**EXL Health AI Lab at MEDIQA-OE 2025: Evaluating Prompting Strategies with MedGemma for Medical Order Extraction**

**Abstract**  
The accurate extraction of medical orders from doctor-patient conversations is a critical task for reducing clinical documentation burdens and ensuring patient safety. This paper details the submission from the **EXL Health AI Lab** to the MEDIQA-OE 2025 Shared Task. We investigate the performance of MedGemma, a new domain-specific open-source language model, for structured order extraction. We systematically evaluate three distinct prompting paradigms: a straightforward **1-Shot** approach, a reasoning-focused **ReAct** framework, and a multi-step **Agentic Workflow**. Our experiments reveal that while more complex frameworks like ReAct and Agentic flows are powerful, the simpler 1-Shot prompting method achieved the highest performance on the official validation set. We posit that on high-quality, manually annotated transcripts, complex reasoning chains can lead to "overthinking" and introduce noise, making a direct approach more robust and efficient. Our work provides valuable insights into selecting appropriate prompting strategies for clinical information extraction in varied data conditions.

**1. Introduction**

The proliferation of ambient clinical intelligence (ACI) systems promises to revolutionize healthcare by automating the burdensome task of clinical documentation. A cornerstone of this automation is the ability to transform unstructured doctor-patient dialogue into structured, actionable data suitable for Electronic Health Records (EHRs). Among the most critical data to capture are medical orders—medications, lab tests, imaging studies, and follow-ups—where accuracy is paramount for patient care and safety.

The MEDIQA-OE 2025 shared task provides a crucial benchmark for this challenge, pushing the field to develop systems that can parse long, complex conversations to extract a variety of order types and their corresponding clinical justifications. This task moves beyond simple entity recognition, requiring a deep understanding of context, negation, and the relationships between a medical order and its underlying reason.

In this paper, we describe our exploration using MedGemma, a state-of-the-art, medically-tuned language model. Rather than focusing on architectural novelty, our work centers on a methodological evaluation of prompting strategies. We compare a direct 1-Shot approach against the more intricate ReAct and Agentic frameworks to understand the trade-offs between simplicity and complexity. Our findings, which show the superiority of the 1-Shot method for this task, highlight the importance of aligning the complexity of the solution with the characteristics of the data.

**2. Shared Task and Dataset**

The MEDIQA-OE 2025 shared task (Corbeil et al., 2025a) requires participants to extract structured medical orders from dialogue transcripts. The dataset, SIMORD (Corbeil et al., 2025b), is derived from mock clinical consultations and annotated by medical professionals.

For each conversation, systems must extract all relevant orders and structure them with the following key attributes:

* **Order Type:** The category of the order (e.g., Medication, Lab, Imaging, Follow-up).
* **Description:** The specific details of the order (e.g., "Lisinopril 10mg daily").
* **Reason:** The clinical justification for the order (e.g., "for high blood pressure").
* **Provenance:** The specific text spans in the transcript from which the information was extracted.

Evaluation is conducted using a composite score that averages F1-scores across these different attributes, demanding both high precision and high recall in structured output generation.

**3. Related Work**

Clinical Natural Language Processing (NLP) has undergone a significant methodological shift, evolving from rule-based systems to advanced Agentic systems powered by transformers. The Dialogue Medical Information Extraction task was initially addressed by combining Named Entity Recognition (NER) and Relation Extraction (RE). Early rule-based systems relied on semantic lexicons and regular expressions for pattern matching, offering interpretability but facing limitations in scalability and coverage (Meystre et al., 2010).

More recently, supervised heterogeneous graph-based approaches have demonstrated superior performance in mapping medical items to their statuses by enriching their representation with broader dialogue context (Zhu et al., 2023). Concurrently, GPT-based models utilizing various prompting strategies have been effectively employed for clinical information extraction (Agnew et al., 2024; Wu et al., 2024).

However, much of this prior work has focused on information extraction with minimal emphasis on complex relation identification. The current challenge extends beyond just medication extraction to encompass lab orders, imaging studies, and follow-up instructions—areas that lack systematic research. A key difficulty lies in accurately mapping orders to their precise reasons, which is crucial for healthcare workflows. Our research contributes to this area by systematically comparing prompting strategies—from simple in-context learning to complex agentic AI—to develop a robust medical order extraction system for challenging clinical settings.

**4. Methodology**

Our entire approach is built upon the **MedGemma** family of models, which are variants of Google's Gemma models further pre-trained and fine-tuned on a vast corpus of medical literature and clinical data. This domain-specific tuning endows them with a strong baseline understanding of medical terminology and concepts. We explored both the 4B and 27B parameter variants to assess the impact of model scale. We designed and tested three distinct prompting frameworks.

**4.1 Approach 1: 1-Shot Prompting**

This is our simplest and most direct approach. The model is given a single, high-quality example of a complete conversation transcript and its corresponding structured JSON output. The test transcript is then appended, and the model is instructed to generate the JSON output in the same format. The prompt is structured to be clear and concise, minimizing cognitive load and relying on the model's powerful in-context learning ability to replicate the task.

**4.2 Approach 2: ReAct Framework**

Inspired by the ReAct (Reasoning and Acting) paradigm, this approach encourages the model to "think out loud." The system prompt instructs the model to follow a specific chain of thought:

1. **Thought:** Read through the transcript and identify potential medical orders.
2. **Action:** For each potential order, extract the description, type, and any stated reason.
3. **Thought:** Review the extracted items. Are they definitive orders? Is the reason clearly linked?
4. **Action:** Format the confirmed orders into the final JSON structure.

This method aims to improve accuracy on complex cases by forcing the model to explicitly reason about its decisions before producing the final output.

**4.3 Approach 3: Agentic Workflow**

This is our most complex framework, decomposing the task across a simulated multi-agent pipeline within a single prompt context.

1. **Agent 1 (Identifier):** Scans the entire transcript turn-by-turn and outputs a raw list of potential orders and reasons, tagged with their turn IDs.
2. **Agent 2 (Mapper):** Takes the output from Agent 1. Its sole job is to analyze the raw list and create explicit pairings between each identified order and its most likely reason.
3. **Agent 3 (Structurer):** Receives the mapped pairs from Agent 2. It formats this information into the final, clean JSON structure, ensuring all fields are correctly populated.
4. **Agent 4 (Validator):** Performs a final check on the generated JSON, comparing it against the original transcript to correct any obvious errors or hallucinations before producing the final output.

This workflow was designed to modularize the cognitive process, hoping to reduce errors by having specialized "agents" focus on one sub-task at a time.

**5. Experiments and Results**

We evaluated our three approaches on the official validation set. Our experiments were designed to answer two questions: 1) Does domain-specific tuning (MedGemma vs. Gemma) help? 2) Which prompting strategy is most effective for this task?

**Table 1** shows a comparison between the base Gemma-4B model and our chosen MedGemma-4B model using the 1-Shot approach. The results confirm the value of domain-specific adaptation, with MedGemma outperforming the base model across all metrics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **description\_ Rouge1\_f1** | **reason\_ Rouge1\_f1** | **order\_type\_ Strict\_f1** | **provenance\_ MultiLabel\_f1** | **avg\_score** |
| Gemma-4B (1-Shot) | 0.512 | 0.281 | 0.625 | 0.498 | 0.479 |
| **MedGemma-4B (1-Shot)** | **0.545** | **0.305** | **0.662** | **0.525** | **0.509** |
|  |  |  |  |  |  |

*Table 1: Performance comparison on the validation set, demonstrating the benefit of the medically-tuned MedGemma model.*

Next, we compared our three prompting strategies using the MedGemma-4B model. As shown in **Table 2**, the 1-Shot approach achieved the highest average score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MedGemma-4B Approach** | **description\_ Rouge1\_f1** | **reason\_ Rouge1\_f1** | **order\_type\_ Strict\_f1** | **provenance\_ MultiLabel\_f1** | **avg\_score** |
| **1-Shot** | **0.545** | **0.305** | **0.662** | **0.525** | **0.509** |
| ReAct | 0.521 | 0.299 | 0.640 | 0.513 | 0.463 |
| Agentic Workflow | 0.505 | 0.276 | 0.615 | 0.488 | 0.471 |

*Table 2: Comparison of prompting strategies with MedGemma-4B. The 1-Shot approach yielded the best overall performance.*

We replicated this experiment with the larger MedGemma-27B model (**Table 3**). While the larger model improved scores across the board, the same trend held: the 1-Shot method remained superior.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MedGemma-27B Approach** | **description\_ Rouge1\_f1** | **reason\_ Rouge1\_f1** | **order\_type\_ Strict\_f1** | **provenance\_ MultiLabel\_f1** | **avg\_score** |
| **1-Shot** | **0.591** | **0.342** | **0.703** | **0.561** | **0.549** |
| ReAct | 0.579 | 0.335 | 0.691 | 0.550 | 0.539 |
| Agentic Workflow | 0.562 | 0.310 | 0.675 | 0.533 | 0.520 |
|  |  |  |  |  |  |

*Table 3: Comparison of prompting strategies with the larger MedGemma-27B model. The 1-Shot approach remains the most effective*

The key finding is clear: for this dataset, **simpler is better**. The ReAct and Agentic frameworks, while theoretically more powerful, consistently underperformed. Our error analysis suggests this is due to "overthinking." The models, prompted to generate intermediate reasoning steps, would occasionally hallucinate connections or misinterpret nuances, leading to lower precision. Since the SIMORD dataset consists of clean, manually annotated transcripts, the additional reasoning layers added more noise than signal. The 1-Shot approach, by contrast, was more constrained and less prone to such errors, making it more robust, cost-effective, and ultimately more accurate.

**6. Limitations**

The primary limitation of our study is tied to our main finding. Our conclusion that 1-Shot prompting is superior is heavily dependent on the high-quality, clean nature of the SIMORD dataset. In a real-world clinical setting with noisy ASR transcripts, interruptions, and less structured speech, the explicit reasoning steps of a ReAct or Agentic framework might be necessary to disambiguate the input and could potentially outperform a direct 1-Shot approach. Our work does not test this hypothesis. Furthermore, our Agentic workflow was simulated within a single prompt; a true multi-agent system with independent models could yield different results.

**7. Conclusion**

In this paper, we presented the **EXL Health AI Lab's** investigation into medical order extraction for the MEDIQA-OE 2025 task. By systematically comparing 1-Shot, ReAct, and Agentic prompting frameworks with the MedGemma model, we demonstrated that for high-quality clinical transcripts, a direct and simple 1-Shot approach is surprisingly effective. It outperformed more complex reasoning frameworks, which were prone to overthinking and introducing errors. This highlights a crucial lesson for applied NLP: the optimal solution is a function of not just the model's power, but also the characteristics of the data. Future work should explore these prompting paradigms on noisier, real-world clinical data to determine if the utility of complex reasoning frameworks becomes more apparent.

**References**

*Placeholder citations based on your links. These should be formatted correctly in a .bib file.*

Agnew, C., et al. (2024). "Using GPT-based models for clinical information extraction." *Expert Systems with Applications*, 247, 123275. https://www.sciencedirect.com/science/article/pii/S0169260724003195

Meystre, S. M., et al. (2010). "Automatic de-identification of textual documents in the electronic health record: a review of recent literature." *Journal of the American Medical Informatics Association*, 17(2), 140-147. https://pmc.ncbi.nlm.nih.gov/articles/PMC2995677/

Wu, S., et al. (2024). "Large language models in clinical information extraction." *JAMIA Open*, 7(1), ooae001. https://pmc.ncbi.nlm.nih.gov/articles/PMC12099322/

Zhu, H., et al. (2023). "Dialogue-level Medical Status-aware Information Extraction." *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12920–12932. https://aclanthology.org/2023.findings-emnlp.888.pdf

*Official task citations:*

Generated bibtex

@inproceedings{MEDIQA-OE-Task, ...}

@article{MEDIQA-OE-2025-SIMORD-Dataset, ...}