Leveraging Large Language Models for Clinical Medication Verification

BT5153 Applied Machine Learning in Business Analytics – Group 4
Ng JingYi Shanice (A0142593A) Ng Qi Hao Joshua (A0173219E) Sun Tian Yu (A0167233H))
Steven Harta Prawira (A0127702M) Yan Siyang (A0177911Y)
Github Link: https://github.com/shanicenjy/BT5153_Group04_2025

Abstract

This project develops an LLM-powered solution to assess medication order safety by analyzing patient profiles and drug monographs. We explored Retrieval-Augmented Generation, few-shot full-context prompting, and supervised fine-tuning, using comprehensive metrics to assess both classification accuracy and reasoning quality. Our solution reduces pharmacist verification workload while maintaining medication safety through high-recall detection of unsafe orders and interpretable justifications.

1. Introduction

1.1. Background

Hospital pharmacists play a pivotal role in safeguarding patient safety by participating in ward rounds and verifying that prescribed medications are appropriate and effective (Babu et al., 2023). Their responsibilities extend beyond medication order verification to include dispensing, patient counseling, procurement, and responding to drug-related inquiries (Studer et al., 2023). However, the demanding nature of hospital environments can make it challenging to consistently identify subtle medication interactions or context-specific risks.

Medication errors remain a significant concern within healthcare systems worldwide. A medication error is defined as a failure in the medication-use process that may or may not result in patient harm (Rodziewicz et al., 2024). According to the World Health Organization (WHO), medication-related harm affects 1 in every 30 patients receiving healthcare, with more than a quarter of these cases classified as severe or life-threatening. Notably, half of all avoidable harm in healthcare is related to medication errors (World Health Organization, 2024; Tariq et al., 2024).

In Singapore, the current pharmacy systems aggregate various types of patient data — including medical history, allergy status, and visit and medication order information — from multiple sources. When a medication order is placed, pharmacists must manually cross-reference patient information and external drug databases to identify potential drug interactions, duplicate therapies, or contraindications. This manual process increases the risk of human error and missed safety alerts. Studies have shown that nearly 50% of all medication errors occur during the prescribing or ordering stages, and nurses and pharmacists detect between 30% to 70% of these errors.

Without an automated mechanism to proactively flag potential risks, patients remain vulnerable to preventable medication-related complications. The integration of advanced technologies, such as Large Language Models (LLMs), into the medication order verification process offers the potential to enhance the accuracy and efficiency of pharmacists' reviews, ultimately improving patient safety outcomes.

1.2. Objective

This project aims to develop a Large Language Model (LLM)-powered solution to assess the safety of medication orders by analyzing specific patient profiles alongside drug monograph information. The system is designed to output: (i) a classification of each medication order as either 'safe' or 'unsafe' and (ii) a justification for the decision, detailing the rationale behind the classification.

Safe orders will be automatically verified and removed from the pharmacist's work-list, while unsafe orders will be flagged for manual review, accompanied by the reasons for failing auto-verification. By reducing the verification workload, pharmacists can devote more time to complex cases and high-value clinical interventions. This approach not only enhances decision support but also minimizes medication-related risks, ultimately improving patient safety.

2. Dataset Preparation

2.1. Dataset Description

2.1.1. Drug Monograph

This project utilized the open-source openFDA Drug Product Labeling API to retrieve comprehensive drug monograph information. The API provides access to structured data submitted by drug manufacturers and distributors, encompassing both prescription and over-the-counter medications. The drug labels are segmented into sections such as indications for use, adverse reactions, and dosage instructions, facilitating targeted information retrieval. The drug labeling is considered a "living document," updated over time to reflect new information regarding a drug's safety and effectiveness. According to openFDA, the API data is refreshed on a weekly basis (Food & Administration, 2025). For this project, API calls were made for 27 unique medications.

2.1.2. Patient and Medication Order

Due to the sensitive and confidential nature of patient data, accessing real-world clinical datasets is often time-consuming, costly, and subject to strict privacy regulations. To address these challenges, synthetic data was used in this project. Synthetic datasets offer several key advantages, including enabling early-stage development, hypothesis testing, and method validation without the need for immediate access to real patient information. This approach not only preserves privacy but also enhances research reproducibility (Kokosi & Harron, 2022). Synthetic patient profiles and medication orders were therefore generated to support the goals of this project while ensuring compliance with privacy standards. These synthetic profiles include demographic information (e.g., height, weight, latest vitals), past medical history, allergy status, concomitant medications, and relevant laboratory or imaging results.

The patient profiles were carefully curated by a member of our project team, a practicing clinical pharmacist with over five years of hospital experience. Real-world patient information was adapted to cover a diverse range of cases (e.g., pediatrics, geriatrics) and various medication-related risks. This approach ensured robust stress-testing of the solution against realistic clinical scenarios.

For the medication order dataset, the features consists of the route, medication name, formulation, dosing and frequency. Each medication order is labeled as "safe" or "unsafe," along with the pharmacist's reasoning behind the classification. Given the time and expertise required to curate these cases, the dataset is limited to 50 medication orders. Our dataset consists of 23 safe medication orders and 27 unsafe medication orders. Table 1 below describes the patient information while Table 2 describes the medication order information.

Feature	Description
patient_id	Patient ID
age	Age
sex	Gender
height_m	Height in metres
weight_kg	Weight in kg
hr_latest	Latest heart rate
sbp_latest	Latest systolic blood pressure
allergy	Allergy status
pmhx	Past medical history
issue	Active Issue for currently
wbc	White blood cell count
hb	Hemoglobin levels
plt	Platelet count
high_sens_crp	C-reactive protein levels
urea	Urea levels
crcl	Creatinine Clearance
egfr	Estimated glomerular filtration rate
sodium	Blood sodium levels
potassium	Blood potassium levels
magnesium	Blood magnesium levels
calcium	Blood calcium levels
bicarb	Bicarbonate levels
ast	Aspartate aminotransferase levels
alt	Alanine aminotransferase levels
albumin	Albumin levels
ck	Creatinine Kinase
glucose	Fasting blood glucose levels
hba1c	Hemoglobin A1C
hdl	High-density lipoprotein
ldl	Low-density lipoprotein
tg	Ttriglycerides levels
microb	Microbiology report summary
imaging	Imaging report summary
concomitant meds	Concomitant Medication

Table 1. Patient Profile Dataset Features

Feature	Description
order_id	Medication Order ID
patient_id	Patient ID
route	Route of administration
medication	Medication name
formulation	Formulation of medication
dose	Dosage quantity
dose_unit	Dosage Unit of Measurement
freq	Frequency
label	Label ("safe" or "unsafe")
reason_fyi	Justfication for label

Table 2. Medication Order Dataset Features

Some examples of drug-related problems for the unsafe orders include *Overdose/ Underdose, Renal dose adjustment, Serious adverse reactions, Boxed warnings, Drug interactions, Pregnancy contraindication, Age-related contraindication, Allergy, Drug administration issues and Missing drug monograph information.*

To err on the side of caution, if an openFDA API call fails (i.e., no drug monograph information is available), the corresponding order is classified as **"**unsafe**"** due to insufficient information. This ensures that such orders are escalated to pharmacists for manual review. Pharmacist rationales for safe and unsafe labels were later used to construct a Chain-of-Thought dataset for fine-tuning; see Methodology)

2.2. Dataset Splitting

As the project explores various LLM methodologies, including supervised fine-tuning, it is necessary to split the dataset into training and testing sets. Splitting was performed based on the type of medication-related problem. For instance, if two medication orders were labeled "unsafe" due to inadequate renal dose adjustment in patients with renal impairment, one order was allocated to the training set and the other to the testing set to ensure balanced representation.

This stratified splitting approach ensures that both the training and testing datasets are representative of the range of medication-related issues observed, thereby enhancing the robustness and generalization of the model (Huo et al., 2023). Out of the 50 medication orders, 25 orders were assigned to the training set and 25 orders were assigned to the testing set.

2.3. Dataset Processing

2.3.1. Patient and Medication Order

The raw patient profiles and medication order information, initially stored in Excel files, were concatenated into standardized strings following a consistent template. This approach ensures uniform formatting across all 50 medication orders, providing structured and predictable input for the LLM. These formatted strings are included as part of the LLM context information during model input.

A sample of the processed patient and medication order string is shown in **Appendix A.**

2.3.2. Drug Monograph

The drug monograph information retrieved from openFDA displayed considerable variability in both the sections included and their content, reflecting the diverse characteristics of individual drug products. The longest monograph retrieved — for Tramadol Hydrochloride — contained 26,354 words, highlighting the significant potential for large, inefficient context lengths when using raw monographs directly. Furthermore, several sections, such as *clinical_pharmacology*, *mechanism_of_action*, and *pharmacodynamics*, primarily provide clinical trial or mechanistic details that are not directly relevant to a pharmacist's verification workflow. These sections are often lengthy and not critical for medication order safety assessment.

To optimize context window usage (critical for LLM efficiency) and focus the model's attention on clinically actionable information, irrelevant sections were removed. Only the following sections were retained: *boxed_warnings*, *warnings*, *contraindications*, *do_not_use*, *dosage_and_administration*, *pregnancy*.

These selections were validated by a clinical pharmacist to ensure that all necessary information for safe medication classification is captured. A sample of the processed drug monograph is shown in **Appendix B**.

3. Methodology

This project explored multiple approaches to determine the most effective strategy for classifying medication orders as 'safe' or 'unsafe.' The approaches included (1) Retrieval-Augmented Generation (RAG), (2) direct context few-shot feeding ("full context prompting"), and (3) supervised fine-tuning of Large Language Models (LLMs) using the Unsloth framework.

3.1. Retrieval Augmented Generation

The Retrieval-Augmented Generation (RAG) framework was adopted to leverage on the strength of both information retrieval into the reasoning process of a large language model (LLM) (Lewis et al., 2020). In the context of medication order safety verification, RAG simulates the typical pharmacist workflow: reviewing the patient profile and order details, retrieving relevant drug information and finally determining the clinical safety of the prescribed medication.

3.1.1. Prompt Design

Several methodological decisions were made to further enhance the performance and reliability of the RAG method. First, a targeted retrieval strategy was implemented ensuring that only the drug monograph segments directly relevant to the medication order were retrieved. Additional strict filtering measures were implemented to prevent irrelevant drug information from being included in the prompt, which could otherwise compromise the quality of the LLM's response. Second, context-aware retrieval was incorporated, whereby patient-specific factors, such as the presence of renal impairment or pregnancy status were considered during the prompt construction to better highlight clinically significant risks during evaluation (Brown et al., 2020). Thrid, a standard-

ized prompt structuring method was adopted across all cases. Each prompt consistently followed a defined sequence:

- **Patient Clinical Context:** Structured information extracted from patient profile, including demographic, diagnoses, laboratory results vital signs and concurrent medications.
- **Medication Order Details:** The specific medication(s) ordered for that patient, including dose, route and frequency
- **Relevant Drug Information:** Retrieved by RAG retriever based on the identified drug from FAISS vector store, where each drug monograph has been preprocessed to include only clinically actionable actions. The retrieving process leverages FAISS for effcient and effective multi-resolution recall (MRR) search.
- **Question:** "Is this medication order safe for the given patient profile? Provide reasons to support your answer."

The prompt structuring reduced interpretation variability that ensures LLM is able to reason systematically. Refer to Appendix C for sample template.

3.1.2. RAG Fine-Tuning

Natrually, the quality of RAG approach response quality is heavily influenced by the quality and accuracy of the retrieved drug information. To optimize information retrieval and ensure clinical relevance, the original drug monograph were first preprocessed to extract only key actional sections, such as boxed warnings, contradictions, pregnancy-related warnings and dosage guidelines. This reduced the risk of introducing unnecessary noise into retrieval process.

Beyond preprocessing, a target maximum character length of 1000 per chunk was adopted. This size was selected based on the need to preserve coherent clinical information while avoiding excessive context length that could dilute retrieval specificity. During the retrieval, a top k strategy was adopted to select the most relevant chunks for each query. The value of k is set as 5 to balance two competing objectives: (i) ensuring sufficient coverage of potential safety issues across different sections of the monograph, and (ii) maintaining a compact prompt size to prevent overwhelming the LLM on its response generation(OpenAI, 2023). Finally, the standardized prompt template was input into three LLMs (Qwen-0.5B, GPT-3.5, and GPT-4o) for evaluation.

3.2. Direct context few-shot prompting

To further improve on the LLM output, we hypothesize that enhancements can be made via 3 main changes. Firstly, instead of using Retrieval-Augmented Generation (RAG) to selectively retrieve chunks of drug monograph data, we provided the entire monograph directly in the context. This would help to ensure that all pertinent monograph information has been fed into the context and helps prevent omission of important during retrieval as imperfect chunking can result in information leakage. Secondly, a larger model "suayptalha/DeepSeek-R1-Distill-Llama-3B" was implemented as DeepSeek has slightly better reasoning capabilities and using a model with size 3B would also improve the output quality having better contextual understanding and generation fluency. However, due to resource constraint, this is the largest model that could be implemented. Thirdly, we provided 4-shot prompting in the context, giving 2 examples of safe and unsafe prescriptions each. This would help the LLM train on the type of output we expect, learn from the reasoning patterns and tones, allowing it to generate a more short and concise answer that is contextually grounded and aligned with our expectations.

3.3. Supervised Finetuning using Unsloth

We have also explored a supervised finetuning approach for this project, using the Unsloth framework to adapt the DeepSeek-R1-Distill-Llama-8B model for the medication safety verification task.

Unsloth is a Python-based framework optimized for efficient Low-Rank Adaptation (LoRA) fine-tuning. It accelerates the fine-tuning process by manually deriving matrix differentials and chaining matrix multiplications, significantly reducing computational overhead. Built atop the Hugging Face Transformers library, Unsloth combines the robustness of Transformers with enhanced optimization for speed and memory efficiency (Han, 2023).

The base model selected was DeepSeek-R1-Distill-Llama-8B, known for its strong chain-of-thought (CoT) reasoning capabilities—a crucial feature for clinical safety tasks. DeepSeek-R1 builds upon DeepSeek-R1-Zero (the first open-source model trained purely via large-scale reinforcement learning) by introducing cold-start supervised pretraining prior to reinforcement learning, improving clarity, readability, and performance on math, coding, and reasoning benchmarks (HuggingFace, 2025; Xu, 2025).

3.3.1. Fine-Tuning Configuration

To accommodate GPU memory constraints, the model was loaded at 4-bit precision. Fine-tuning was further optimized through Low-Rank Adaptation (LoRA) applied to key attention weight matrices (e.g., query and value projections) with a rank of 16. This selective tuning approach enabled adaptation of the model's reasoning pathways without exceeding a 12 GB VRAM limit.

3.3.2. FINE-TUNING DATASET

The SFT training dataset consisted of three columns:

- **Question**: A concatenation of patient profile information, medication order information, and relevant drug monograph sections for each of the 25 training set orders.
- **Complex_CoT**: Pharmacist-authored, step-by-step Chain-of-Thought (CoT) rationales explaining the "safe" or "unsafe" classification.
- **Response**: The expected final structured output from the model.

The dataset was embedded into prompts using a structured template, explicitly instructing the model to generate intermediate reasoning steps before providing a final classification. There is zero-shot prompting for this approach. During inference, only the model's final classification was extracted to simulate a production workflow.

Fine-tuning the model on our custom dataset was crucial, as the specific structure of the pharmacist's reasoning and response style is highly specialized to the medication verification process, differing significantly from generic LLM behavior (see Appendix D for an example of Pharmacist's CoT). Hyperparameter tuning was also employed, with temperatures and number of epochs adjusted.

Unsloth's optimizations enabled fine-tuning of the 8B-parameter model on a single GPU at approximately 2× faster training speeds compared to vanilla Transformers (Unsloth, 2025). By coercing the model to generate CoT rationales, we maintained transparency in the decision-making process. Unlike generic RAG or full context prompting, supervised fine-tuning allowed the model to internalize the reasoning structure of pharmacists, thereby reducing reliance on large context injections and improving scalability. Given the small training dataset (25 orders), this fine-tuning was exploratory in nature, aimed primarily at assessing the feasibility of adapting reasoning models to the medication safety verification task. Future work should focus on scaling the training corpus to enhance generalization and performance.

4. Evaluation Metric Selection

4.1. Prediction Quality

We defined unsafe medication orders as the positive class. Model predictions (i.e., *safe* or *unsafe*) were compared against pharmacist-labeled ground truth to compute:

- **True Positives (TP)** – Unsafe orders correctly classified as unsafe
- **False Negatives (FN)** – Unsafe orders incorrectly classified as safe
- **False Positives (FP)** – Safe orders incorrectly classified as unsafe
- **True Negatives (TN)** – Safe orders correctly classified as safe

Given the critical importance of medication safety and the potential for patient harm, minimizing **False Negatives** was prioritized. Therefore, **Recall** (also known as sensitivity) for unsafe orders was selected as the primary evaluation metric. This ensures the model effectively captures unsafe prescriptions, thereby reducing the risk of medication errors.

While **False Positives**—unnecessarily flagging safe orders—pose minimal direct clinical risk due to subsequent pharmacist review, reducing them is desirable to enhance workflow efficiency. Consequently, **Specificity** (True Negative Rate) was chosen as the secondary evaluation metric, measuring the model's ability to correctly identify truly safe orders and avoid unnecessary pharmacist interventions.

To balance these considerations, a custom weighted score was developed:

$$\text{Weighted Score} = \alpha \times \text{Recall} + (1-\alpha) \times \text{Specificity} \quad (1)$$

Given the higher priority of Recall, α was set to 0.7. This weighted scoring approach provides a holistic and clinically grounded evaluation, translating technical model performance into a single metric for both business and clinical stakeholders clarity.

4.2. Justification Quality

Beyond prediction accuracy, the quality of the model's reasoning was evaluated to ensure interpretability and clinical trustworthiness. Two complementary evaluation methods were employed:

- **Embedding-Based Similarity:** Ground-truth rationales authored by pharmacists and model-generated rationales were embedded using the `all-mpnet-base-v2` Sentence Transformer model. Cosine similarity was calculated between each pair to measure semantic alignment. This automated, quantitative assessment gauges how closely the model's explanations match expert reasoning.
- **Human Expert Scoring:**
  A qualitative evaluation was conducted by a licensed pharmacist to assess the interpretability of model-generated justifications. Each explanation was reviewed and scored on a 1–4 scale across seven criteria: logical coherence, correctness of reasoning, con-

ciseness, clarity, relevance, duplication, and hallucination. These dimensions were carefully chosen to capture medical validity, clarity, and clinical utility of the justifications beyond what can be measured by embedding-based metric like cosine similarity. Given the labor-intensive nature of this human evaluation process, it was selectively applied to the outputs from the highest-performing models, as determined by their cosine similarity scores. By incorporating expert review in this targeted manner, the evaluation process remains practical while still offering a nuanced lens into the quality and safety of model reasoning in high-stakes scenarios. Most evaluation dimensions are rated on a scale from 1 to 4, where 4 represents the highest quality and 1 the lowest. However, for logical coherence and hallucination, a binary scoring system is applied: a score of 4 indicates acceptable output, while a score of 1 denotes an unacceptable response. To enhance interpretability, final scores were normalized on a scale from 0 to 1. Average score of the 7 metrics was calculated as the final score for human evaluation. Table 3 lists the detailed explanation of the 7 metrics for human evaluation. The guideline framework is shown in **Appendix E.**

Metrics	Description
Logical Coherence	Does the reasoning logically lead to the final decision or recommendation stated, regardless of medical correctness?
Correctness	Do the statements in the generated reasoning factually align with the medical guideline?
Conciseness	Linguistic efficiency and brevity - Is the reasoning clear and concise without unnecessary or filler content?
Clarity	Is the reasoning process easy for a human to follow and understand?
Relevance	Is the reasoning relevant to the true reasoning thought process?
Duplication	Is there any repeated statement in the reasoning?
Hallucinations	Does the reasoning invent unsupported facts or claims not present in the medical evidence?

Table 3. Human Evaluation Metrics

5. Discussion of Result

This project evaluated various methods for classifying medication orders as "safe" or "unsafe," with a focus on assessing prediction quality and justification quality.

5.1. Prediction Quality

The prediction quality of each model was evaluated using recall, specificity, and a weighted recall-specificity score.

Model	Recall / Specificity
Base Model	0.2308 / 0.5833
Base Model with Monograph	0.4615 / 0.8333
Fine-Tuned (temp = 1)	0.6923 / 0.6667
Fine-Tuned (temp = 0.5)	0.6923 / 0.7500
Qwen0.5b_RAG	0.0000 / 1.0000
GPT3.5_RAG	0.7692 / 0.5000
GPT4o_RAG	0.6154 / 0.5833
Few-shot Prompt	0.9231 / 0.1667

Table 4. Recall/Specificity evaluations. (i) Base Model: DeepSeek-R1-Distill-Llama-8B without monograph information; (ii) Base Model with Monograph: Base Model with drug monograph information in context; (iii) Fine-Tuned (temp = 1): Base Model with Monograph fine-tuned with SFT using Unsloth at temperature 1; (iv) Fine-Tuned (temp = 0.5): Fine-tuned model at temperature 0.5; (v) Qwen0.5b with RAG; (vi) GPT-3.5 with RAG; (vii) GPT-4o with RAG; (viii) Direct Context Few-Shot Prompting.

Model	Weighted Score
Base Model	0.3365
Base w Monograph	0.5731
Fine-Tuned (temp = 1)	0.6846
Fine-Tuned (temp = 0.5)	0.7096
Qwen0.5b_RAG	0.3000
GPT3.5_RAG	0.6885
GPT4o_RAG	0.6058
Few-Shot Prompt	0.6962

Table 5. Weighted Score Evaluations. (i) Base Model: DeepSeek-R1-Distill-Llama-8B without monograph information; (ii) Base Model w Monograph: Base Model with drug monograph information in context; (iii) Fine-Tuned (temp = 1): Base Model w Monograph with SFT using Unsloth, temperature = 1; (iv) Fine-Tuned (temp = 0.5): Base Model w Monograph with SFT using Unsloth, temperature = 0.5; (v) Qwen0.5b with RAG; (vi) GPT-3.5 with RAG; (vii) GPT-4o with RAG; (viii) Direct Context Few-Shot Prompting.

From Table 4, we observed that the base model, both with and without monograph information, performed poorly in terms of recall. The purpose of including the base model without monograph information was to assess DeepSeek-R1-Distill-Llama-8B's inherent ability to verify medication orders without external drug references. This configuration achieved a recall of 0.2308. Incorporating monograph information increased recall to 0.4615, indicating that structured drug knowledge helps the model better identify unsafe prescriptions. The improvement in both recall and the weighted

score (from 0.3365 to 0.5731, as shown in Table 5) suggests that access to clinical context enhances the model's decision-making.

Fine-tuning the model using pharmacist-authored Chain-of-Thought (CoT) rationales further improved performance. The best results were observed at temperature 0.5, with a recall of 0.6923 and specificity of 0.7500, leading to a weighted score of 0.7096. Lowering the temperature reduced randomness, producing more deterministic outputs—crucial for a safety-critical task. In contrast, temperature 1.0 yielded more variable outputs and a slightly lower weighted score (0.6846). These findings emphasize the importance of both fine-tuning and careful hyperparameter tuning to strike a balance between identifying unsafe cases (recall) and minimizing false alarms (specificity).

We also explored a few-shot prompting strategy using full-context prompts to leverage the model's reasoning ability without additional training. This approach achieved the highest recall across all models (0.9231), demonstrating strong sensitivity in identifying unsafe prescriptions. However, its specificity was the lowest (0.1667), reflecting a tendency to over-predict unsafe cases. Despite this, the method attained a competitive weighted score of 0.6962, as shown in Table 4, positioning it close to the top-performing fine-tuned configuration.

Lastly, RAG-based models such as Qwen0.5b_RAG, GPT3.5_RAG, and GPT4o_RAG showed mixed results. Qwen0.5b_RAG had perfect specificity but zero recall, making it unsuitable for detecting unsafe orders. GPT3.5_RAG and GPT4o_RAG offered better balance, but their specificity still lagged behind fine-tuned configurations. Table 4 helps to clearly rank these models based on the weighted score, highlighting the trade-offs between high recall and reliable specificity across modeling strategies.

5.2. Justification Quality

Justification quality was assessed using cosine similarity between model-generated rationales and ground truth explanations written by clinical experts. A higher cosine similarity score indicates greater semantic alignment and, consequently, better justification quality. Table 6 summarizes the average cosine similarity achieved by each model configuration.

5.2.1. Embedding-Based Similarity

The base model achieved a cosine similarity of 0.7199, and the inclusion of drug monograph information led to a noticeable improvement to 0.7639. This suggests that structured clinical references help the model generate more accurate and relevant explanations, aligning more closely with expert reasoning.

Model	Cosine Similarity
Base Model	0.7199
Base w Monograph	0.7639
Fine-Tuned (temp = 1)	0.7713
Fine-Tuned (temp = 0.5)	0.7319
Qwen0.5b_RAG	0.5957
GPT3.5_RAG	0.7376
GPT4o_RAG	0.7337
Few-Shot Prompt	0.6575

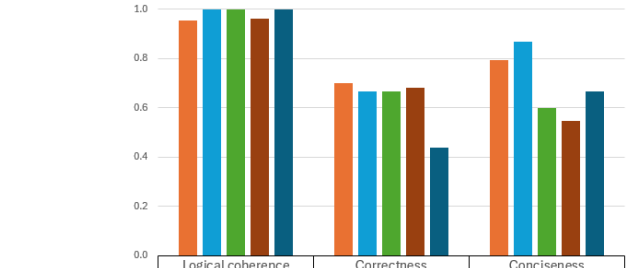
Table 6. Cosine Similarity Evaluations. (i) Base Model: DeepSeek-R1-Distill-Llama-8B without monograph information; (ii) Base Model w Monograph: Base Model with drug monograph information in context; (iii) Fine-Tuned (temp = 1): Base Model w Monograph with SFT using Unsloth, temperature = 1; (iv) Fine-Tuned (temp = 0.5): Base Model w Monograph with SFT using Unsloth, temperature = 0.5; (v) Qwen0.5b with RAG; (vi) GPT-3.5 with RAG; (vii) GPT-4o with RAG; (viii) Direct Context Few-Shot Prompting.

Fine-tuning the model further improved performance. The fine-tuned model with a temperature of 1 produced the highest similarity score (0.7713), indicating that this configuration generated the most expert-aligned and coherent justifications. In contrast, lowering the temperature to 0.5 slightly reduced justification quality (0.7319), possibly due to decreased generation flexibility. While deterministic outputs are beneficial for consistency, they may also suppress the nuanced reasoning required for high-quality clinical explanations.

The RAG-based models showed varied performance. GPT-3.5_RAG and GPT-4o_RAG achieved moderate cosine similarities of 0.7376 and 0.7337, respectively, while Qwen0.5b_RAG lagged significantly behind at 0.5957. These results suggest that although RAG approaches can incorporate external knowledge effectively, they may not consistently produce justifications that reflect expert-level reasoning unless well-aligned with the downstream task.

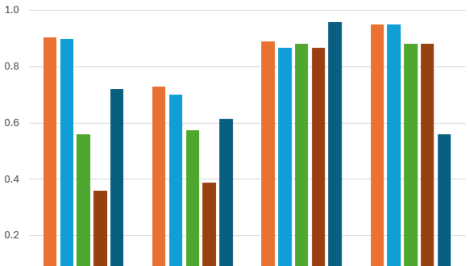
The few-shot prompting method, which provides full context without fine-tuning, achieved a cosine similarity of 0.6575. While this outperformed Qwen0.5b_RAG, it remained lower than most other methods, reflecting challenges in guiding the model toward structured, expert-like reasoning solely through prompting. Nevertheless, this result shows that prompting alone can produce moderately aligned justifications when carefully constructed, even without additional training.

Overall, these findings underscore the value of both content enrichment and targeted fine-tuning in improving the alignment of generated justifications with expert clinical reasoning.



	Logical coherence	Correctness	Conciseness
Fine-Tuned (temp=1)	0.953	0.700	0.793
Fine-Tuned (temp=0.5)	1.000	0.667	0.867
GPT3.5_RAG	1.000	0.667	0.600
GPT4o_RAG	0.960	0.680	0.547
Few-shot prompting	1.000	0.440	0.667

Figure 1. Average Human Scoring for Logical Coherence, Correctness and Conciseness per model. (i) Fine-Tuned (temp = 1): DeepSeek-R1-Distill-Llama-8B with SFT using Unsloth, temperature = 1 (ii) Fine-Tuned (temp = 0.5): DeepSeek-R1-Distill-Llama-8B with SFT using Unsloth, temperature = 0.5 (iii) GPT3.5 with RAG (iv) GPT4o with RAG (v) Direct Context Few-Shot Prompting



	Clarity	Relevance	Duplication	Hallucination
Fine-Tuned (temp=1)	0.903	0.730	0.890	0.950
Fine-Tuned (temp=0.5)	0.900	0.700	0.867	0.950
GPT3.5_RAG	0.560	0.573	0.880	0.880
GPT4o_RAG	0.360	0.387	0.867	0.880
Few-shot prompting	0.720	0.613	0.960	0.560

Figure 2. Average Human Scoring for Clarity, Relevance, Duplication and Hallucination per model. (i) Fine-Tuned (temp = 1): DeepSeek-R1-Distill-Llama-8B with SFT using Unsloth, temperature = 1 (ii) Fine-Tuned (temp = 0.5): DeepSeek-R1-Distill-Llama-8B with SFT using Unsloth, temperature = 0.5 (iii) GPT3.5 with RAG (iv) GPT4o with RAG (v) Direct Context Few-Shot Prompting

5.2.2. Human Evaluation

Following human evaluation of the test cases, the average scores across seven evaluation metrics were computed for each model.

As shown in Figure 1 and Figure 2, the Fine-Tuned (temp=0.5) model stands out overall, achieving the highest average score of 0.850, driven by perfect performance in Logical Coherence and consistently strong results in Conciseness, Clarity, Duplication, and Hallucination. This outcome is aligned with expectations, as lower temperature settings typically result in more deterministic and focused outputs.

Between the base models, GPT-3.5 consistently outperforms GPT-4o in most categories. Notably, GPT-4o demonstrated a tendency toward verbosity, generating more nuanced but less concise responses. Its lowest score was in Clarity (0.360), where the model frequently included multiple aspects within its reasoning, not all of which were directly relevant. If deploying GPT-4o in similar contexts, it would be advisable to consider reducing the temperature setting or applying stricter output constraints to improve conciseness and clarity.

When comparing the two fine-tuned models, Fine-Tuned (temp=1) performed well, with Logical Coherence (0.952) and Hallucination Control (0.950) on par with the Fine-Tuned (temp=0.5) model. However, it trailed slightly in Clarity (0.900) and Conciseness (0.794 versus 0.867). These results suggest that lowering the temperature enhanced generation consistency without compromising correctness or hallucination control.

Few-shot prompting model outperformed the rest on Duplication(0.960), while it perform worst in terms of Hallucination (0.560) and Correctness (0.440).

Across all models, Logical Coherence scores were consistently high, indicating that this dimension is less sensitive to temperature adjustments or model type in this evaluation setup. Correctness scores were moderately consistent (0.67–0.70), suggesting that while models effectively avoid hallucinations, minor factual inaccuracies persist. Both Duplication and Hallucination Control demonstrated stable performance across models, reflecting effective prompt engineering.

However, an important limitation was identified in both Fine-Tuned (temp=1) and Fine-Tuned (temp=0.5) models. In certain cases, these models provided a final decision and recommended action without accompanying reasoning. This indicates a potential shortcoming in the models' ability to transparently justify their outputs, which may undermine interpretability and trustworthiness in high-stakes or decision-critical applications. To address this, it is recommended to revise both the model prompts and training data to explicitly enforce structured reasoning prior to presenting final decisions. This would ensure that each recommendation is consistently supported by clear, stepwise justification.

Additional observations and improvement areas from human evaluation:

- **Hallucination**: In Order 45, Fine-Tuned (temp=1) incorrectly assumed a male patient was pregnant and provided reasoning based on this error.

- **Incomplete Reasoning**: In certain instances, Fine-Tuned (temp=1) generalized conclusions about safety risks without explicitly detailing the underlying rationale, opting instead to list medication usage and patient conditions.
- **Incoherence**: For Order 34, a logical inconsistency was observed where an issue flagged in the reasoning was not reflected in the final decision.
- **Under-dosing Detection**: A recurring weakness across all four models was the failure to identify under-dosing cases, highlighting a priority area for future iterations.
- **Ambiguous Justification**: In Order 26, Fine-Tuned (temp=0.5) presented non-committal reasoning, citing either insufficient information or inappropriate dosing without clarifying which applied, thus offering little actionable insight.

Overall, while the fine-tuned models demonstrate substantial improvements over base models, particularly at lower temperature settings, targeted refinements in prompt structure, reasoning requirements, and specific clinical safety checks (e.g., under-dosing detection) will be critical to achieving reliable, interpretable decision support.

5.3. Time taken for inference

The time taken for inference for each medication order in the testing dataset is recorded across all 3 approaches (RAG, Few-shot prompting, SFT). The average time taken is shown in Table 7.

Model	Avg inference time per order (seconds)
Fine-Tuned (temp = 1)	79.29
Fine-Tuned (temp = 0.5)	43.74
Qwen0.5b_RAG	7.76
GPT3.5_RAG	5.30
GPT4o_RAG	10.33
Few-Shot Prompt	12.60

Table 7. Average inference time per medication order per model. (i) Fine-Tuned (temp = 1): DeepSeek-R1-Distill-Llama-8B with Monograph and SFT using Unsloth, temperature = 1; (ii) Fine-Tuned (temp = 0.5): DeepSeek-R1-Distill-Llama-8B with Monograph and SFT using Unsloth, temperature = 0.5; (iii) Qwen0.5b with RAG; (iv) GPT-3.5 with RAG; (v) GPT-4o with RAG; (vi) Direct Context Few-Shot Prompting.

A higher temperature setting (e.g., 1.0) encourages more diverse output sampling, which increases token generation time and often results in more verbose responses. Consequently, the average inference time for the Fine-Tuned model at temperature 1 exceeded one minute per medication order, rendering the approach impractical for real-time clinical use. Lowering the temperature allowed the model to produce more focused and concise outputs, thereby improving response time. However, despite this improvement, inference remained slower than retrieval-based or API-hosted methods due to the large size of the fine-tuned model—DeepSeek-R1-Distill-Llama-8B—and the computational overhead of running it locally, even at 4-bit precision. In contrast, Retrieval-Augmented Generation (RAG) approaches using hosted models like GPT-3.5 and GPT-4o demonstrated the fastest inference times. This performance advantage stemmed from more efficient prompt construction, shorter input lengths enabled by targeted retrieval, and the use of highly optimized infrastructure maintained by OpenAI.

Interestingly, direct few-shot prompting with commercial GPT models also outperformed locally fine-tuned inference in terms of speed. These API-hosted models benefit from hardware acceleration, dynamic batching, and other system-level optimizations that reduce latency. Although they do not offer the same level of control or customization as supervised fine-tuning, they provided quick and consistent responses suitable for prototyping and low-latency deployment. Model size also clearly influenced inference performance: smaller models like Qwen0.5b yielded faster results, albeit with a potential tradeoff in reasoning depth. Ultimately, the evaluation revealed a fundamental tradeoff between interpretability, responsiveness, and scalability—highlighting the importance of aligning model architecture and deployment strategy with clinical and operational requirements.

5.4. Finetuning or RAG

From the overall performance of the evaluated models, we can postulate several reasons why fine-tuning outperforms RAG in this task.

RAG models fundamentally depend on retrieval quality. In RAG, the model first retrieves external documents (e.g., sections of the drug monograph) and then generates a response based on them. If the retrieved chunks are not perfectly relevant or precisely matched to the patient's clinical context, the model tends to generate weaker or more generic explanations. Retrieval errors or omissions in fine-grained clinical details can severely limit reasoning depth, even when the base model is strong.

In contrast, fine-tuned models are explicitly trained on examples demonstrating how to reason about patient-medication safety, using pharmacist-style clinical justifications. This training not only imparts factual knowledge but also teaches the structure, prioritization, and language style expected in clinical settings. Fine-tuning thus enables the model to

internalize both what to say and how to say it — something RAG alone cannot achieve unless retrieval and generation are perfectly optimized together, which is challenging in practice.

Moreover, fine-tuned models integrate patient information and drug properties into a cohesive reasoning chain, allowing them to generate more organically adapted and case-specific outputs. By comparison, RAG-based reasoning often feels "stitched together" — retrieved chunks are treated more independently, resulting in less fluid, less patient-tailored justifications.

Finally, the model size matters. Smaller models like Qwen-0.5B have limited reasoning capacity compared to larger fine-tuned models such as the 8B DeepSeek-R1. Even when relevant documents are retrieved, smaller models may lack the capacity to deeply synthesize the facts into strong, clinically coherent justifications.

Overall, these factors explain why fine-tuning, particularly when aligned with clinical Chain-of-Thought supervision, edges out RAG for medication safety verification tasks in our experiments.

6. Limitations and Future Work

6.1. Small Dataset Size

One key limitation of this project is the small dataset size used for both model training and evaluation. The set of 50 medication orders, while curated to cover diverse safety issues, is insufficient to capture the full complexity and variability of real-world prescribing scenarios. This constrains the generalizability of the findings and limits the statistical significance of performance metrics. Future work could focus on expanding the dataset through collaboration with healthcare institutions or leveraging anonymized prescription datasets to enable more robust fine-tuning and evaluation.

6.2. Limited Expert Justifications

Another limitation lies in the use of single-expert ground truth justifications for assessing explanation quality. While these justifications provide a clinically sound benchmark, they reflect the reasoning of one individual and may not account for the range of valid perspectives that exist in clinical practice. Additionally, the cosine similarity metric, though useful for measuring semantic alignment, may not capture more nuanced aspects such as clinical correctness, contextual clarity, or hallucinated content. To address this, future work could involve multi-expert annotations and a more comprehensive human evaluation rubric to better reflect the interpretability and trustworthiness of model outputs.

6.3. Potential for Human-in-the-Loop Feedback Reinforcement Learning

Incorporating human-in-the-loop feedback offers a promising pathway to enhance model performance. By involving medical professionals to systematically review and annotate outputs — flagging deficiencies in justification, reasoning clarity, or safety considerations — the fine-tuning process can be iteratively refined. This approach would align model outputs more closely with clinical standards, improving reliability and trust. However, the implementation of such feedback-driven reinforcement learning is resource-intensive and may pose significant operational challenges.

7. Conclusion

In conclusion, while supervised fine-tuning showed the most promising performance in reasoning accuracy, its scalability is limited. The model was trained on only 25 examples, raising concerns about overfitting and generalizability. Additionally, generating Chain-of-Thought rationales is labor-intensive, slowing data expansion. The current solution is not ready for clinical deployment, where even a single false negative—i.e., an unsafe order misclassified as safe—could result in significant patient harm. Future work should focus on scaling data collection, enhancing the robustness of supervised models, and further optimizing the Retrieval-Augmented Generation (RAG) pipeline to balance performance, safety, and practicality for clinical adoption.

References

Babu, D., Rowett, D., Lim, R., Marotti, S., Wisdom, A., and Kalisch Ellett, L. Clinical pharmacists' participation in ward rounds in hospitals: responses from a national survey. *International Journal of Pharmacy Practice*, 31(4):409–416, 2023. doi: 10.1093/ijpp/riac112. URL https://academic.oup.com/ijpp/article/31/4/409/7161614.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

DataCamp. Fine-tuning deepseek r1 (reasoning model), 2025. URL https://www.datacamp.com/tutorial/fine-tuning-deepseek-r1-reasoning-model. Accessed: 2025-04-18.

Food, U. and Administration, D. openfda drug labeling api, 2025. URL https://open.fda.gov/apis/drug/label/. Accessed: 2025-04-18.

Han, D. Introducing unsloth: 30x faster llm training, 2023. URL https://unsloth.ai/introducing. Accessed: 2025-04-18.

HuggingFace. Deepseek-r1 distill model, 2025. URL https://huggingface.co/unsloth/DeepSeek-R1-Distill-Llama-8B. Accessed: 2025-04-18.

Huo, T., Glueck, D. H., Shenkman, E. A., and Muller, K. E. Stratified split sampling of electronic health records. *BMC Medical Research Methodology*, 23(1):128, 2023. doi: 10.1186/s12874-023-01938-0. URL https://doi.org/10.1186/s12874-023-01938-0.

Kokosi, T. and Harron, K. Synthetic data in medical research. *BMJ Medicine*, 1(1):e000167, 2022. doi: 10.1136/bmjmed-2022-000167. URL https://doi.org/10.1136/bmjmed-2022-000167.

Lewis, P., Perez, E., Piktus, A., Karpukhin, V., Guu, K., Chen, W.-t., Oguz, B., Min, S., Yih, W.-t., and Riedel, S. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

OpenAI. Best practices for prompt engineering with openai api. https://platform.openai.com/docs/guides/prompt-engineering, 2023.

Rodziewicz, T. L., Houseman, B., Vaqar, S., and Hipskind, J. E. Medical error reduction and prevention. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2024. URL https://pubmed.ncbi.nlm.nih.gov/29763131/. PMID: 29763131.

Studer, H., Imfeld-Isenegger, T., Beeler, P., et al. The impact of pharmacist-led medication reconciliation and interprofessional ward rounds on drug-related problems at hospital discharge. *International Journal of Clinical Pharmacy*, 45:117–125, 2023. doi: 10.1007/s11096-022-01496-3. URL https://link.springer.com/article/10.1007/s11096-022-01496-3.

Tariq, R. A., Vashisht, R., Sinha, A., et al. Medication dispensing errors and prevention. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2024. URL https://www.ncbi.nlm.nih.gov/books/NBK519065/. Updated 2024 Feb 12, cited 2025 Apr 18.

Unsloth. Fine-tuning guide, 2025. URL https://docs.unsloth.ai/get-started/fine-tuning-guide. Accessed: 2025-04-18.

World Health Organization. Patient safety, 2024. URL https://www.who.int/news-room/fact-sheets/detail/patient-safety. Accessed: 2025-04-18.

Xu, S. The complete guide to deepseek models: From v3 to r1 and beyond, 2025. URL bentoml.com DeepSeek Guide. Accessed: 2025-04-18.

Appendix

A. Sample processed Patient Information and Medication Order

Patient is a 71-year-old female, height 1.56 m and weight 51 kg. Latest heart rate is 92 bpm. Latest systolic blood pressure is 113. Patient has a past medical history of Parkinson's disease, Type 2 Diabetes Mellitus, and Stable Ischemic Heart Disease. Currently, the patient is admitted for Vertigo and Gastroesophageal Reflux Disease.

Patient is also taking the following medications: oral Glipizide 15 mg twice daily before meals, oral Clopidogrel 75 mg every morning, oral Benserazide 25 mg / Levodopa 100 mg twice a day.

Patient has known allergy to: *none*.

Some recent lab results and reports are shown below:

White Blood Cell = 2.32 $\times 10^9$/L

Haemoglobin = 11.3 g/dL

Platelet = 56.0 g/dL

Creatinine Clearance = 61.0 mL/min

eGFR = 55.0 mL/min/1.73 m^2

Sodium = 141 mmol/L

Potassium = 4.0 mmol/L

Magnesium = 0.8 mmol/L

Fasting Glucose = 7.1 mmol/L

HbA1c = 6.5%

HDL = 0.5 mmol/L

LDL = 2.0 mmol/L

TG = 2.0 mmol/L.

Microbiology Report: *none*
Imaging Report (Chest X-ray): No significant lung consolidation, pleural effusion, or pneumothorax. Cardiomediastinal contour unremarkable.

The doctor ordered oral Metoclopramide 10 mg three times a day.

B. Sample processed Drug Monograph Information

Drug Monography Information as below:

WARNING: TARDIVE DYSKINESIA Treatment with metoclopramide can cause tardive dyskinesia, a serious movement disorder that is often irreversible. The risk of developing tardive dyskinesia increases with duration of treatment and total cumulative dose. Metoclopramide therapy should be discontinued in patients who develop signs or symptoms of tardive dyskinesia. There is no known treatment for tardive dyskinesia. In some patients, symptoms may lessen or resolve after metoclopramide treatment is stopped. Treatment with metoclopramide for longer than 12 weeks should be avoided in all but rare cases where therapeutic benefit is thought to outweigh the risk of developing tardive dyskinesia. See WARNINGS DOSAGE AND ADMINISTRATION Therapy with metoclopramide oral solution should not exceed 12 weeks in duration. For the Relief of Symptomatic Gastroesophageal Reflux Administer from 10 mg to 15 mg metoclopramide orally up to 4 times daily 30 minutes before each meal and at bedtime, depending upon symptoms being treated and clinical response (see CLINICAL PHARMACOLOGY and INDICATIONS AND USAGE). If symptoms occur only intermittently or at specific times of the day, use of metoclopramide in single doses up to 20 mg prior to the provoking situation may be preferred rather than continuous treatment. Occasionally, patients (such as elderly patients) who are more sensitive to the

therapeutic or adverse effects of metoclopramide will require only 5 mg per dose. Experience with esophageal erosions and ulcerations is limited, but healing has thus far been documented in one controlled trial using 4 times daily therapy at 15 mg/dose, and this regimen should be used when lesions are present, so long as it is tolerated (see ADVERSE REACTIONS). Because of the poor correlation between symptoms and endoscopic appearance of the esophagus, therapy directed at esophageal lesions is best guided by endoscopic evaluation. Therapy longer than 12 weeks has not been evaluated and cannot be recommended. For the Relief of Symptoms Associated with Diabetic Gastroparesis (Diabetic Gastric Stasis) Administer 10 mg of metoclopramide 30 minutes before each meal and at bedtime for two to eight weeks, depending upon response and the likelihood of continued well-being upon drug discontinuation. The initial route of administration should be determined by the severity of the presenting symptoms. If only the earliest manifestations of diabetic gastric stasis are present, oral administration of metoclopramide may be initiated. However, if severe symptoms are present, therapy should begin with metoclopramide injection (consult labeling of the injection prior to initiating parenteral administration). Administration of metoclopramide injection up to 10 days maybe required before symptoms subside at which time oral administration may be instituted. Since diabetic gastric stasis is frequently recurrent, metoclopramide therapy should be reinstituted at the earliest manifestation. Use in Patients with Renal or Hepatic Impairment Since metoclopramide is excreted principally through the kidneys, in those patients whose creatinine clearance is below 40 mL/min, therapy should be initiated at approximately one-half the recommended dosage. Depending upon clinical efficacy and safety considerations, the dosage may be increased or decreased as appropriate. See OVERDOSAGE section for information regarding dialysis. Metoclopramide undergoes minimal hepatic metabolism, except for simple conjugation. Its safe use has been described in patients with advanced liver disease whose renal function was normal. WARNINGS Mental depression has occurred in patients with and without prior history of depression. Symptoms have ranged from mild to severe and have included suicidal ideation and suicide. Metoclopramide should be given to patients with a prior history of depression only if the expected benefits outweigh the potential risks. Extrapyramidal symptoms, manifested primarily as acute dystonic reactions, occur in approximately 1 in 500 patients treated with the usual adult dosages of 30 to 40 mg/day of metoclopramide. These usually are seen during the first 24 to 48 hours of treatment with metoclopramide, occur more frequently in pediatric patients and adult patients less than 30 years of age and are even more frequent at the higher doses. These symptoms may include involuntary movements of limbs and facial grimacing, torticollis, oculogyric crisis, rhythmic protrusion of tongue, bulbar type of speech, trismus, or dystonic reactions resembling tetanus. Rarely, dystonic reactions may present as stridor and dyspnea, possibly due to laryngospasm. If these symptoms should occur, inject 50 mg diphenhydramine hydrochloride intramuscularly, and they usually will subside. Benztropine mesylate, 1 to 2 mg intramuscularly, may also be used to reverse these reactions. Parkinsonian-like symptoms have occurred, more commonly within the first 6 months after beginning treatment with metoclopramide, but occasionally after longer periods. These symptoms generally subside within 2 to 3 months following discontinuance of metoclopramide. Patients with preexisting Parkinsonâ€™s disease should be given metoclopramide cautiously, if at all, since such patients may experience exacerbation of parkinsonian symptoms when taking metoclopramide. Tardive Dyskinesia (see Boxed Warnings) Treatment with metoclopramide can cause tardive dyskinesia (TD), a potentially irreversible and disfiguring disorder characterized by involuntary movements of the face, tongue, or extremities. The risk of developing tardive dyskinesia increases with duration of treatment and the total cumulative dose. An analysis of utilization patterns showed that about 20% of patients who used metoclopramide took it for longer than 12 weeks. Treatment with metoclopramide for longer than the recommended 12 weeks should be avoided in all but rare cases where therapeutic benefit is thought to outweigh the risk of developing TD. Although the risk of developing TD in the general population may be increased among the elderly, women, and diabetics, it is not possible to predict which patients will develop metoclopramide-induced TD. Both the risk of developing TD and the likelihood that TD will become irreversible increase with duration of treatment and total cumulative dose. Metoclopramide should be discontinued in patients who develop signs or symptoms of TD. There is no known effective treatment for established cases of TD, although in some patients, TD may remit, partially or completely, within several weeks to months after metoclopramide is withdrawn. Metoclopramide itself may suppress, or partially suppress, the signs of TD, thereby masking the underlying disease process. The effect of this symptomatic suppression upon the long term course of TD is unknown. Therefore, metoclopramide should not be used for the symptomatic control of TD. Neuroleptic Malignant Syndrome (NMS) There have been rare reports of an uncommon but potentially fatal symptom complex sometimes referred to as Neuroleptic Malignant Syndrome (NMS) associated with metoclopramide. Clinical manifestations of NMS include hyperthermia, muscle rigidity, altered consciousness, and evidence of autonomic instability (irregular pulse or blood pressure, tachycardia, diaphoresis and cardiac arrhythmias). The diagnostic evaluation of patients with this syndrome is complicated. In arriving at a diagnosis, it is important to identify cases where the clinical presentation includes both serious medical illness (e.g., pneumonia, systemic infection, etc.) and untreated or inadequately treated extrapyramidal signs and symptoms (EPS). Other important considerations in the differential diagnosis include central anticholinergic toxicity, heat stroke, malignant hyperthermia, drug fever and primary central nervous system

(CNS) pathology. The management of NMS should include 1) immediate discontinuation of metoclopramide and other drugs not essential to concurrent therapy, 2) intensive symptomatic treatment and medical monitoring, and 3) treatment of any concomitant serious medical problems for which specific treatments are available. Bromocriptine and dantrolene sodium have been used in treatment of NMS, but their effectiveness have not been established (see ADVERSE REACTIONS).

C. Sample RAG Prompt Template

Patient Profile:
Patient is a 31-year-old female, height 1.68m and weighs 49kg. Latest heart rate is 87 bpm. Latest systolic blood pressure is 119 mmHg. Patient has a past medical history of *null*. Currently admitted for *Trigger Finger*.

Patient is also taking the following medications: *null*.
Patient has known allergy to: **tramadol**.

Recent Laboratory Results:

- White Blood Cell = 9.2 $\times 10^9$/L
- Haemoglobin = 12.5 g/dL
- Platelet = 56.0 g/dL
- Creatinine Clearance = 80.0 mL/min
- eGFR = 64.0 mL/min/1.73m^2
- Sodium = 140 mmol/L
- Potassium = 4.1 mmol/L
- Magnesium = 0.9 mmol/L

Microbiology Report: *null*
Imaging Report: *null*

Medication Order:
Oral tramadol hydrochloride 50mg two times a day.

Relevant Drug Information: (drug: TRAMADOL HYDROCHLORIDE_data) Tablets Once-Daily and 50 mg Tramadol IR Tablets Every 6 Hours. Food Effects After a single dose administration of 200 mg tramadol hydrochloride extended-release tablet with a high fat meal, the C_{max} and AUC of tramadol decreased 28% and 16%, respectively, compared to fasting conditions. Mean T max was increased by 3 hr (from 14 hr under fasting conditions to 17 hr under fed conditions). While tramadol hydrochloride extended-release tablets may be taken without regard to food, it is recommended that it be taken in a consistent manner [see Dosage and Administration (2.1)]. Distribution The volume of distribution of tramadol was 2.6 and 2.9 L/kg in male and female subjects, respectively, following a 100 mg intravenous dose. The binding of tramadol to human plasma proteins is approximately 20% and binding also appears to be independent of concentration up to 10 mcg/mL. Saturation of plasma protein binding occurs only at concentrations outside the clinically relevant range.

(drug: TRAMADOL HYDROCHLORIDE_data) If concomitant use is warranted, carefully observe the patient, particularly during treatment initiation and dose adjustment. Discontinue tramadol hydrochloride extended-release tablets if serotonin syndrome is suspected.Selective serotonin reuptake inhibitors (SSRIs), serotonin and norepinephrine reuptake inhibitors (SNRIs), tricyclic antidepressants (TCAs), triptans, 5-HT3 receptor antagonists, drugs that affect the serotonin neurotransmitter system (e.g., mirtazapine, trazodone, tramadol), certain muscle relaxants (i.e., cyclobenzaprine, metaxalone), monoamine oxidase (MAO) inhibitors (those intended to treat psychiatric disorders and also others, such as linezolid and intravenous methylene blue).

(drug: TRAMADOL HYDROCHLORIDE_data) A total of 3108 patients were studied during trials conducted in the U.S., including four double-blind studies in patients with osteoarthritis and/or chronic low back pain and one open-label study in patients with chronic non-malignant pain. A total of 901 patients were 65 years or older. The frequency of adverse reactions generally increased with doses from 100 mg to 400 mg in the two pooled, twelve-week, randomized, double-blind, placebo-controlled studies in patients with chronic non-malignant pain (see Table 1). The most common adverse reactions from Table 1 occurring in $\geq$ 10% and $\geq$ 2 times the placebo rate of patients treated with tramadol hydrochloride extended-release tablets were dizziness (not vertigo), nausea, constipation, headache, somnolence, flushing, pruritus, vomiting, insomnia, and dry mouth.

(drug: TRAMADOL HYDROCHLORIDE_data) PACKAGE LABEL.PRINCIPAL DISPLAY PANEL 300 mg label NDC 47335-861-83 (Once Daily) Tramadol Hydrochloride Extended-release Tablets, USP CIV 300 mg The tablets should be swallowed whole with liquid and not split, chewed, dissolved or crushed. PHARMACIST: Please dispense with Medication Guide provided separately to each patient. Rx only 30 Tablets SUN PHARMA tramadol-label-300mg

(drug: TRAMADOL HYDROCHLORIDE_data) addiction, abuse, and misuse. Consider prescribing naloxone for the emergency treatment of opioid overdose [see Dosage and Administration (2.2), Warnings and Precautions (5.2)]. Abuse or misuse of tramadol hydrochloride extended-release tablets by cutting, breaking, chewing, crushing, snorting, or injecting the dissolved product will result in the uncontrolled delivery of tramadol and can result in overdose and death [see Overdosage (10)]. Opioids are sought for nonmedical use and are subject to diversion from legitimate prescribed use. Consider these risks when prescribing or dispensing tramadol hydrochloride extended-release tablets. Strategies to reduce these risks include prescribing the drug in the smallest appropriate quantity and advising the patient on careful storage of the drug during the course of treatment and proper disposal of unused drug. Contact local state professional licensing board or state-controlled substances authority for information on how to prevent and

(drug: TRAMADOL HYDROCHLORIDE_data) and possibly signs and symptoms of opioid withdrawal in patients who had developed physical dependence to tramadol. Intervention: If concomitant use is necessary, consider dosage reduction of tramadol hydrochloride extended-release tablets until stable drug effects are achieved. Inform patients and caregivers of this potential interaction, educate them on the signs and symptoms of seizures, serotonin syndrome, and signs of respiratory depression and sedation. If a CYP3A4 inhibitor is discontinued, consider increasing the tramadol hydrochloride extended-release tablets dosage until stable drug effects are achieved and evaluate patients at frequent intervals for signs and symptoms of opioid withdrawal. Examples Macrolide antibiotics (e.g., erythromycin), azole-antifungal agents (e.g. ketoconazole), protease inhibitors (e.g., ritonavir) CYP3A4 Inducers Clinical Impact: The concomitant use of tramadol hydrochloride extended-release tablets and CYP3A4 inducers can decrease the plasma

(drug: TRAMADOL HYDROCHLORIDE_data) from Other Opioids to Tramadol Hydrochloride Extended-Release Tablets When tramadol hydrochloride extended-release tablets therapy is initiated, discontinue all other opioid analgesics other than those used on an as needed basis for breakthrough pain when appropriate. 2.4 Titration and Maintenance of Therapy Individually titrate tramadol hydrochloride extended-release tablets by 100 mg every five days to a dose that provides adequate analgesia and minimizes adverse reactions. The maximum daily dose of tramadol hydrochloride extended-release tablets is 300 mg per day. Continually reevaluate patients receiving tramadol hydrochloride extended-release tablets to assess the maintenance of pain control, signs and symptoms of opioid withdrawal, and other adverse reactions as well as to reassess for the development of addiction, abuse, or misuse [see Warnings and Precautions (5.1, 5.18)] . Frequent communication is important among the prescriber, other members of the healthcare team, the

(drug: TRAMADOL HYDROCHLORIDE_data) Pharmacokinetic Parameter Values (n=32) Tramadol M1 Metabolite Pharmacokinetic Parameter Tramadol hydrochloride extended-release 200 mg Tablet Tramadol hydrochloride 50 mg Tablet Tramadol hydrochloride extended-release 200 mg Tablet Tramadol hydrochloride 50 mg Tablet Once-Daily Every 6 Hours Once-Daily Every 6 Hours AUC 0-24 (ng·h/mL) 5975 (34) 6613 (27) 1890 (25) 2095 (26) C max (ng/mL) 335 (35) 383 (21) 95 (24) 104 (24) C min (ng/mL) 187 (37) 228 (32) 69 (30) 82 (27) T max (h) 12 (27) 1.5 (42) 15 (27) 1.9 (57) % Fluctuation 61 (57) 59 (35) 34 (72) 26 (47)

AUC 0-24: Area Under the Curve in a 24-hour dosing interval; C max: Peak Concentration in a 24-hour dosing interval; C min: Trough Concentration in a 24-hour dosing interval; T max: Time to Peak Concentration

Figure 1: Mean Steady-State Tramadol (a) and M1 (b) Plasma Concentrations on Day 8 Post Dose after Administration of 200 mg Tramadol Hydrochloride Extended-Release Tablets Once-Daily and 50 mg Tramadol IR Tablets Every 6 Hours.

Question:
Is this medication order safe for the given patient profile? Provide reasons to support your answer.

D. Sample Pharmacist's Chain-of-Thought for the order in Appendix A

Okay, so I'm trying to figure out if the doctor's order for metoclopramide is safe for this patient. Let me start by going through all the information provided.

The patient is a 71-year-old woman with Parkinson's disease, Type 2 Diabetes, and stable ischemic heart disease. She's currently admitted for vertigo and gastroesophageal reflux disease. Her medications include glipizide, clopidogrel, and levodopa/benserazide. She has no known allergies.

She's being treated for vertigo and GERD with metoclopramide 10mg three times a day.

Looking at the drug monograph, it mentioned that patients with preexisting Parkinson's disease should be given metoclopramide cautiously, if at all, since such patients may experience exacerbation of parkinsonian symptoms when taking metoclopramide. This patient has Parkinson's disease. This is a significant risk factor. This order is already appearing to be not safe.

Looking at her lab results: her eGFR is 55, which is just above the threshold for renal impairment (usually 60 or above is okay, so 55 might be a bit low but not necessarily problematic yet). Her sodium is 141, which is within normal range. Her potassium is 4.0, which is within normal range. Magnesium is 0.8, which seems normal.

Her fasting glucose is 7.1, and HbA1C is 6.5%, which are both within target range for diabetes management. Her cholesterol levels are normal as well.

The drug monography mentioned that treatment with metoclopramide can cause tardive dyskinesia. This risk increases with duration of treatment and total dose. The warning says to avoid using it for longer than 12 weeks unless absolutely necessary. Let me see if this is relevant to the patient. Here, the patient is being treated for vertigo and GERD. These diagnosis are usually acute in nature. Will the patient be on metoclopramide for 12 weeks? For an acute diagnosis, I don't think the patient will be on long term metoclopramide. I am assuming metoclopramide is newly initiated.

Also, the patient's age is a factor. The elderly are more prone to developing tardive dyskinesia, and the risk increases with age. Her age is 71, which is quite advanced, so this adds another layer of concern.

The patient's current issue of vertigo and GERD makes metoclopramide a reasonable treatment, but the risk of worsening Parkinsonism symptoms and tardive dyskinesia is significant, especially given her age and history of Parkinson's disease.

In conclusion, the risk of metoclopramide significantly outweighs the benefit due to the patient's medical history of Parkinson's disease and age. Use should be avoided in patients with Parkinson disease and other patients being treated with antiparkinsonian drugs.

E. Human Evaluation Guideline

Below is a guideline for human evaluation, designed to minimize subjectivity in scoring. Most evaluation dimensions are rated on a scale from 1 to 4, where 4 represents the highest quality and 1 the lowest. However, for logical coherence and hallucination, a binary scoring system is applied: a score of 4 indicates acceptable output, while a score of 1 denotes an unacceptable response.

1. Logical coherence: Does the reasoning logically lead to the final decision or recommendation stated regardless of medical correctness? Score 4: Reasoning logically supports and aligns with the final answer; Score 1: Reasoning contradicts or fails to justify the final answer

2.Correctness of reasoning: Do the statements in the generated reasoning factually align with the medical guideline? Score 4: Fully align without any mistake; Score 3: Minor misalignment but main idea still the same; Score 2: Major misalignment but some part is correct; Score 1: Totally incorrect

3. Conciseness: Linguistic efficiency and brevity - Is the reason clear and concise without unnecessary or filler content? Score 4: The reasoning is clear, concise and meaningful, and free of generic statements; Score 3: Minor parts include generic statements or filler content without obscuring the decision; Score 2: Noticeable part of the reason is generic or filler,

but key conclusion remains visible; Score 1: The major part or entire statement is generic or unnecessary, drowning the critical information

4.Clarity: Is the reasoning process easy for human to follow and understand? Score 4: Flow is very easy to follow; Score 3: Generally understandable but may require minor re-reading for complex parts; Score 2: Require significant effort to read; Score 1: The flow is not easy to follow and unclear

5.Relevant: Is the reasoning relevant to the true reasoning thought process? Score 4: The entire reasoning addresses the patient and medication-specific risk with precise evidence; Score 3: reasoning covers key factors but some overgeneralization; Score 2: Some parts address the patient and medication but key factors are ignored; Score 1: The reasoning is unrelated to the patient condition or medication

6.Duplication: Is there any repeated statement in the reasoning? Score 4: Zero repetition; every sentence adds unique value; Score 3: Minor repetition (single restatement for emphasis); Score 2: Core argument repeated $\geq$2 times without progression; Score 1: Excessive copy-paste or circular reasoning

7.Hallucinations: Does the reasoning invent unsupported facts or claims not present in the monograph or known medical evidence? Score 4: No hallucinations — all statements evidence-based; Score 1: Contains hallucinated or fabricated clinical facts