MasonNLP at MEDIQA-OE 2025: Assessing Large Language Models for Structured Medical Order Extraction

A H M Rezaul Karim
George Mason University, VA, USA
akarim9@gmu.edu

Özlem Uzuner
George Mason University, VA, USA
ouzuner@gmu.edu

Abstract

Medical order extraction is essential for structuring actionable clinical information, supporting decision-making, and enabling downstream applications such as documentation and workflow automation. Orders may be embedded in diverse sources, including electronic health records, discharge summaries, and multi-turn doctor–patient dialogues, and can span categories such as medications, laboratory tests, imaging studies, and follow-up actions. The MEDIQA-OE 2025 shared task focuses on extracting structured medical orders from extended conversational transcripts, requiring the identification of order type, description, reason, and provenance. We present the MasonNLP submission, which ranked 5th among 17 participating teams with 105 total submissions. Our approach uses a general-purpose, instruction-tuned LLaMA-4 17B model without domain-specific fine-tuning, guided by a single in-context example. This few-shot configuration achieved an average F_1 score of 37.76, with notable improvements in reason and provenance accuracy. These results demonstrate that large, non-domain-specific LLMs, when paired with effective prompt engineering, can serve as strong, scalable baselines for specialized clinical NLP tasks. [1]

1 Introduction

Clinical free-text notes in electronic health records (EHRs) contain essential information such as diagnoses, medications, procedures, and treatment plans (Wang et al., 2018; Demner-Fushman et al., 2009). Extracting structured medical orders, including medications, labs, imaging, and procedures, from such unstructured text is critical for enabling downstream applications like decision support and Computerized Physician Order Entry (CPOE) (Sutton et al., 2020; Kuperman and Gibson, 2003). However, despite the adoption of CPOE systems, errors in order entry persist (Kinlay et al., 2021; Campbell et al., 2006), and medication mistakes often arise during care transitions (Vira et al., 2006). This highlights the need for reliable methods to extract structured medical orders from clinical documentation.

To support such downstream tasks and reduce error rates, clinical information extraction (IE) methods have been developed to automatically identify entities and relations from free-text EHRs (Uzuner et al., 2010; Hahn and Oleynik, 2020). These systems have enabled large-scale mining of clinical concepts for applications such as cohort identification, adverse event detection (ADE), and case surveillance (Sarmiento and Dernoncourt, 2016; Landolsi et al., 2023; Ford et al., 2016). Within this domain, *medical order extraction (MOE)* focuses specifically on identifying medical orders, such as medications, lab tests, or imaging, and structuring them into machine-readable formats (Xu et al., 2010). Automating this process can reduce transcription burden, enhance care quality, and minimize errors in clinical workflows.

The **MEDIQA-OE 2025 Shared Task on Medical Order Extraction (OE)** (Corbeil et al., 2025b) introduced a new benchmark to address this need. The task provides annotated multi-turn doctor–patient conversations and evaluates systems on their ability to extract structured medical orders, including medications, laboratory tests, imaging studies, and follow-up procedures, from conversational transcripts. In addition to identifying the order, systems must also extract the corresponding description and the reason or justification provided by the physician. This reflects real-world clinical documentation scenarios, where accurate interpretation of both the order and its rationale is essential.

In this paper, we describe our participation in the MEDIQA-OE task, which involves identifying and structuring various medical orders and their

[1] Implementation can be found here: https://github.com/AHMRezaul/MEDIQA-OE-2025

Proceedings of the 7th Clinical Natural Language Processing Workshop, pages 57–67
October 30, 2025 ©2025 Association for Computational Linguistics

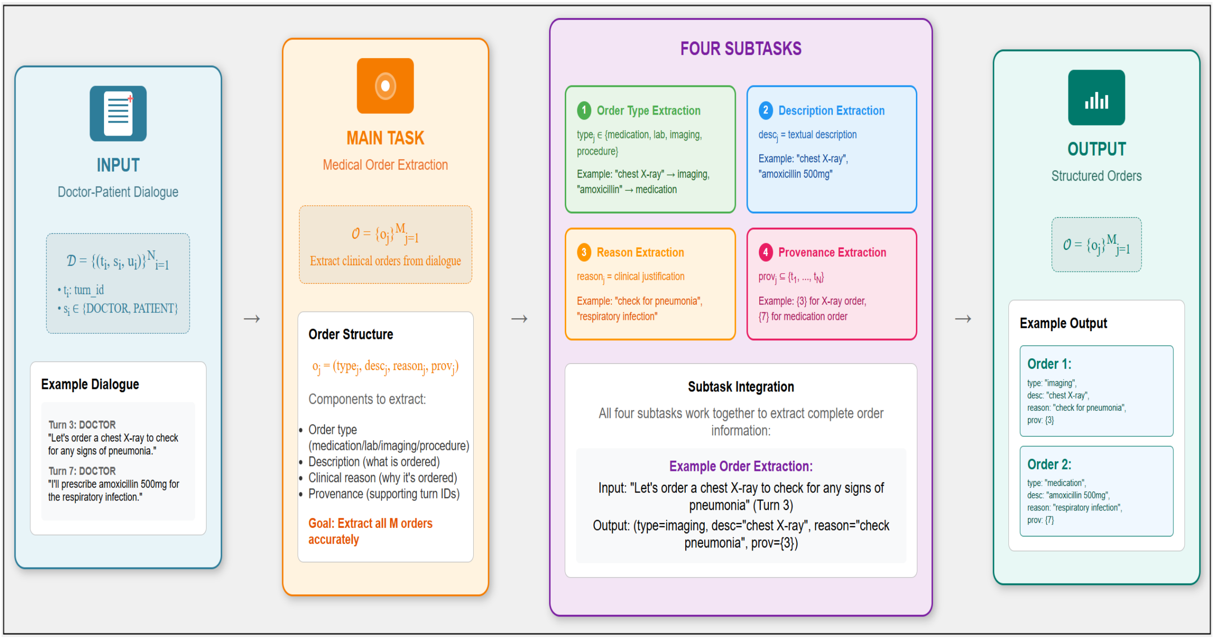

INPUT
Doctor-Patient Dialogue
$\mathcal{D} = \{(t_i, s_i, u_i)\}^N_{i=1}$
• t_i: turn_id
• $s_i \in$ {DOCTOR, PATIENT}
Example Dialogue
Turn 3: DOCTOR
"Let's order a chest X-ray to check for any signs of pneumonia."
Turn 7: DOCTOR
"I'll prescribe amoxicillin 500mg for the respiratory infection."
→
MAIN TASK
Medical Order Extraction
$\mathcal{O} = \{o_j\}^M_{j=1}$
Extract clinical orders from dialogue
Order Structure
o_j = (type$_j$, desc$_j$, reason$_j$, prov$_j$)
Components to extract:
• Order type (medication/lab/imaging/procedure)
• Description (what is ordered)
• Clinical reason (why it's ordered)
• Provenance (supporting turn IDs)
Goal: Extract all M orders accurately
→
FOUR SUBTASKS
1 Order Type Extraction
type$_j$ ∈ {medication, lab, imaging, procedure}
Example: "chest X-ray" → imaging, "amoxicillin" → medication
2 Description Extraction
desc$_j$ = textual description
Example: "chest X-ray", "amoxicillin 500mg"
3 Reason Extraction
reason$_j$ = clinical justification
Example: "check for pneumonia", "respiratory infection"
4 Provenance Extraction
prov$_j$ ⊆ {t_1, ..., t_N}
Example: {3} for X-ray order, {7} for medication order
Subtask Integration
All four subtasks work together to extract complete order information:
Example Order Extraction:
Input: "Let's order a chest X-ray to check for any signs of pneumonia" (Turn 3)
Output: (type=imaging, desc="chest X-ray", reason="check pneumonia", prov={3})
→
OUTPUT
Structured Orders
$\mathcal{O} = \{o_j\}^M_{j=1}$
Example Output
Order 1:
type: "imaging",
desc: "chest X-ray",
reason: "check for pneumonia",
prov: {3}
Order 2:
type: "medication",
desc: "amoxicillin 500mg",
reason: "respiratory infection",
prov: {7}

Figure 1: Overview of the MEDIQA-OE 2025 task: input transcripts, four subtasks (order type, description, reason, provenance), and the final structured output.

reasons with provenance grounding. Our approach uses **Meta's LLaMA-4 Scout 17B** (Meta, 2025) model and relies on *few-shot prompt engineering*, without any domain-specific fine-tuning or external knowledge sources. We curated exemplar prompts that capture conversational structures and medical order patterns. This approach allows us to evaluate the capabilities of general-purpose LLMs on domain-specific extraction tasks.

Our contributions are as follows:

- We assess the feasibility and limitations of prompt-based approaches for structured information extraction in complex, safety-critical clinical domain tasks.
- We investigate the model's reasoning capabilities by analyzing how well it can identify the clinical justification (reason) for prescribed medications, lab tests, imaging, and follow-ups.
- We evaluate the ability of a non-medical, instruction-tuned LLM to perform medical order extraction from clinical text without any domain-specific fine-tuning.

Our findings contribute to the growing body of work comparing general and domain-specific LLMs for clinical applications, highlighting prompting as a lightweight yet effective approach to structured prediction.

2 Related Work

The extraction of structured medical orders from unstructured clinical narratives has been a long-standing challenge in clinical Natural Language Processing (NLP), motivated by its potential to streamline clinical workflows, enhance decision support, and improve patient safety (Lussier et al., 2001; Patrick and Li, 2010; Uzuner et al., 2010, 2011). Early approaches were predominantly rule-based systems leveraging hand-crafted patterns, lexicons, and regular expressions to identify clinical entities and actions. Examples include systems built on platforms such as MedLEE (Friedman, 2000) and MetaMap (Aronson and Lang, 2010), which mapped text spans to controlled vocabularies like the UMLS (Bodenreider, 2004). These methods demonstrated high precision in restricted domains but suffered from limited transferability and scalability across institutions due to variations in clinical language and documentation styles.

The next generation of systems shifted toward statistical and machine learning approaches, which incorporated features from linguistic preprocessing (e.g., tokenization, POS tagging, dependency parsing) into classifiers such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). Early examples in medication extraction, such as the 2009 i2b2 challenge systems (Patrick and Li, 2010), demonstrated improved adaptability over purely rule-based methods, though their reliance on manually engineered features still posed challenges for transferability.

With the advent of deep learning, feature engineering was largely replaced by distributed representations learned directly from Gan et al.. Recurrent Neural Networks (RNNs), particularly LSTMs (Hochreiter and Schmidhuber, 1997) and BiLSTMs (Schuster and Paliwal, 1997), became popular for sequence labeling in clinical NLP, including medication extraction (Jagannatha and Yu, 2016; Huang et al., 2015; Narayanan et al., 2022; Christopoulou et al., 2020). Attention mechanisms and hierarchical architectures further improved the capture of long-range dependencies, which is critical for modeling multi-turn dialogues and long EHR notes.

The introduction of transformer-based models (Vaswani et al., 2017) marked a significant leap in performance. Domain-specific transformers such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and BlueBERT (Peng et al., 2019) fine-tuned on biomedical corpora demonstrated substantial gains in extracting entities and relations from EHR data. These models leveraged self-attention to capture contextual relationships across long sequences, making them highly suitable for MOE from extended clinical narratives.

More recently, large language models (LLMs) such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023) have shown strong zero- and few-shot capabilities across domains, including clinical tasks. While most LLM work in healthcare has focused on general summarization, question answering, and entity recognition (Singhal et al., 2023; Moor et al., 2023), some studies have explored their application to structured MOE (Yang et al., 2020; Peng et al., 2023; Mahajan et al., 2023). These approaches typically involve prompt engineering, in-context learning, or retrieval-augmented generation (RAG), sometimes without any domain-specific fine-tuning, to leverage LLMs' reasoning and language understanding abilities.

Ford et al.; Spasic and Nenadic; Grouin et al. highlight a persistent gap between domain-specific models trained on narrow datasets and general-purpose LLMs that can adapt to new tasks with minimal supervision. While prior research has shown that LLMs trained on biomedical data can excel at extraction tasks, little is known about how non-biomedical, general-purpose LLMs perform in high-stakes domains like MOE when only minimal in-context examples are provided (Landolsi et al., 2023; Peng et al., 2023).

In this study, we position our work within the evolving trajectory from rule-based to LLM-based systems by focusing on the capabilities of a non-domain-specific model for the shared task. Our analysis complements prior work by quantifying how far prompt-engineered, general-purpose LLMs can go in structured clinical IE, without the cost and complexity of large-scale biomedical pretraining or fine-tuning, while identifying key strengths and weaknesses across different MOE subtasks.

3 Task Description

The MEDIQA-OE 2025 Shared Task (Corbeil et al., 2025b) targets the extraction of structured medical orders from extended, doctor–patient conversations. The objective is to streamline clinical documentation, reduce provider workload, and ensure reliable capture of essential patient information from lengthy conversations.

Given a dialogue $\mathcal{D} = \{(t_i, s_i, u_i)\}_{i=1}^{N}$, where t_i is the $\texttt{turn\_id}$, $s_i \in \{\texttt{DOCTOR}, \texttt{PATIENT}\}$ denotes the speaker, and u_i is the utterance text, the goal is to predict a set of medical orders $\mathcal{O} = \{o_j\}_{j=1}^{M}$.

Each medical order o_j is divided into four subtasks as a tuple $(\texttt{type}_j, \texttt{desc}_j, \texttt{reason}_j, \texttt{prov}_j)$, where $\texttt{type}_j \in \{\texttt{medication}, \texttt{lab}, \texttt{imaging}, \texttt{followup}\}$, $\texttt{desc}_j$ is the textual description, $\texttt{reason}_j$ is the clinical justification, and $\texttt{prov}_j \subseteq \{t_1, \ldots, t_N\}$ contains the supporting turn IDs.

Multiple medical orders may be present per dialogue, and systems must extract all relevant orders with accurate structure and provenance grounding. Success in this task requires models to handle long-range dependencies, differentiate between clinically relevant and incidental information, and produce outputs in a consistent, structured format that can be directly integrated into electronic health record (EHR) systems. Figure 1 illustrates the input format, subtask definitions, and expected output.

4 Dataset

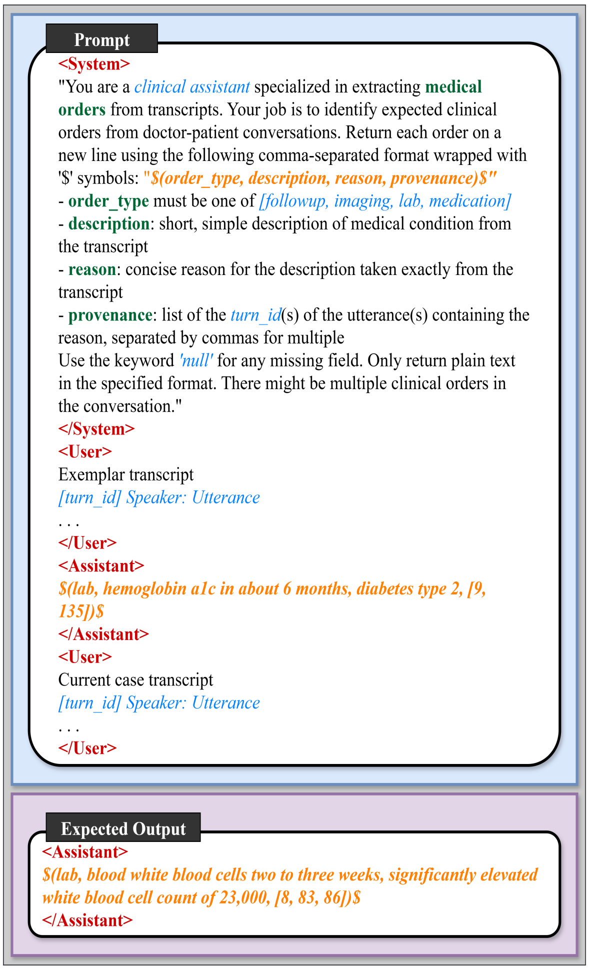
The MEDIQA-OE dataset (Corbeil et al., 2025a) consists of multi-turn doctor–patient conversations annotated with medical orders. Each instance is a JSON object containing an id, a list of expected_orders, and a transcript of turns (t_i, s_i, u_i), where t_i is the turn ID, s_i denotes the speaker, and u_i is the utterance. Orders are annotated as $(\texttt{type}, \texttt{desc}, \texttt{reason}, \texttt{prov})$, with prov representing the supporting turn IDs.

Set	#Enc	Follow-Up	Imaging	Lab	Medication
Train	63	25	14	29	75
Dev	100	41	26	71	117
Test	100	-	-	-	-

Table 1: Number of encounters and order types per set. Gold labels for the test set have not been released.

Transcripts were sourced from the PriMock57 (Papadopoulos Korfiatis et al., 2022) and ACI-Bench (Yim et al., 2023) datasets, with annotations merged using the official preprocessing script. The dataset, derived from the SIMORD corpus with an inter-annotator agreement of *0.768κ*, was curated by experts following post-encounter documentation practices, capturing both explicit and implicit orders that often require multi-turn reasoning.

Dataset Analysis. Table 1 provides the distribution of the dataset with a breakdown of each `order_type`. *Medication* orders dominate in both training and development sets, followed by *lab*, *followup*, and *imaging*, reflecting a clear class imbalance that may bias models toward frequent types in case of model fine-tuning. Dialogues are long, averaging 95.4 turns in training, 102.1 in development, and 101.6 in test, with the longest spanning 290 turns and over 2,900 tokens, posing challenges for models with limited context windows. Across all sets, doctors produce the majority of content; for example, in the test set, they contribute 6,123 turns and 89,449 tokens compared to 4,037 turns and 39,362 tokens from patients. Follow-up suggestions appear in roughly one-third of encounters and align with annotated follow-up orders. Incomplete annotations are also present, with roughly one-fifth of orders lacking a `reason` field. `Provenance` spans averaging only 1–2 turns, making both reason capture and evidence attribution challenging. These long contexts, skewed label distribution, implicit or missing reasons, and brief evidence spans, motivate models that can (i) maintain long-range dialogue state, (ii) generalize with little task-specific supervision, and (iii) ground outputs to cited turns. LLMs can handle extended inputs, adapt with few-shot prompts, and return structured fields with explicit provenance, making them the ideal candidate for this task.



```
Prompt
<System>
"You are a clinical assistant specialized in extracting medical
orders from transcripts. Your job is to identify expected clinical
orders from doctor-patient conversations. Return each order on a
new line using the following comma-separated format wrapped with
'$' symbols: "$(order_type, description, reason, provenance)$"
- order_type must be one of [followup, imaging, lab, medication]
- description: short, simple description of medical condition from
the transcript
- reason: concise reason for the description taken exactly from the
transcript
- provenance: list of the turn_id(s) of the utterance(s) containing the
reason, separated by commas for multiple
Use the keyword 'null' for any missing field. Only return plain text
in the specified format. There might be multiple clinical orders in
the conversation."
</System>
<User>
Exemplar transcript
[turn_id] Speaker: Utterance
. . .
</User>
<Assistant>
$(lab, hemoglobin a1c in about 6 months, diabetes type 2, [9,
135])$
</Assistant>
<User>
Current case transcript
[turn_id] Speaker: Utterance
. . .
</User>

Expected Output
<Assistant>
$(lab, blood white blood cells two to three weeks, significantly elevated
white blood cell count of 23,000, [8, 83, 86])$
</Assistant>
```

Figure 2: Few-shot prompt showing system instructions, exemplar input/output, the input query, and the expected model output.

5 Methodology

We use several general-domain Meta-Llama (Touvron et al., 2023) models for MEDIQA-OE 2025, moving from zero-shot with a smaller model to few-shot with larger ones. We select Meta-Llama for its open weights, long context window, and strong instruction following ability, which is necessary for this task.

Input transcripts from the dataset were converted from JSON to a plain-text format with one utterance per line:

```
[turn_id] Speaker: Utterance
```

This ensured that turn order and speaker roles were preserved for downstream reasoning.

5.1 Model Configurations

We evaluated three LLM configurations:

1. LLaMA 3 Inference (Zero-Shot): First, *meta-llama/Llama-3-8B-Instruct* (Grattafiori et al., 2024) with no in-context examples. The system prompt defined the model's role as a clinical assistant and specified the schema and constraints.

2. LLaMA-4 Inference (Zero-Shot): Next, *meta-llama/Llama-4-Scout-17B-16E-Instruct* (Meta, 2025) with the same prompt design, leveraging a larger model for potentially better reasoning and grounding.

3. LLaMA-4 Inference (Few-Shot): Finally, added a single in-context example from the training set, formatted as a user–assistant exchange preceding the inference case. The assistant's example output illustrated the correct schema and provenance formatting, providing the model with a domain-specific reference.

5.2 Prompt Design

The prompt, illustrated in Figure 2, defined the model's role as a clinical assistant tasked with extracting all medical orders from a doctor–patient conversation. It explicitly described the target output schema and the allowable values for `order_type`, and it stated what each subtask should return. Additional guidance required `null` for missing fields, allowed multiple orders per conversation, and constrained `provenance` to turn IDs for the supporting utterances. The transcript was provided line by line with turn IDs and speaker roles, followed by either no example (zero-shot) or one exemplar with its gold output (few-shot). We reached this final prompt through iterative refinement: an initial version asked only for `order_type`, `description`, `reason`, and `provenance` without role assignment, field definitions, or a fixed format, which led to long free-form text, difficult post-hoc parsing, and generic reasons that did not cite explicit spans. Assigning the clinical-assistant role and explaining each field improved grounding and produced more concise outputs. Requesting strict JSON next proved unreliable, as the model often added extra keys or commentary. We therefore switched to a comma-separated line format that the model followed more consistently. This process ensured clear instructions, a faithful mapping to the schema, and outputs that were both grounded and easy to parse.

5.3 Post-Processing

Following inference, raw model outputs were normalized and structured to match the required format. This process involved removing any extraneous text such as preambles or explanations, ensuring that all four fields were present, and explicitly assigning `null` to missing fields. The `order_type` field was standardized to the allowed set, and the `provenance` field was validated to contain only integer turn IDs within the valid range for each conversation. Outputs were then serialized into JSON for evaluation. When predictions contained minor formatting issues, such as concatenated fields or misplaced delimiters, these were corrected automatically; predictions that could not be repaired were discarded to avoid evaluation errors.

5.4 Experimental Setup

All experiments were conducted on a High-Performance Computing (HPC) environment with NVIDIA A100 80GB GPUs using mixed-precision (`bfloat16`) inference to optimize memory usage and runtime. The maximum context length was set to 8,192 tokens with a generation limit of 1,024 tokens. Decoding parameters were fixed across runs (`temperature=0.2`, `top_p=0.9`) to balance determinism and variability. Random seeds were fixed across the model, tokenizer, and generation routines for reproducibility.

5.5 Evaluation

System outputs are first aligned to gold-standard orders through a description-based pairing process. Matching is performed on the `description` field after normalization, which lowercases text and removes selected punctuation. Orders are excluded from evaluation if they have an empty `description` or an `order_type` outside the allowed set `medication`, `lab`, `followup`, `imaging`.

Once aligned, each field is scored with a metric suited to its content. The `description` field is evaluated with ROUGE-1 F_1, rewarding unigram overlap with the reference and granting partial credit for preserving key clinical terms even if phrasing differs. The `reason` field is also scored with ROUGE-1 F_1, capturing semantic similarity despite surface variation in justifications. The `order_type` field uses a STRICT F_1, counting only exact matches among the four permissible categories to penalize misclassification. The `provenance` field is evaluated with a MULTILABEL F_1, treating provenance

Team Name	description	reason	order_type	provenance	avg_score
MasonNLP	39.05	19.78	50.91	41.32	37.76
WangLab	66.77	29.49	81.45	63.04	60.19
silver_shaw	64.06	41.30	74.74	60.44	60.14
MISo KeaneBeanz	57.99	35.64	71.56	48.38	53.39
EXL Health AI Lab	54.45	30.50	66.17	52.47	50.90
HerTrials	19.61	8.99	29.59	5.61	15.95

Table 2: MEDIQA-OE 2025 leaderboard results (F1 in %). Top six systems, rows sorted by average score; MasonNLP shown first for reference.

System	description	reason	order_type	provenance	avg_score
LLaMA-3 8B (Zero-shot)	30.20	13.95	40.79	27.10	28.01
LLaMA-4 17B (Zero-shot)	36.82	15.60	47.23	30.32	32.49
LLaMA-4 17B (Few-shot)	**39.05**	**19.78**	**50.91**	**41.32**	**37.76**

Table 3: Performance across experimental setups (F1 in %). Best values are in bold.

as a set of turn IDs and balancing precision (excluding unrelated turns) with recall (capturing all relevant turns).

The final shared-task score is the unweighted mean of the four primary field-level F_1 scores (description_ROUGE1_f1, reason_ROUGE1_f1, order_type_Strict_f1, and provenance_MultiLabel_f1).

6 Results and Discussion

6.1 Leaderboard Performance

The MEDIQA-OE 2025 shared task attracted participation from **17 teams**, producing a total of **105 submissions**. Our **MasonNLP** system, based on a few-shot prompting setup with the general-purpose LLaMA-4 17B model and no domain-specific fine-tuning, achieved an average F_1 score of **37.76**, placing competitively among the top-ranked systems. Table 2 presents the top six leaderboard with subtask-specific scores, with our system listed first for clarity. Notably, this performance was obtained without incorporating clinical-domain pretraining or retrieval augmentation, competing against systems that leveraged specialized architectures or domain-specific resources.

6.2 Ablation Study

To better understand the impact of model scale and prompting strategy, we evaluated three configurations: LLaMA-3 8B zero-shot, LLaMA-4 17B zero-shot, and LLaMA-4 17B few-shot (final submission). Results in Table 3 show a clear progression in average F_1 across configurations. Moving from LLaMA-3 to LLaMA-4 improved performance in all subtasks, especially description and order_type, reflecting the larger model's stronger capacity for identifying and categorizing medical orders in long transcripts. This aligns with the dataset's high average turn count and doctor-heavy content, which demand robust long-context processing. Introducing a single in-context example further improved all four subtasks, with the largest relative gain in provenance, suggesting that even minimal task-specific guidance helps the model ground predictions more accurately and follow the required structured format.

6.3 Discussion and Implications

These findings confirm our initial hypothesis that larger, instruction-tuned LLMs provide measurable benefits for MOE from long, multi-turn dialogues, even without domain-specific fine-tuning. The improvements from the few-shot configuration validate our contribution, which claims that minimal in-context supervision can close part of the performance gap between general-purpose LLMs and domain-adapted systems. However, reason extraction remains the most challenging subtask, likely due to the implicit nature of many clinical justifications in the dataset. Similarly, while provenance accuracy improved, grounding still lags behind other subtasks, reflecting the difficulty of linking orders to scattered and sometimes indirect evidence in the dialogue.

Overall, the results suggest that combining large general-purpose LLMs with carefully designed prompts and minimal in-context examples can yield competitive performance in structured clinical IE. Future gains may require integrating retrieval-based grounding or domain adaptation to better handle implicit reasoning and improve evidence alignment.

7 Error Analysis

Building on the results in Section 6, we conducted a detailed error analysis of our best-performing `LLaMA-4 17B` few-shot system to better understand its strengths and remaining challenges across the four subtasks. The development set offers gold-standard annotations for all fields, enabling both quantitative and qualitative assessment. The test set, lacking gold-standard annotations, is analyzed only for schema validity.

Metric	Score
`description_ROUGE1_f1`	44.53
`reason_ROUGE1_f1`	25.13
`order_type_Strict_f1`	57.28
`provenance_MultiLabel_f1`	40.17
`avg_score`	41.78

Table 4: Development set scores for the LLaMA-4 17B few-shot system (F1 in %).

7.1 Development Set Analysis

Table 4 summarizes the official shared-task metrics for the development set. Consistent with leaderboard results, the model showed strong performance in `order_type` classification and `description` extraction, while `reason` and `provenance` remained more challenging. The model also broke down a single order into multiple orders in some cases, as illustrated in Figure 3. To explore why, we examined a few samples of matched and unmatched predictions, categorizing representative patterns for each subtask.

Description. The model was able to identify the correct target of an order in most cases, even in multi-turn, context-heavy transcripts. Many predictions contained the correct general test or medication, but lacked finer details such as timing or exact test subtype. For example, *"blood work"* was produced for the gold-standard *"blood white blood cells two to three weeks"*. This indicates that the model successfully locates the core clinical action but sometimes omits modifiers, an area that could be enhanced by incorporating temporal and entity-specific cues.

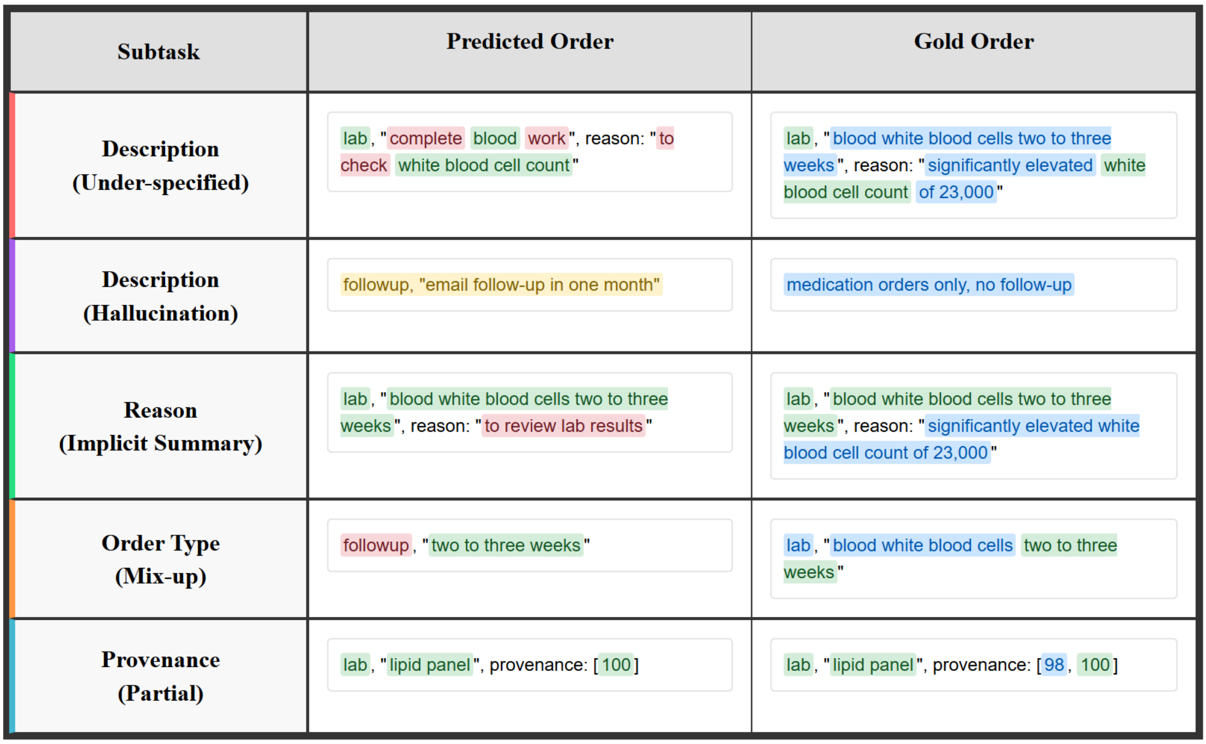
Reason. In most cases, the model provided a plausible reason aligned with the overall clinical context. For instance, it correctly linked a lab order to white blood cell count monitoring, though it occasionally summarized the reason more generally (*"to review lab results"*) instead of including explicit values. This shows that the model is capable of long-context integration to capture the essence of clinical justification, with potential for refinement through methods that encourage inclusion of specific numeric and temporal evidence.

Order Type. Order type classification was generally strong, but certain linguistic patterns led to confusion. Scheduling phrases (e.g., *"two to three weeks"*) were sometimes interpreted as follow-up visits rather than scheduled labs. Invalid `order_type`, present in 8 instances, included mentions of *surgery (3), referral (1), and null_type (4)*. Such mix-ups likely arise when multiple order-like actions occur in close proximity, and can be addressed by fine-tuning with examples emphasizing subtle category distinctions.

Provenance. The model demonstrated the ability to identify at least one correct evidence turn for most orders, as in the case where it predicted provenance `[100]` while the gold-standard label included both `[98, 100]`. This partial grounding suggests that the model can reliably find the key confirmation turn, but may miss earlier reason turns when information is distributed. Expanding its retrieval capacity for dispersed evidence could close this gap.

7.2 Test Set Analysis

Out of all predicted orders, 20 (4.7%) lacked a `description`, 8 (1.9%) contained an invalid `order_type` that included the same three keywords as we saw in the predicted orders of development set *(surgery, referral, null_type)*, 57 (13.3%) were missing a `reason`, and 46 (10.8%) omitted `provenance` identifiers. These results show that the system generally produces well-structured outputs with relatively few schema violations, though systematic omissions and field-level incompleteness directly reduce evaluation scores. Invalid `order_type` predictions typically arose from ambiguous dialogue phrasing that led the model to select categories outside the permitted set `medication`, `lab`, `followup`, `imaging`. For `description`, beyond the 20 missing fields, 11 cases involved text not present in the transcript,



Subtask	Predicted Order	Gold Order
Description (Under-specified)	lab, "complete blood work", reason: "to check white blood cell count"	lab, "blood white blood cells two to three weeks", reason: "significantly elevated white blood cell count of 23,000"
Description (Hallucination)	followup, "email follow-up in one month"	medication orders only, no follow-up
Reason (Implicit Summary)	lab, "blood white blood cells two to three weeks", reason: "to review lab results"	lab, "blood white blood cells two to three weeks", reason: "significantly elevated white blood cell count of 23,000"
Order Type (Mix-up)	followup, "two to three weeks"	lab, "blood white blood cells two to three weeks"
Provenance (Partial)	lab, "lipid panel", provenance: [100]	lab, "lipid panel", provenance: [98, 100]

Figure 3: Examples of different error types for each subtask.

reflecting hallucination or paraphrasing of plausible but unsupported orders. For reason, omission was dominant, with 57 missing values and 4 ungrounded justifications, indicating persistent difficulty in capturing implicit or distributed reason. For provenance, there are 46 missing spans. Partial grounding, common in the development set, likely persists here, underscoring the need for stronger evidence attribution.

Overall Observations. The analysis shows that instruction-tuned LLMs, even without domain-specific fine-tuning, can handle complex, multi-turn clinical dialogues to extract actionable orders with reasonable accuracy. While finer details (e.g., exact timing, numeric values, dispersed evidence) are sometimes omitted, the model frequently identifies the correct order, reason, and at least one key supporting turn. Hallucinations, as with most LLMs, are still present, highlighting the potential benefit of RAG (Lewis et al., 2020). With targeted enhancements, these strengths can be leveraged to develop robust clinical NLP systems capable of supporting real-world documentation workflows.

8 Conclusion

We addressed medical order extraction from multi-turn doctor–patient conversations using general-domain Meta-Llama models without domain-specific fine-tuning. The setup began with zero-shot prompting on a smaller model and then moved to few-shot prompting on larger models. A simple structured prompt returned order type, description, reason, and provenance with cited turns. The error analysis shows that the model struggles with temporal and numeric specificity, occasional hallucination, under-specific reasons, and partial provenance spans. These gaps narrowed with larger models and a few clear exemplars. The findings indicate that general domain LLMs are a viable base when guided by domain cues, retrieval to reduce hallucination, and schema validators for strict JSON. Overall, the study shows that instruction-tuned LLMs can handle long clinical dialogues with minimal adaptation and provides a practical template for grounded multi-field clinical IE with clear next steps on reason modeling, tighter provenance, and better balance across order types.

Limitations

Our approach avoids any domain-specific pretraining or fine-tuning on clinical corpora. While integrating such specialization could potentially yield further gains, our goal was to assess the adaptability of a general-purpose, instruction-tuned LLM in a highly specialized medical order extraction task using only prompt engineering. This choice enables a fair evaluation of the model's zero- and few-shot capabilities, providing insights into its out-of-the-box performance without reliance on costly domain-specific data or retraining. The strong results achieved by our few-shot LLaMA-4 system demonstrate that competitive baselines can be established under these conditions, laying the groundwork for future enhancements through targeted domain adaptation.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American medical informatics association*, 17(3):229–236.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Emily M Campbell, Dean F Sittig, Joan S Ash, Kenneth P Guappone, and Richard H Dykstra. 2006. Types of unintended consequences related to computerized provider order entry. *Journal of the American Medical Informatics Association*, 13(5):547–556.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeeson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025a. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. *CoRR*.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025b. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient conversations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270.

Qiwei Gan, Mengke Hu, Kelly S Peterson, Hannah Eyre, Patrick R Alba, Annie E Bowles, Johnathan C Stanley, Scott L DuVall, and Jianlin Shi. 2023. A deep learning approach for medication disposition and corresponding attributes extraction. *Journal of biomedical informatics*, 143:104391.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Cyril Grouin, Natalia Grabar, and 1 others. 2023. Year 2022 in medical natural language processing: availability of language models as a step in the democratization of nlp in the biomedical area. *Yearbook of Medical Informatics*, 32(01):244–252.

Udo Hahn and Michel Oleynik. 2020. Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, 29(01):208–220.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Abhyuday N Jagannatha and Hong Yu. 2016. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing*.

conference on empirical methods in natural language processing, volume 2016, page 856.

Madaline Kinlay, Wu Yi Zheng, Rosemary Burke, Ilona Juraskova, Rebekah Moles, and Melissa Baysari. 2021. Medication errors related to computerized provider order entry systems in hospitals and how they change over time: a narrative review. *Research in Social and Administrative Pharmacy*, 17(9):1546–1552.

Gilad J Kuperman and Richard F Gibson. 2003. Computer physician order entry: benefits, costs, and issues. *Annals of internal medicine*, 139(1):31–39.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2023. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yves A Lussier, Lyudmila Shagina, and Carol Friedman. 2001. Automating snomed coding using medical language underchen2020jointstanding: a feasibility study. In *Proceedings of the AMIA Symposium*, page 418.

Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432.

AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Sankaran Narayanan, Kaivalya Mannam, Pradeep Achan, Maneesha V Ramesh, P Venkat Rangan, and Sreeranga P Rajan. 2022. A contextual multi-task neural approach to medication and adverse events identification from clinical text. *Journal of biomedical informatics*, 125:103960.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527.

Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.

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*.

Raymond Francis Sarmiento and Franck Dernoncourt. 2016. Improving patient cohort identification using natural language processing. *Secondary analysis of electronic health records*, pages 405–417.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

T Vira, M Colquhoun, and E Etchells. 2006. Reconcilable differences: correcting medication errors at hospital admission and discharge. *BMJ Quality & Safety*, 15(2):122–126.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and 1 others. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American medical informatics Association*, 17(1):19–24.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.