**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: In Context Learning (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 descriptions, 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 relation between each identified order and its most likely descriptions.
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.