Saha. 2025. Agentic ai framework for end-to-end medical data inference. *arXiv preprint arXiv:2507.18115*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Robin Willard-Grace, Margaret Knox, Bruce Huang, Heather Hammer, Chris Kivlahan, and Kevin Grumbach. 2019. Burnout and health care workforce turnover. *Annals of Family Medicine*, 17(1):36–41.

Qianqi Yan, Huy Nguyen, Sumana Srivatsa, Hari Bandi, Xin Eric Wang, and Krishnaram Kenthapadi. 2026. Cite-while-you-generate: Training-free evidence attribution for multimodal clinical summarization. *arXiv preprint arXiv:2601.16397*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Musarrat Zeba, Abdullah Al Mamun, Kishoar Jahan Tithee, Debopom Sutradhar, Mohaimenul Azam Khan Raiaan, Saddam Mukta, Reem E Mohamed, Md Rafiqul Islam, Yakub Sebastian, Mukhtar Hussain, et al. 2025. Mitigating hallucinations in healthcare llms with granular fact-checking and domain-specific adaptation. *arXiv preprint arXiv:2512.16189*.

Xiaoxuan Zhan, Michael Humbert-Droz, Partha Mukherjee, and Olivier Gevaert. 2021. Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*, 2(7):100289.

Zhiyu Zhang and Arbee LP Chen. 2022. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC bioinformatics*, 23(1):458.