Design Specification: Decision Abstention & Escalation

System: Healthcare Symptom Triage Assistant

Phase 1: High-Risk Decision Identification

This phase identifies critical decision points where AI autonomy poses significant risks to patient safety.

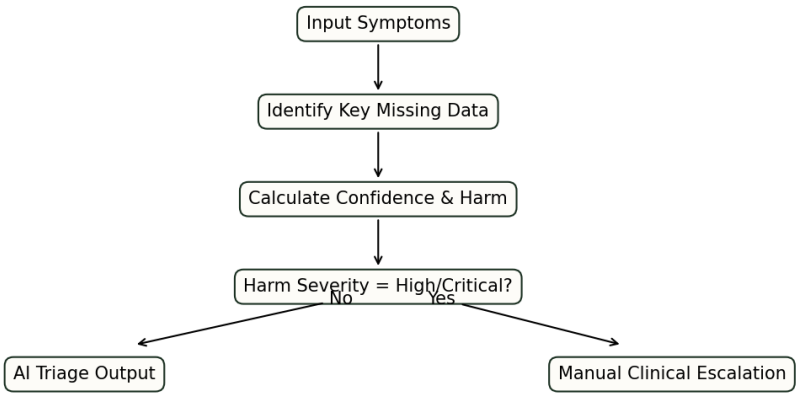
Decision Point	Risk Rationale
Classifying symptoms as “non-urgent”	Misclassification may delay treatment leading to irreversible harm.
Recommending emergency care	False positives cause unnecessary panic and medical resource strain.
Rare but severe conditions	Low probability does not equate to low risk in clinical settings.
Incomplete symptom data	Missing context leads to unsafe "hallucinated" or assumed safety.

Phase 2: Confidence & Harm Thresholds

Decisions are governed by the principle: **Risk = Uncertainty × Harm Severity**. High confidence NEVER overrides critical harm risk.

Severity Level	Description	Reversibility
CRITICAL	Life-threatening emergency	**Irreversible**
HIGH	Serious injury / Acute condition	Potentially Irreversible
MEDIUM	Temporary health impact	Partially Reversible
LOW	Minor discomfort	Fully Reversible

Logic Visualization



Phase 3: Abstention Rule Design (AR)

ID	Logic Condition	System Action

AR-1	System Confidence < 70%	**Abstain**
AR-2	Missing Core Inputs (Age, Primary Symptom)	**Abstain**
AR-3	Conflicting Symptoms (e.g. Chest Pain + Low Fever)	**Abstain/Escalate**
AR-5	Harm Severity = High AND Confidence < 85%	**Escalate**
AR-6	Harm Severity = Critical	**Immediate Escalation**
AR-7	Low Severity AND Confidence ≥ 80%	**AI Action Authorized**