Leveraging Large Language Models for Clinical Medication Verification

Group 4: Ng JingYi Shanice, Ng Qi Hao Joshua, Sun Tian Yu
Steven Harta Prawira, Yan Siyang

The Role of Hospital Pharmacists

Participation in ward rounds

Verify medication appropriateness

Counsel patients

Respond to drug inquiries

Manage procurement & dispensing

The Role of Hospital Pharmacists

Verify medication appropriateness

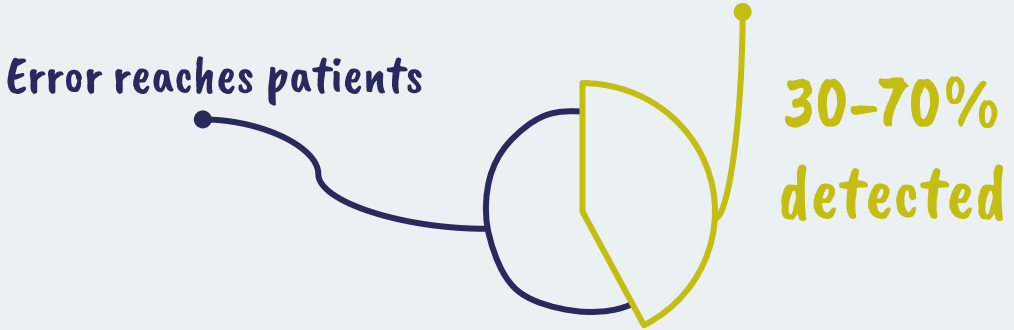
- Pharmacists need to manually cross-reference patient information and external drug database to identify drug-related problems.

High Workload, High Risks

Subtle interactions and context-specific risks can be easily missed.

Medication Errors are a Global Concern

- Medication error: A failure in the medication-use process that may or may not result in patient harm
- Half of all medication errors occur during prescribing or ordering stages
  - Nurses and pharmacists detect **30%-70%** of these errors.



Our proposed solution

Patient Information

Medication Order Information

Drug Monograph Information

LLM

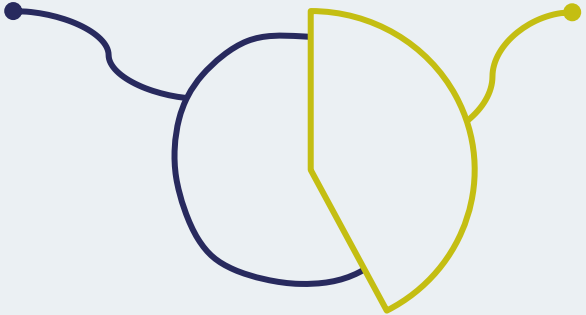
Order is Safe

Justifications of why medication order is considered “Safe”.

Order is Unsafe

Justifications of why medication order is considered “Unsafe”.

Error reach patient

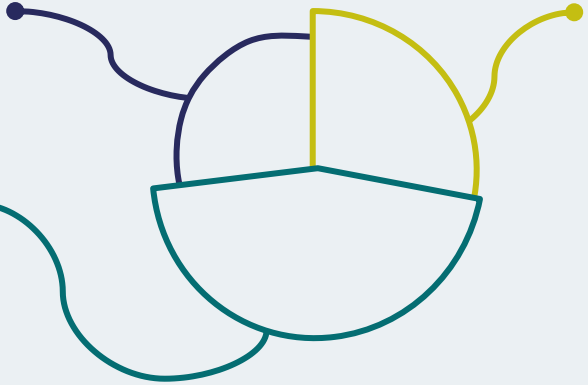


30-70% detected by nurses/ pharmacist

Our goal

Reduced medication error reach patients

LLM-detected medication errors



Reduced verification workload on pharmacists

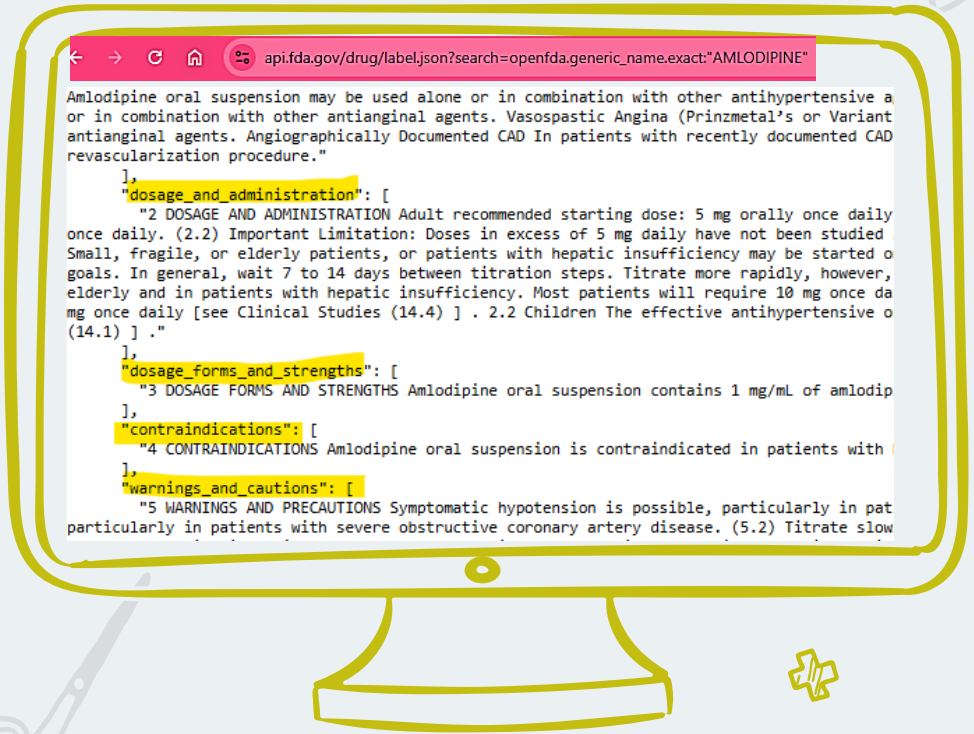
01

Dataset

Drug Monograph • Patient and Medication Order • Processing • Splitting

Drug Monograph

- openFDA Drug Product Labelling API
- Contains drug labels provided by manufacturers
- Segmented into sections
- To optimize context window usage (critical for LLM efficiency), irrelevant sections were removed.



Patient information

- Synthetic datasets preserves privacy and enhances research reproducibility
- Patient profiles are curated by a practising clinical pharmacist with > 5 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.
- Data includes: demographic information, past medical history, allergy status, concomitant medications, relevant lab or imaging results

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
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

Example of Patient Profile after processing

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
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

Patient is a 71 years old female, height 1.56m and weighs 51kg. Latest heart rate is 92 bpm. Latest systolic blood pressure is 113. Patient has past medical history of Parkinson disease, Type 2 Diabetes Mellitus, Stable Ischemic Heart Disease. Currently patient is admitted for Vertigo, Gastroesophageal reflux disease.
Patient is also taking the following medications: Oral Glipizide 15mg two times a day before meal, Oral Clopidogrel 75mg every morning, Oral Benserazide 25 mg/levodopa 100 mg twice a day.
Patient has known allergy to: null.
Some recent lab results and reports are shown below:
White Blood Cell = 2.32 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 mÂ²
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: null
Imaging Report: Chest X-ray: No significant lung consolidation, pleural effusion or pneumothorax. Cardio mediastinal contour unremarkable.

Medication order information

- Features consists of the route, medication name, formulation, dosing and frequency

Example of Medication Order after processing

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

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

Medication order information

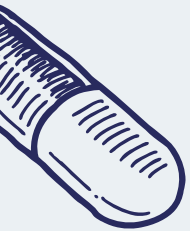
- Each medication order is labeled as ”safe” or ”unsafe,” along with the pharmacist’s reasoning behind the classification.
- Dataset size: 50 medication orders

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

Dataset Splitting

- Stratified splitting based on medication-related problem.
  - Ensures that both the training and testing datasets are representative of the range of medication-related issues observed

Train-Test Split



02

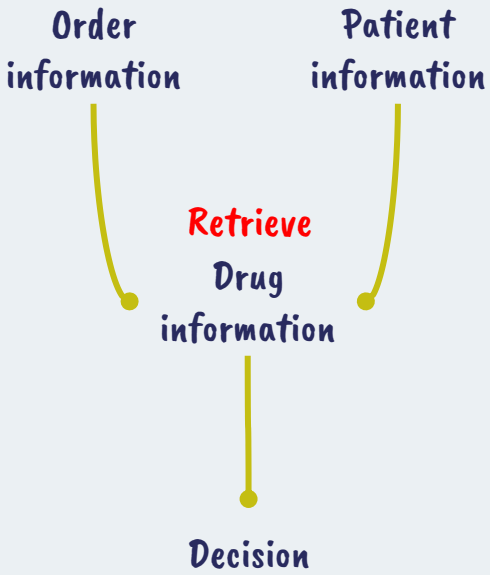
Methodology

Retrieval Augmented Generation •
Few-Shot Prompting • Supervised Finetuning

Approach #1: Retrieval-Augmented Generation (RAG)

RAG Combines information retrieval with LLM reasoning

- Mitigates hallucination issues, improving the professional reliability of LLM-generated responses.
- Mirrors the typical pharmacist workflow.
- RAG potentially enhances efficiency by relieving the LLM from internal retrieval, enabling faster and more focused reasoning.



Approach #1: Retrieval-Augmented Generation (RAG)

RAG approach's responses is heavily influenced by the quality and accuracy of the retrieved drug information

- Drug monograph were first pre-processed to extract only key actional sections such as boxed warnings, contradictions, dosage guidelines.
- Chunk length should preserve coherent clinical information
- FAISS is used to store the vectorized chunks for efficient retrieval
- Additional filtering condition is being used to ensure only relevant drug is retrieved

Approach #1: RAG fine tuning

Competing Objectives

- Ensuring sufficient coverage of potential safety issues across different sections within the monograph
- Maintaining a compact prompt size to prevent overwhelming the LLM with too much noise

Fine Tuning

- A maximum of 1000 characters per chunk was targeted to ensure clinical information coherently preserved
- Only the top 5 results are retrieved to balance the prompt size and coverage

Lastly, the constructed prompt template is fed to three different LLMs (Qwen0.5B, GPT-3.5 and GPT-4o)

Approach #1: RAG prompt construction

In our RAG setup, the prompt template consist of **four** main segments:

- **Patient Clinical Context:** Structured information extracted from patient profile
- **Medication Order Details:** The specific medication(s) ordered for that patient
- **Relevant Drug Information:** Retrieved by RAG retriever based on the identified order
- **Question:** “Is this medication order safe for the given patient profile? Provide reasons to support your answer.”

Approach #1: Sample prompt for RAG

Patient Profile:
46-year-old male, 1.68m, 86kg.
Vital signs: HR 87 bpm, SBP 119 mmHg.
Past medical history: Vertigo.
Current admission: Type 2 Diabetes.
Allergy: None reported.

Recent Labs:
WBC: 13.6 $\times 10^9$/L, Hb: 11.3 g/dL, Platelet: 355 $\times 10^9$/L
Creatinine Clearance: 47.0 mL/min,
eGFR: 40.0 mL/min/1.73m^2
CRP: 271.8 mg/L, Sodium: 141 mmol/L,
Potassium: 3.8 mmol/L
Microbiology: Heavy growth of methicillin-sensitive Staph aureus.

Order details:
The doctor ordered intravenous cefazolin 2g every 8 hours.

Relevant Drug Information:
(CEFAZOLIN_data)
with severe renal impairment (creatinine clearance of 20 to 5 mL/min.)
may be given 10 percent of the normal daily dose every 24 hours. All dosage recommendations apply after an initial loading dose. RECONSTITUTION Preparation of Parenteral Solution: Parenteral drug products should be SHAKEN WELL when reconstituted, and inspected visually for particulate matter prior to administration. If particulate matter is evident in reconstituted fluids, the drug solutions should be discarded. When reconstituted or diluted according to the instructions below, Cefazolin for Injection is stable for 24 hours at room temperature or for 10 days if stored under refrigeration (5¬∞C or 41¬∞F). Reconstituted solutions may range in color from pale yellow to yellow without a change in potency. Single-Dose Vials: For I.M. injection, I.V. direct (bolus) injection or I.V. infusion, reconstitute with Sterile Water for Injection according to the following table. SHAKE WELL. Discard unused portion. Vial Size Amount

Approach #2: Direct context few-shot prompting

- Refinement of RAG & Prompt Construction
  - **Patient Clinical Context and Medication Order Details:** Unchanged
  - **Relevant Drug Information:** Full drug monograph retrieved matching drug name rather than using a retrieval system to find relevant chunks
  - **Few shot prompt:** 4 examples provided (2 safe and 2 unsafe)
  - **Question:** “Is this medication order safe for the given patient profile? Provide reasons to support your answer.”
- Larger model DeepSeek-R1-Distill-Llama-3B

Approach #3: Supervised Fine-Tuning with Unsloth

- General LLMs lack domain-specific reasoning structure
  - Model: **DeepSeek-R1-Distill-Llama-8B** (strong CoT reasoning)
  - **Unsloth**, a lightweight LoRA-based fine-tuning framework
  - This approach requires a fine-tuning dataset

Approach #3: Supervised Fine-Tuning with Unsloth

- This approach requires a fine-tuning dataset
  - **Question**: Patient Information + Med Order Information + Drug Monograph
  - **Complex Chain-of-Thought**: 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

Approach #3: Supervised Fine-Tuning with Unsloth

- Specific structure of the pharmacist's reasoning is highly specialized to the medication verification process.
  - Background
  - Linked past medical history to drug monograph
  - Continues to assess labs, draws link between reason of current hospital stay to drug monograph
  - Take everything from above into consideration and weigh risks vs benefit

Excerpt from sample of Pharmacist's Chain-of-Thought rationales

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. …
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. …

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.

Approach #3: Supervised Fine-Tuning with Unsloth

Sample Response in Finetuning Dataset

This medication order is **not safe** because symptoms of Parkinson disease may be exacerbated with metoclopramide use. The doctor should discontinue the medication.

- Pharmacists don't have time to read lengthy reasoning during rounds.
- The model is trained to generate clear, structured outputs: Safe or Unsafe, with concise explanation if needed.

Approach #3: Supervised Fine-Tuning with Unsloth

- Configuration:
  - 4-bit precision loading
  - LoRA on attention layers (rank = 16)
  - 12 GB VRAM-constrained setup

- CoT generation enforced to mimic pharmacist reasoning and maintains transparency.
- Zero-shot prompting during inference to minimize context window length and simulate real-world deployment.
- ~2× faster than Hugging Face baseline

03

Evaluation Metric Selection

Prediction Quality • Reasoning Quality

03 Evaluation Metrics Selection

2 Key Metrics

Assessing accuracy and relevance of LLM-generated response

Prediction Quality

- 'Unsafe' = positive case
- Recall (Minimizing false negatives)
- Specificity (Minimizing false positives
- Prioritizing Recall vs Specificity (weighted score)

Reasoning Quality

- Embedding-based similarity vs ground-truth reasoning.
- Human expert scoring (7 key criteria)

Human Expert Scoring

01
Logical Coherence
Reasoning logically lead to final decision or recommendation

02
Correctness
Reasoning factually align with medical guidelines

03
Conciseness
Reasoning clear and concise without unnecessary filler

04
Clarity
Reasoning easy for human to follow and understand

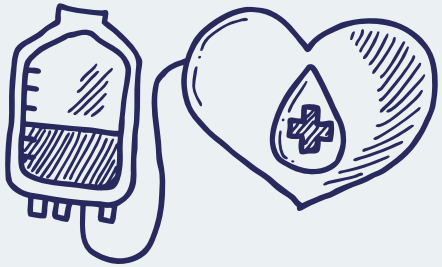
05
Relevance
Reasoning relevant to true thought process

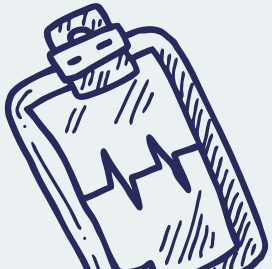
06
Duplicate
Reasoning comes with repeated statements?

07
Hallucinations
Reasoning invent unsupported claims with no medical evidence?

04

Result Discussion





4.1 Prediction Quality

Model	Recall	Specificity	Weighted Score
Base Model (Deepseek-R1-Llama-8B)	0.2308	0.5833	0.3365
Base Model w Monograph	0.4615	0.8333	0.5731
Fine-tuned* (temp = 1)	0.6923	0.6667	0.6846
Fine-tuned (temp = 0.5)	0.6923	0.7500	0.7096
Qwen0.5b_RAG	0.0000	1.0000	0.3000
GPT3.5_RAG	0.7692	0.5000	0.6885
GPT4o_RAG	0.6154	0.5833	0.6058
Few Shot Prompt**	0.9231	0.1667	0.6962

* Base model with monograph fine-tuned with SFT using Unsloth
** Direct context few shot prompting model

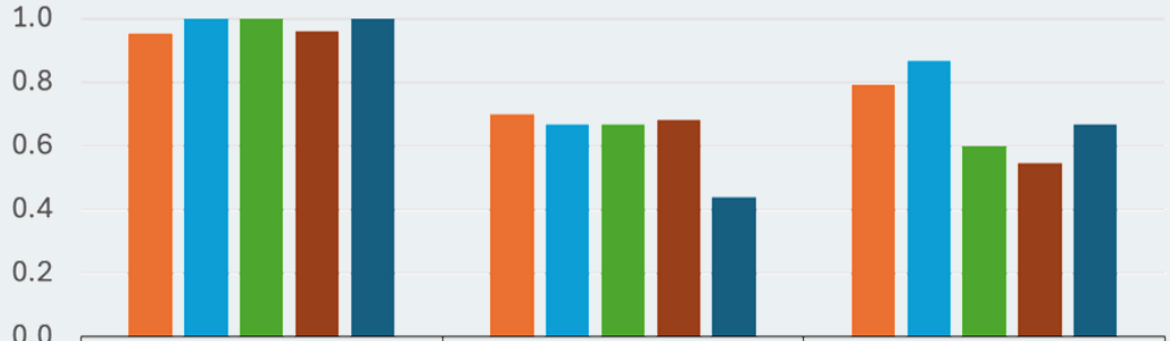
4.2 Reasoning Quality – Cosine Similarity

Model	Cosine Similarity Score
Base Model (Deepseek-R1-Llama-8B)	0.7199
Base Model 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

* Base model with monograph fine-tuned with SFT using Unsloth

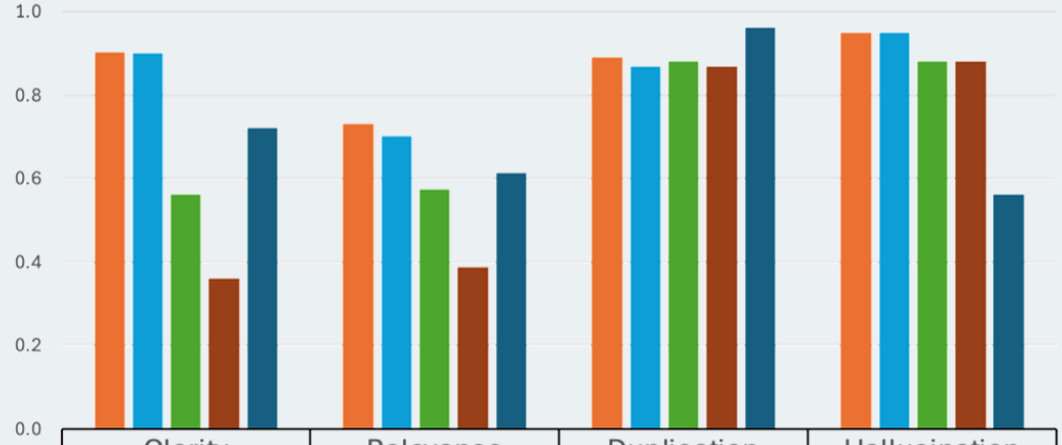
** Direct context few shot prompting model

4.3 Human Evaluation Results



	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

4.3 Human Evaluation Results



	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

4.3 Human Evaluation Results

Criteria	Fine-Tuned (temp=1)	Fine-Tuned (temp=0.5)	GPT3.5_RAG	GPT4o_RAG	Few-shot prompting	Avg.
Logical coherence	0.953	1.000	1.000	0.960	1.000	0.983
Correctness	0.700	0.667	0.667	0.680	0.440	0.631
Conciseness	0.793	0.867	0.600	0.547	0.667	0.695
Clarity	0.903	0.900	0.560	0.360	0.720	0.689
Relevance	0.730	0.700	0.573	0.387	0.613	0.601
Duplication	0.890	0.867	0.880	0.867	0.960	0.893
Hallucination	0.950	0.950	0.880	0.880	0.560	0.844
Avg.	0.846	0.850	0.737	0.669	0.709	

4.3 Human Evaluation Results

Criteria	Fine-Tuned (temp=1)	Fine-Tuned (temp=0.5)	GPT3.5_RAG	GPT4o_RAG	Few-shot prompting	Avg.
Logical coherence	0.953	1.000	1.000	0.960	1.000	0.983
Correctness	0.700	0.667	0.667	0.680	0.440	0.631
Conciseness	0.793	0.867	0.600	0.547	0.667	0.695
Clarity	0.903	0.900	0.560	0.360	0.720	0.689
Relevance	0.730	0.700	0.573	0.387	0.613	0.601
Duplication	0.890	0.867	0.880	0.867	0.960	0.893
Hallucination	0.950	0.950	0.880	0.880	0.560	0.844
Avg.	0.846	**0.850**	0.737	0.669	0.709	

4.3 Human Evaluation Results

Criteria	Fine-Tuned (temp=1)	Fine-Tuned (temp=0.5)	GPT3.5_RAG	GPT4o_RAG	Few-shot prompting	Avg.
Logical coherence	0.953	1.000	1.000	0.960	1.000	0.983
Correctness	0.700	0.667	0.667	0.680	0.440	0.631
Conciseness	0.793	0.867	0.600	0.547	0.667	0.695
Clarity	0.903	0.900	0.560	0.360	0.720	0.689
Relevance	0.730	0.700	0.573	0.387	0.613	0.601
Duplication	0.890	0.867	0.880	0.867	0.960	0.893
Hallucination	0.950	0.950	0.880	0.880	0.560	0.844
Avg.	0.846	**0.850**	0.737	0.669	0.709	

Lower temperature settings -> more deterministic and focused outputs

4.3 Human Evaluation Results

Criteria	GPT3.5_RAG	GPT4o_RAG
Logical coherence	1.000	0.960
Correctness	0.667	0.680
Conciseness	0.600	0.547
Clarity	0.560	0.360
Relevance	0.573	0.387
Duplication	0.880	0.867
Hallucination	0.880	0.880
Avg.	0.737	0.669

* Higher Score vs Lower Score

Sample reasoning from GPT3.5

Based on the patient profile provided, the medication order for intravenous cefazolin 1g every 6 hours may not be safe for this patient. The patient has a creatinine clearance of 47.0 mL/min, which falls within the range where dosage adjustments are recommended. According to the drug information provided, patients with creatinine clearance rates of 35 to 54 mL/min should receive full doses but dosage should be restricted to at least 8 hour intervals.

In this case, the doctor has ordered cefazolin every 6 hours, which may not be in line with the recommended dosage adjustments for patients with reduced renal function. It would be advisable for the doctor to reconsider the dosing frequency or adjust the dosage based on the patient's renal function to ensure the medication is safe and effective for the patient.

Sample reasoning from GPT4o

The medication order for intravenous cefazolin 1g every 6 hours is not safe for the given patient profile due to the patient's reduced renal function.

Reasons to support:

1. **Renal Function**: The patient has a creatinine clearance of 47.0 mL/min and an eGFR of 40.0 mL/min/1.73 m¬≤, indicating moderate renal impairment. According to the cefazolin dosage guidelines, patients with a creatinine clearance rate of 35 to 54 mL/min should receive full doses but at intervals of at least 8 hours, not every 6 hours as currently ordered.

2. **Dosage Adjustment**: The current order of cefazolin 1g every 6 hours does not adhere to the recommended dosing interval for patients with the patient's level of renal function. The dosing interval should be adjusted to every 8 hours to prevent potential accumulation of the drug and toxicity.

3. **Safety and Efficacy**: Administering cefazolin more frequently than recommended for this level of renal function can lead to higher serum concentrations, increasing the risk of adverse effects and toxicity.

Therefore, the medication order should be revised to cefazolin 1g every 8 hours to align with the patient's renal function and ensure safe and effective treatment.

4.3 Human Evaluation Results

Criteria	Fine-Tuned (temp=1)	Fine-Tuned (temp=0.5)
Logical coherence	0.953	1.000
Correctness	0.700	0.667
Conciseness	0.793	0.867
Clarity	0.903	0.900
Relevance	0.730	0.700
Duplication	0.890	0.867
Hallucination	0.950	0.950
Avg.	0.846	**0.850**

* Higher Score vs Lower Score

Sample reasoning from Temp=1

The medication order for enalapril is considered safe for this patient with careful monitoring and precautions in place. Key considerations include the patient's history of atrial fibrillation, impaired kidney function, and use of a nasogastric tube. Starting with a lower dose and closely monitoring blood pressure and kidney function are recommended.

Final Answer: The medication order is safe with close monitoring of blood pressure and kidney function.

Sample reasoning from Temp=0.5

The medication order for enalapril 5mg every morning via nasogastric tube is not safe for this patient. The patient is already on two antihypertensive medications (amlodipine and bisoprolol), and adding enalapril could increase the risk of hypotension and other side effects. Additionally, his age and eGFR may increase the risk of side effects. Therefore, the medication order might not be safe for this patient.

Answer: The medication order is not safe for this patient.

4.3 Human Evaluation Results

- **Limitation** for finetuned models

  Provide decision + recommendation action **w/o reasoning**

- **Solution**

  Revise prompts and training data to explicitly enforce structured reasoning

The medication order of oral enalapril 10 mg every morning is not safe for this patient in his current condition. The doctor should discontinue this order and consider intravenous enalapril instead, along with close monitoring of blood pressure and renal function.

4.3 Human Evaluation Results

(1) Hallucination

- **Context (Order 45)**

"Patient is a 55 years old **male**, height 1.75m and weighs 78kg……"

- **Fine-Tuned (temp=1) reasoning**

"…Mycophenolic acid is contraindicated in **pregnancy** due to risks of pregnancy loss and congenital malformations…"

(2) Incomplete Reasoning

No underlying rationale for reasoning

Only list of medication usage and patient conditions.

4.3 Human Evaluation Results

(3) Logical Incoherence

Order 34: “Ensuring the patient can **take the medication orally is crucia**l…” -> But did not classify as unsafe.

(4) Ambiguous Justification

Order 26: “The medication order is unsafe due to **insufficient information** or **inappropriate dosing**.”

(5) Weak in Under-dosing Detection

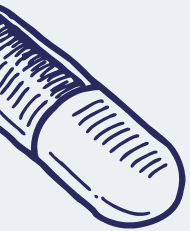
Across models -> priority area for future iterations

4.4 Inferencing Time

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

4.5 Choice between Finetuning & RAG

- **Dependence on Retrieval**
  RAG models rely heavily on retrieval quality; errors weaken reasoning.
- **Clinical Reasoning Internalization**
  Fine-tuning teaches both *what* to say and *how* to say it, matching clinical expectations.
- **Cohesive Case-Specific Outputs**
  Fine-tuned models integrate patient and drug data seamlessly; RAG often feels "stitched together."
- **Model Size Impact**
  Larger fine-tuned models synthesize complex reasoning better than smaller RAG models.
- **Bottom Line**
  Fine-tuning with **Chain-of-Thought** supervision better supports clinical depth and accuracy.



05

Limitation and Future Work

6.1 Limitations

- **Small sample medication orders**
  - Sample of 50 medication orders limits model generalizability
- **Limited Expert Justification**
  - No standardized assessment of explanation quality
  - Does not account for the wide range of clinical judgement
  - Unable to capture hallucinated content

6.2 Future Work

- **Small sample medication orders**
  - Expand dataset via collaboration or anonymized data
- **Limited Expert Justification**
  - Involve multi-expert annotations
  - Utilise a more comprehensive human evaluation rubric
- **Reinforcement Learning**
  - Incorporate human-in-the-loop feedback

Conclusion

- Promising performance in reasoning accuracy
- Limited Scalability: Chain-of-Thought rationales is labor-intensive
- Current solution is not ready for clinical deployment

Thank you!