Safety Guardrails Implementation

Constraints that ensure AI systems operate only within safe boundaries

Input Guardrails

- Restrict allowed input range
- Detect malicious prompts
- Automatically remove personal information
- Format validation

Output Guardrails

- Filter harmful content
- Validate medical accuracy
- Specify uncertainty
- Add disclaimers

Action Guardrails

- Execute only permitted actions
- Permission-based access control
- Automatically record audit logs
- Threshold-based alerts

Dynamic Guardrails

- Context-based adjustment
- User role-based restrictions
- Real-time risk assessment
- Adaptive thresholds