Data Breathing Life into Emergencies

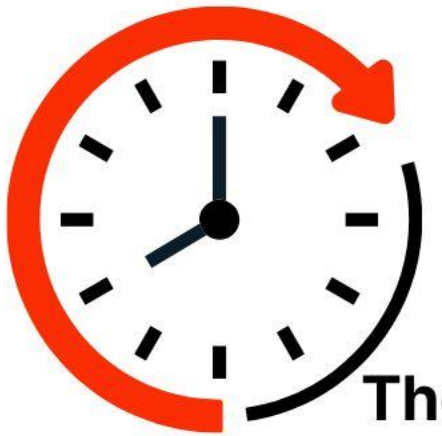
Korea Clinical Datathon 2025

ED multi-agent clinician

2025.10.18. SAT 13:00 | SEOUL, KOREA | TEAM CODE_NOVA

Team Codenova

Introduction_Objectives



Within 5 minutes

The patient arrives at the emergency department

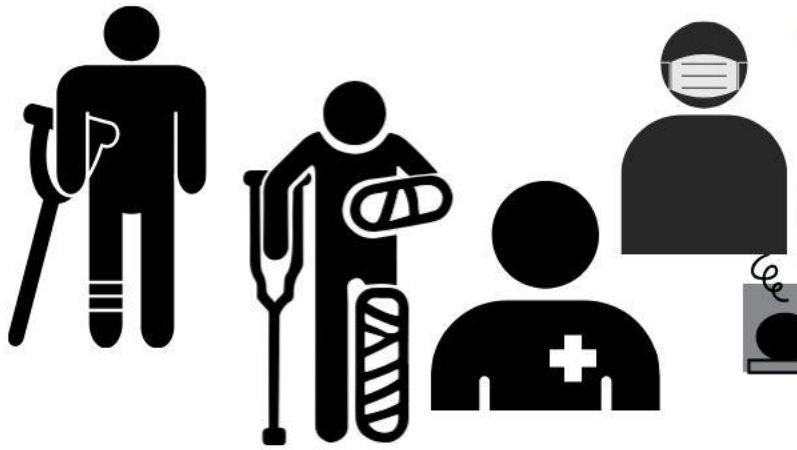
receives a first impression assessment

undergoes an interview and evaluation

is guided to the treatment or waiting area

and treatment begins

Introduction_Objectives



waiting patients.....

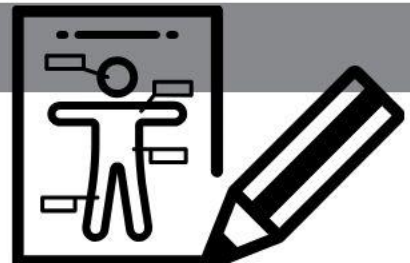
receives a first impression assessment

The patient arrives at the emergency department

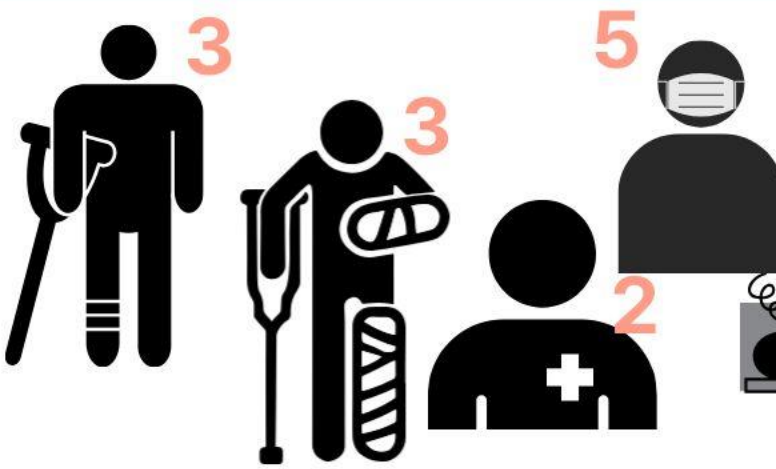
undergoes an interview and evaluation

is guided to the treatment or waiting area

and treatment begins



Introduction_Objectives



waiting patients.....

The patient arrives at the emergency department

The nurse and agent conduct a first impression assessment together.

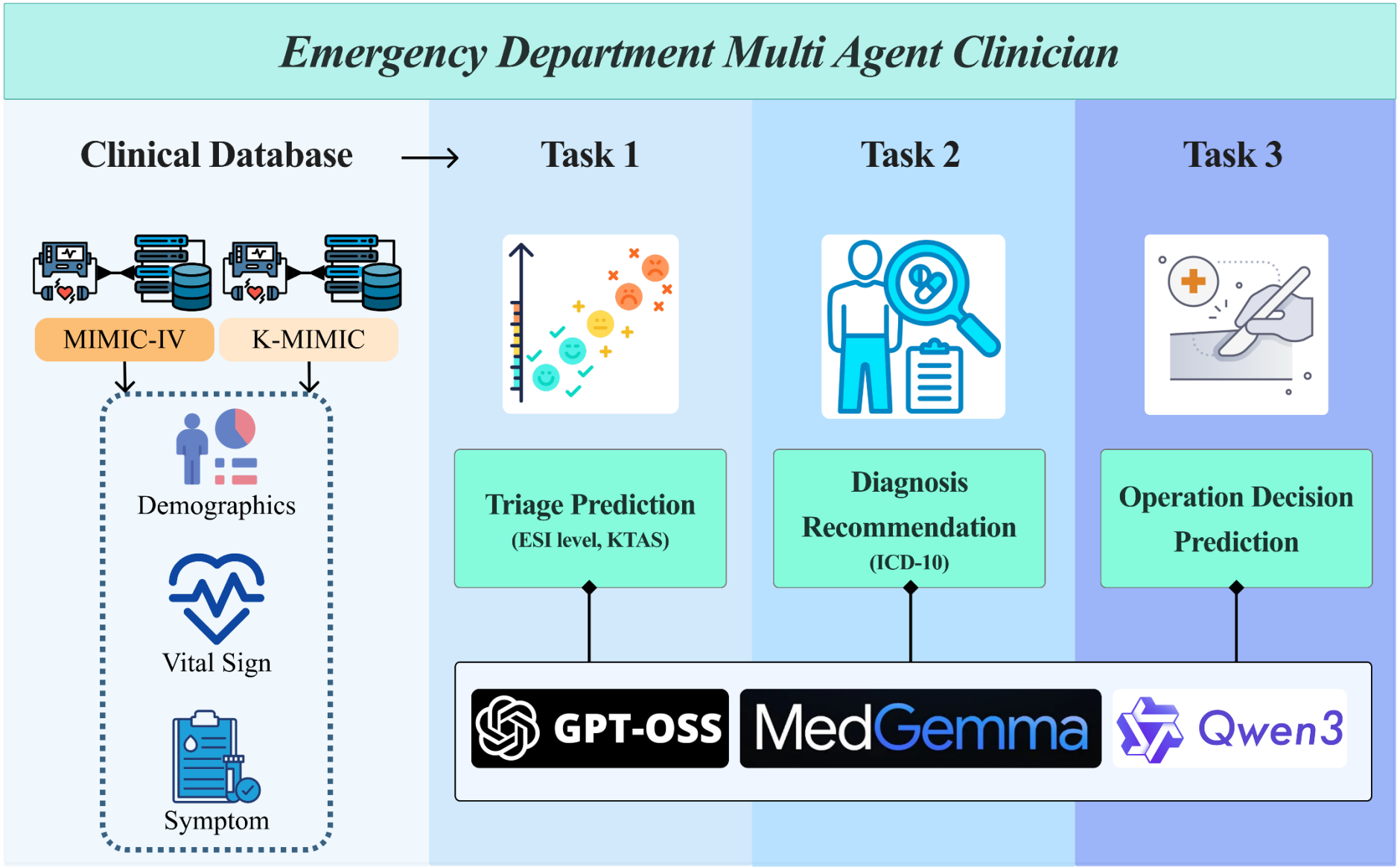
undergoes an interview and evaluation

is guided to the treatment or waiting area

and treatment begins

patient safety

Overview


Emergency Department Multi Agent Clinician
Clinical Database
Task 1
Task 2
Task 3
MIMIC-IV
K-MIMIC
Demographics
Vital Sign
Symptom
Triage Prediction
(ESI level, KTAS)
Diagnosis
Recommendation
(ICD-10)
Operation Decision
Prediction
GPT-OSS
MedGemma
Qwen3

Task 1: Triage Prediction

Motivation

- To reduce critical errors and delays in high-pressure EDs through AI-based triage prediction.

Method

- Benchmarking LLMs to predict KTAS using patient information and initial vitals.

Expected Benefit

- Optimized patient prioritization and improved ED resource allocation

Task 1: Triage Prediction

Prediction Results

Input	Dataset	Model (zero-shot)	Exact Match	Accuracy
V/S, HPI	MIMIC-IV	Claude 3.5 Sonnet + RAG	65.8	77.1
V/S, HPI	MIMIC-IV	Claude 3.5 Sonnet	64.4	82.4
V/S	K-MIMIC	gpt-oss-20B	14	14
V/S	K-MIMIC	gpt-oss-120B	28	26
V/S	K-MIMIC	medgemma-27B	27	26

- K-MIMIC models showed significantly lower accuracy (Exact Match ≤ 28%) compared to reference MIMIC-IV models (≈ 65%)

Task 2: Diagnosis Recommendation

Objective

- To enhance diagnostic efficiency and accuracy at emergency department, while evaluating the clinical applicability of three LLMs (GPT-OSS-20B, Med-Gemma-27B, and Qwen3-Next-80B) in supporting diagnostic decision making

Input Data

- Patient Profile: Gender, Age, Race
- Vital Signs: Temperature, HR, RR, SBP, DBP, O2 saturation
- Clinical Notes: Chief complaint, HPI (history of present illness)

Process Flow

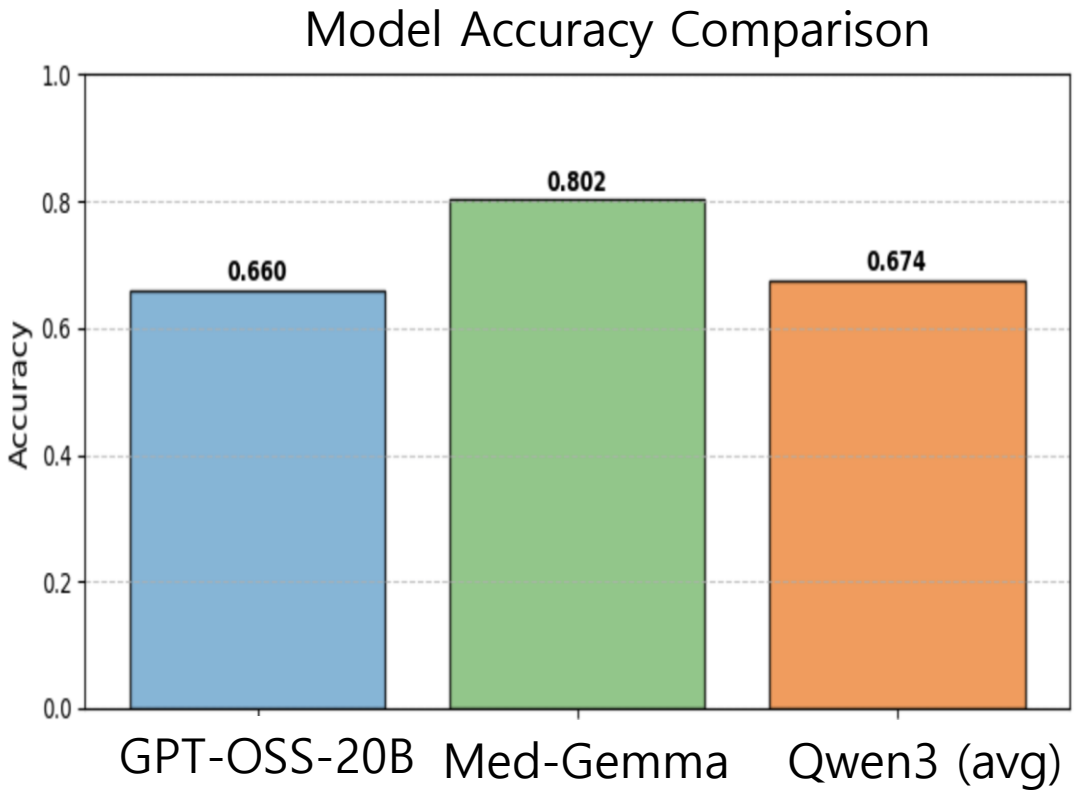
- Input patient data
- Extract ground truth (diagnosis records) from Clinical Notes
- Predict diagnosis (by GPT-OSS, Med-Gemma, Qwen3)
- Evaluate was conducted by by three clinicians

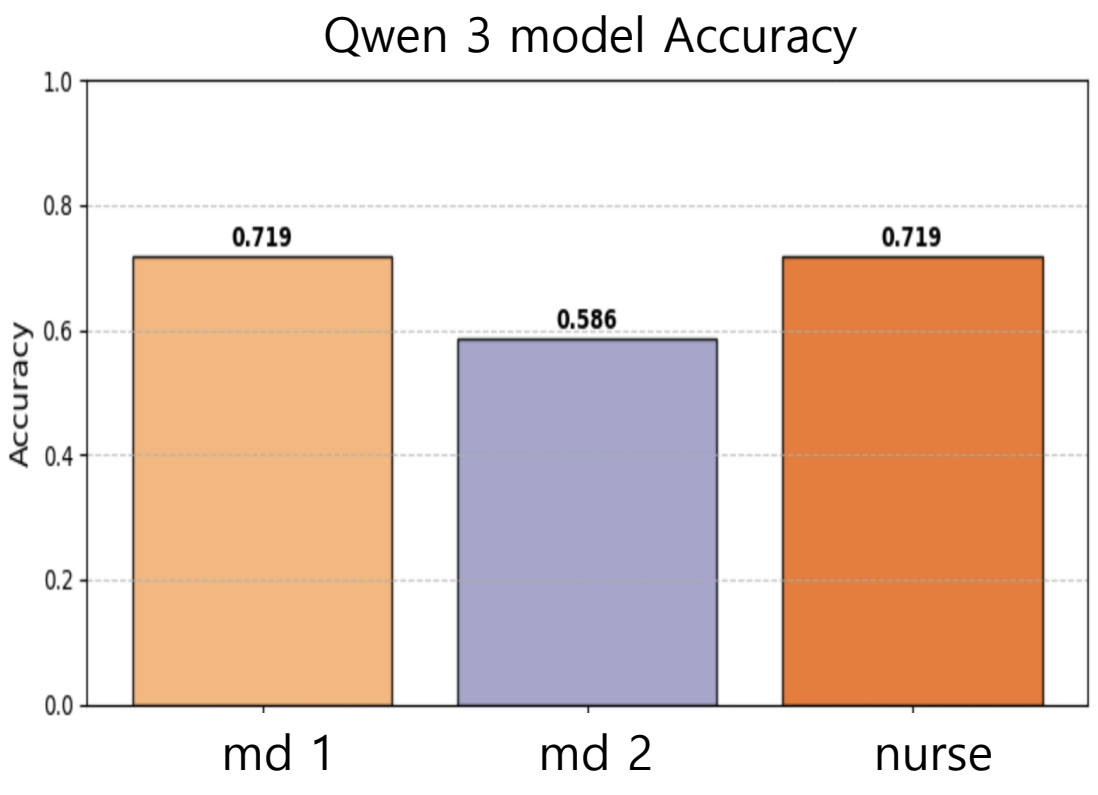
Interpretability

- Token SHAP[1] and analysis of uncertainty distributions by entropy

[1] Goldshmidt, R., & Horovicz, M. (2024). Tokenshap: Interpreting large language models with monte carlo shapley value estimation. *arXiv preprint arXiv:2407.10114.*

Task 2: Diagnosis Recommendation





- Med-Gemma achieved the highest diagnostic accuracy (0.802)
- Among Qwen3 variants, MD1 and nurse showed equal evaluation performance (0.719), while MD2 performed the lowest (0.586).

Temperature = 0 at all experiments

Task 2: Diagnosis Recommendation

- The following case represents a **correct prediction** made by all three clinicians.

```
You are a clinical decision support model. Given the following patient data, predict the most likely primary and secondary diagnosis. Patient info: Gender: Female, Race: WHITE - RUSSIAN, Age: 72 Initial vitals: Temperature: 97.2, Heartrate: 84.0, resprate: 16.0, o2sat: 94.0, sbp: 114.0, dbp: 72.0 HPI: ___ year old woman without significant past medical history who is s/p colonoscopy and polypectomy on ___, presenting with blood per rectum. On colonoscopy, a sessile 8mm benign-appearing polyp and sessile 2cm multilobular poly were completely removed from the proximal transverse and mid -ascending colon respectively. After the colonoscopy she had two episodes "like flowing blood", slept through the night, and then at 8 am on day of presentation had two bloody BMs within 30 minutes where the blood was noticeably darker. She has had occasional dizziniess and weakness recently. Return pred_primary_diagnosis, pred_secondary_diagnosis.
```

Ground truth: Post-polypectomy bleed
Predicted diagnosis (GPT-OSS): Post-polypectomy bleeding

- The following case represents an **incorrect prediction** made by all three clinicians.

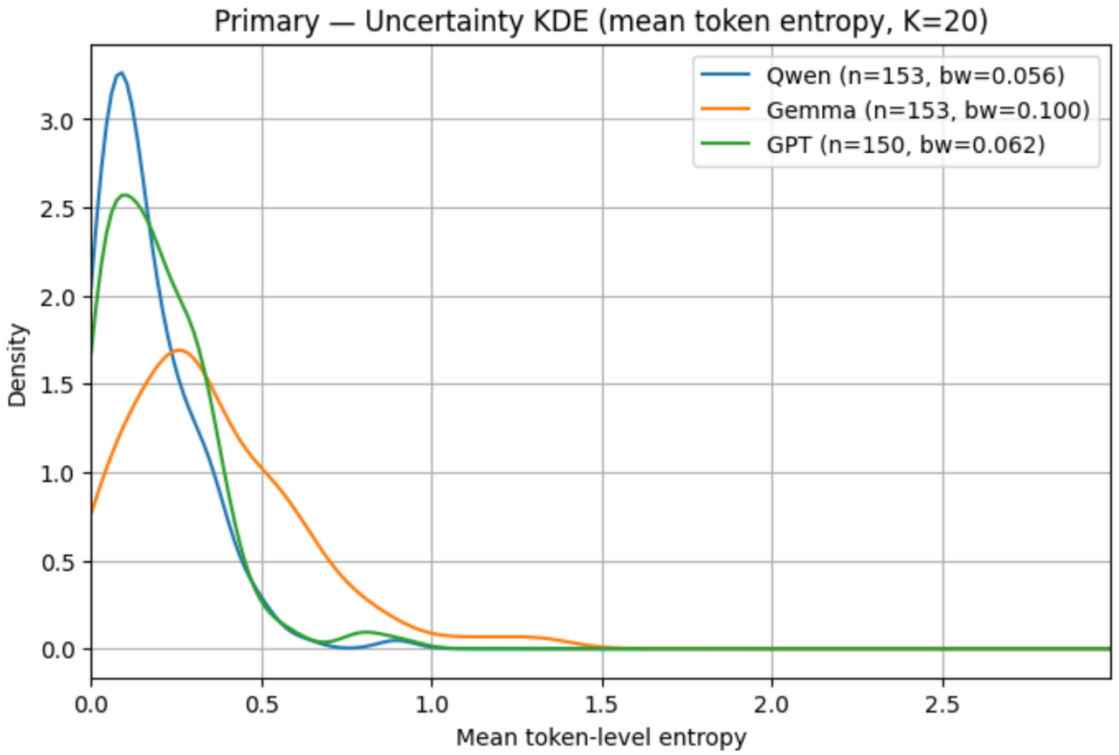
```
You are a clinical decision support model. Given the following patient data, predict the most likely primary and secondary diagnosis. Patient info: Gender: Male, Race: UNKNOWN, Age: 88 Initial vitals: Temperature: 97.2, Heartrate: 80.0, resprate: 16.0, o2sat: 97.0, sbp: 125.0, dbp: 86.0 HPI: HISTORY OF PRESENT ILLNESS: Patient is an ___ year old male with history of DM, CAD and A.Fib anticoagulated with coumadin who was initially transferred from an OSH after a mechanical fall. Head CT at ___ concerning for cerebellar bleed so patient transferred to ___ for neurosurgical evaluation. Patient reports he was carrying a pitcher of water when he slipped on some spilled water falling backwards and pitcher landing on him. He denies any head strike, loss of conciousness. He denies prodrome of lightheadedness, dizziness, chest pain, shortness of breath or other prodromal symptoms. Return pred_primary_diagnosis, pred_secondary_diagnosis.
```

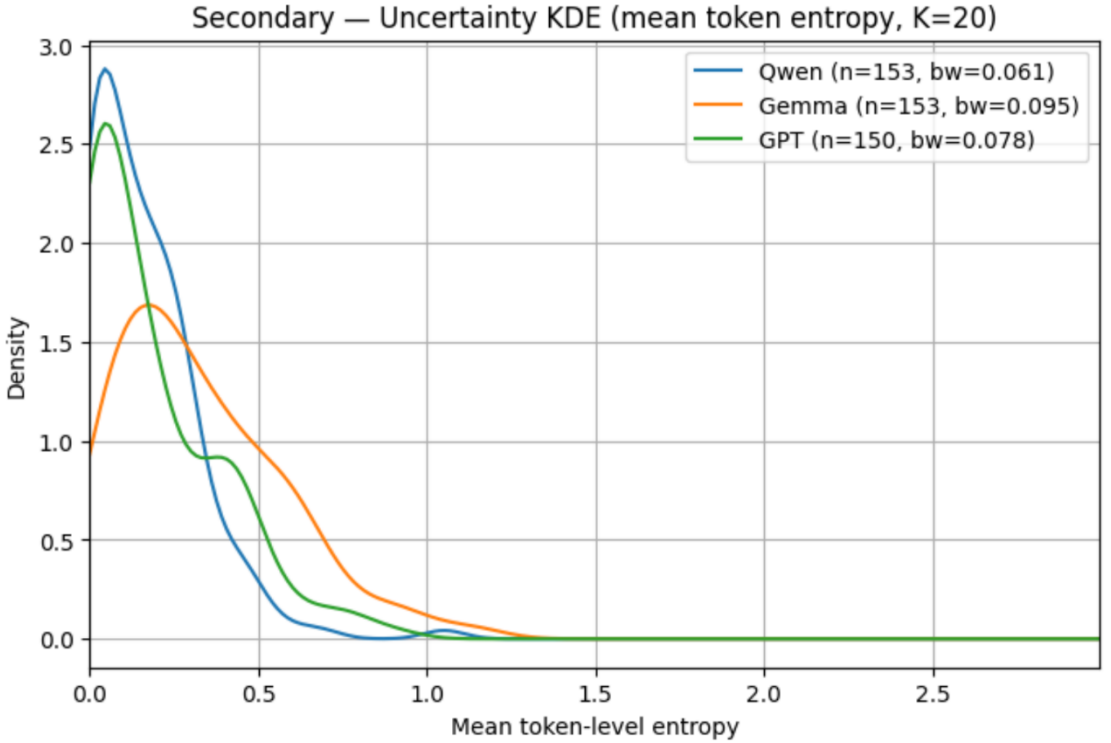
Ground truth: T11 compression fracture
Predicted diagnosis (GPT-OSS): Cerebellar hemorrhage

Clues

Token SHAP was evaluated using GPT-OSS

Task 2: Diagnosis Recommendation





- **Uncertainty distribution (mean token-level entropy)** of three LLMs
- Qwen (blue) and GPT (green) exhibit lower entropy, indicating Computational Overconfidence compared to Gemma (orange), even though both Qwen and GPT demonstrate lower performance.

Task 3: Operation Decision Prediction

Objective

- To assess how different prompting strategies (Zero-shot, CoT, Multi-Agent, RLM) affect the accuracy of predicting surgical i ntervention during ED visits, as part of the ED Multi-Agent Clinician framework.

Input Data

- Source**:** MIMIC-IV-ED
- Rows: 2,160 ED visits → 200 used for validation
- Input fields
  - Demographics: age / sex / race
  - Initial vital sign: first ED vital signs
  - HPI (history of present illness): present illness summary → from *discharge note*
- Ground Truth
  - 1 = Operation / 0 = No Operation → from *discharge note*

Methods

- **A. Zero-Shot:** Direct binary classification prompt
- **B. Chain-of-Thought (CoT):** Adds step-by-step reasoning instruction
- **C. Multi-Agent:** 3-round specialist voting consensus (ED / Surgeon / Physician)
- **D. Recursive Language Model (RLM):** Recursively analyzes complex context before deciding

Model **:** gpt-oss-20b

Task 3: Operation Decision Prediction

A. Zero-Shot	B. Chain-of-Thought (CoT)	C. Multi-Agent	D. Recursive Language Model (RLM)
Fast but conservative (often underestimates surgery)	Adds step-by-step reasoning instruction, still limited recall	3-round specialist voting con sensus (final by majority), reducing false negatives	Recursively manages long context, enhancing deci sion robustness
Clinical reasoning depth : Minimal	Clinical reasoning depth : Moderate	Clinical reasoning depth : High	Clinical reasoning depth : Adaptive~Very High

A.

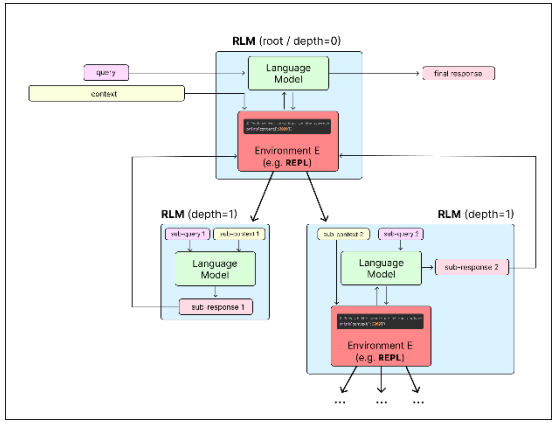
```
[task]
Determine whether the following emergency department (ED) patient
required surgical intervention during this visit.
[classification]
 - 0: No operation
    (managed medically, discharged, or admitted without surgery)
 - 1: Operation
    (underwent or was prepared for surgery during this visit)
```

C.

```
AGENT_ROLES = {
"Emergency Doctor": "balances triage and urgency.",
"Surgeon": "emphasizes early operative management
        when surgical pathology is suspected.",
"Internal Medicine Physician": "favors medical management
 but recognizes when non-operative therapy is insufficient or unsafe;
 recommends surgery if conservative approach risks deterioration."
}
```

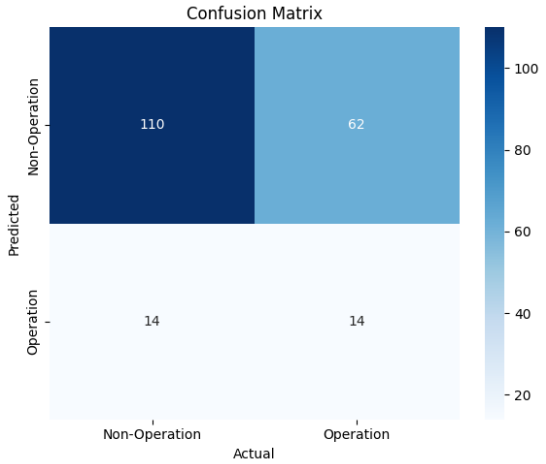
C.

D.



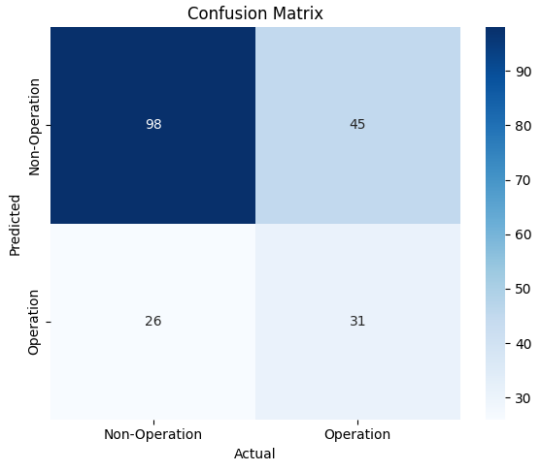
Task 3: Operation Decision Prediction

A. Zero-shot



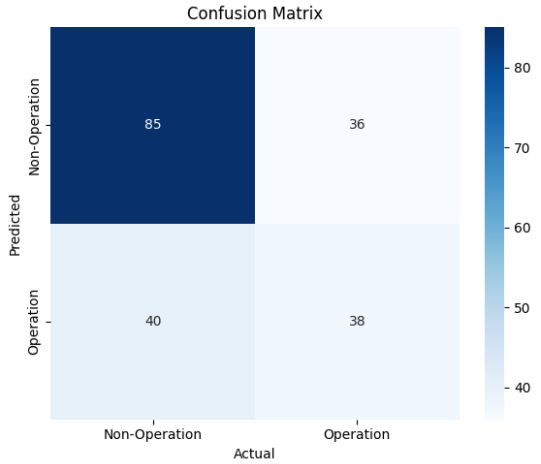
	precision	recall	F1-score
0	0.64	0.89	0.74
1	0.50	0.18	0.27

B. Chain-of-Thought (CoT)



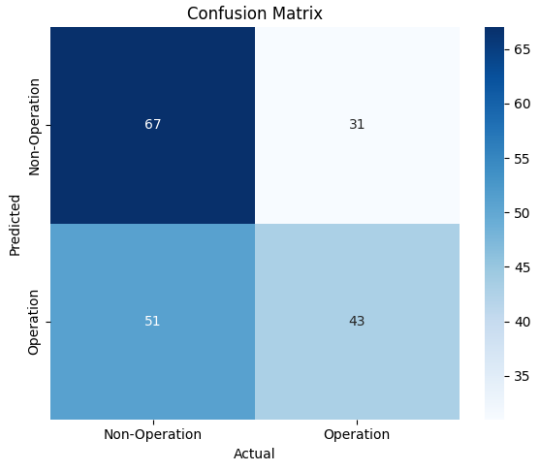
	precision	recall	F1-score
0	0.70	0.80	0.74
1	0.55	0.42	0.48

C. Multi-Agent



	precision	recall	F1-score
0	0.70	0.68	0.69
1	0.49	0.51	0.50

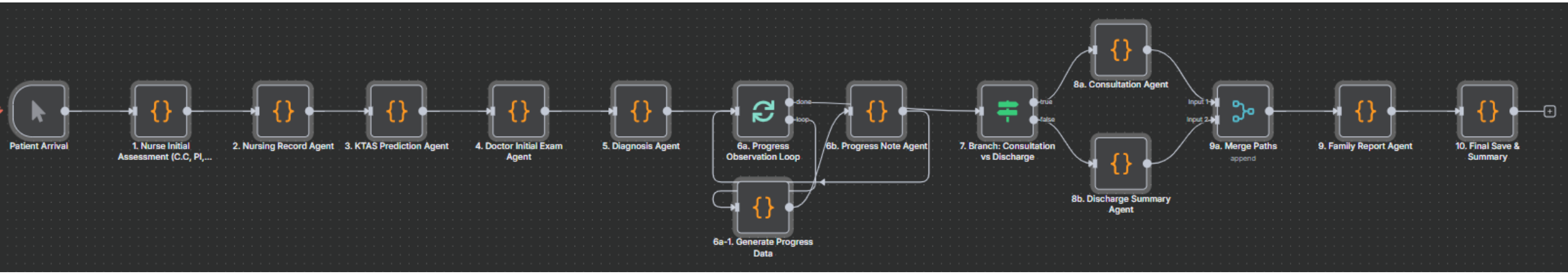
D. Recursive Language Model (RLM)



	precision	recall	F1-score
0	0.68	0.57	0.62
1	0.46	0.58	0.51

- Reasoning evolved from prompt-based → stepwise → collaborative → recursive, progressively emulating real-world clinical decision f low, as reasoning depth and collaboration increased.
- However, several ground-truth *operation* labels were likely overgenerated from discharge-note keyword extraction, leading to cases l abeled as operation even when procedures were only mentioned or planned — explaining some prediction mismatches.

Workflow for Clinical Decision Support System



n8n workflow