

On-Device Parallel AI Triage System

소형 병렬 AI 기반 의료 안내 및 분류 아키텍처 제안서

Author: Cho Hyunwoo

Date: 2025-12-13

Version: v1.0

License: MIT

Abstract

본 문서는 **진단이나 처방을 수행하지 않는 의료 안내(Triage) 전용 AI 시스템**을 제안한다.

본 시스템은 단일 대형 모델 대신 **다수의 소형 전문 AI를 병렬로 운용**하고,

중앙 관리 AI가 결과를 취합하여 **행동 지침(Action Guidance)**만을 제공한다.

이 접근은 다음의 문제를 동시에 해결한다.

- 온디바이스 환경에서의 연산·전력 제약
- 의료 AI의 규제 및 책임 문제
- 병원 접수 및 환자 분류 과정의 병목
- 사용자의 “어디로 가야 하는지 모르는” 구조적 불편

본 제안은 **개인병원, 메디컬 빌딩, 모바일 디바이스**에 즉시 적용 가능한 현실적 아키텍처를 목표로 한다.

1. 문제 정의 (Problem Statement)

1.1 의료 AI가 실패해 온 이유

기존 의료 AI 접근은 다음 가정을 전제로 한다.

- AI가 진단을 내려야 한다
- 정확도는 의료인 수준 이상이어야 한다
- 책임은 시스템 제공자가 져야 한다

이 전제는 현실과 맞지 않는다.

- 진단은 법적·윤리적 책임이 수반된다
- 정확도 요구는 과도하며 비용이 폭증한다
- 의료 현장의 실제 병목은 **진단 이전 단계**에 있다

1.2 실제 병원의 병목 지점

현장에서는 다음 문제가 반복된다.

- 환자는 **자신의 증상은 알지만 어느 과로 가야 할지 모른다**
- 접수처 간호사는 반복적인 안내 업무에 묶인다
- 잘못된 과 접수는 이후 진료 과정에서 수정된다 (전과)

즉, 완벽한 분류가 아니라 '충분히 보수적인 1차 분류'만 있어도
시스템은 정상적으로 작동한다.

2. 제안 개념 (Core Concept)

2.1 진단이 아닌 Triage

본 시스템은 다음을 수행하지 않는다.

- 질병명 확정
- 치료 방법 제안
- 약물 처방

대신 다음만 수행한다.

- 중증도 분류 (위험 / 비위험)
- 방문 시점 판단 (즉시 / 지연 가능)
- 적절한 진료과 안내

이는 병원 접수 시스템 및 안내 키오스크의 역할과 동일하다.

3. 시스템 아키텍처 개요

3.1 병렬 소형 AI 구조

```
[ 사용자 증상 입력 ]  
    ↓  
[ 0.5~0.6B 전문 AI x N ]  
(각 분과 / 역할별)  
    ↓  
[ 7B 중앙 관리 AI ]  
    ↓  
[ 행동 지침 출력 ]
```

3.2 구성 요소

(1) Specialist AI (0.5~0.6B)

- 각 AI는 하나의 분과 또는 역할만 담당
- 예:
 - 호흡기
 - 소화기
 - 심혈관
 - 피부
 - 응급 신호 감지
- 수행 작업:

- 자기 분과 관련 증상인지 판단
- Red Flag 존재 여부 판단
- 자기 분과 아님 판단

(2) Manager AI (7B)

- 전문적 판단 수행 ✖
- 결과 취합 및 설명 수행 ○
- 책임:
 - Specialist 결과 집계
 - 사용자에게 이해 가능한 언어로 안내

4. 판단 로직 (Fail-Safe Design)

4.1 Red Flag 우선 원칙

본 시스템은 다수결을 사용하지 않는다.

- N개의 Specialist 중 1개라도 위험 신호(Red Flag)를 반환하면
 - 즉시 병원 방문 권고

이 로직은 의료·항공·산업 안전 시스템에서 사용되는 보수적 안전 설계(Fail-Safe)와 동일하다.

4.2 판단 예시

| Specialist 결과 | 판단 |
|-----------------------|----------------|
| 1개라도 Red Flag | 즉시 병원 |
| Red Flag 없음, 다수 일반 증상 | 당일 또는 익일 방문 |
| 대부분 “내 분과 아님” | 경과 관찰 또는 일반 진료 |

5. 온디바이스 구현 전략

5.1 왜 대형 모델이 필요 없는가

- 본 시스템은 추론(reasoning)보다 규칙 실행(rule execution)에 가깝다
- 각 Specialist는 제한된 컨텍스트만 필요
- 대형 모델의 장점(창의성, 추론)은 오히려 위험

5.2 Cold Load + Modular Activation

- 기본 상태:
 - 최소 모델만 상주
- 필요 시:
 - 해당 분과 Specialist 모듈 로딩
- 처리 종료 후:
 - 메모리 해제

이 구조는 모바일 환경에서 전력·발열을 최소화한다.

6. 적용 시나리오

6.1 메디컬 빌딩 1층 키오스크

- 증상 입력 → 과 안내
- 틀릴 경우 진료실에서 전과
- 책임 주체: 병원 원장

6.2 개인병원 외래 접수

- 간호사 접수 부담 감소
- 환자 대기 시간 감소

6.3 모바일 디바이스 (On-Device)

- 검색·전화 불필요
- 프라이버시 보호
- 병원 방문 전 사전 판단

7. 규제 및 책임 구조

7.1 의료기기가 아닌 이유

| 항목 | 본 시스템 |
|-------|-------|
| 진단 | × |
| 처방 | × |
| 치료 결정 | × |
| 행정·안내 | ○ |

7.2 책임 주체

- 시스템 제공자: 도구 제공
- 최종 책임: 병원 (사용자)

이는 키오스크, 접수 시스템과 동일한 책임 구조다.

8. 확장 가능성

- 의료 외 영역:
 - 법률 안내
 - 행정 분류
 - 고객 응대
 - 동일 구조 적용 가능:
 - 다수 소형 전문가 + 중앙 관리자
-

9. 결론

본 제안은 의료 AI를 “의사를 대체하는 기술”이 아니라
“병원 문 앞에서 길을 알려주는 기술”로 재정의한다.

이 관점 전환만으로 다음이 가능해진다.

- 소형 모델 사용
- 온디바이스 구현
- 규제 회피
- 즉시 상용화

이 시스템은 완벽하지 않아도 작동한다.

그리고 의료 현장은 이미 그 전제를 받아들이고 운영되고 있다.