

WERKA

Offline Medical Assistant POC

Intake + Triage + Optional Scan Reasoning

Built for constrained care settings

Problem

- Many communities have delayed or limited access to physicians and diagnostics.
- Frontline care workers still need structured triage support.
- Existing assistants are often online-only or not context-aware for local constraints.

Goal: provide actionable, safe-first recommendations from limited inputs.

What WERKA Does

1. Verifies patient identity context.
2. Collects iterative intake (symptoms, duration, optional camera/scans).
3. Uses local text model (Qwen) + triage state to decide MedGemma usage.
4. Produces urgency, immediate actions, and next step.
5. Handles follow-up questions in the same session.

LangGraph Flow

Stage 1: Verification

Verify User -> Identity Mismatch?

Yes -> Verification Follow-up -> Generate Intake + Baseline Triage

No -> Generate Intake + Baseline Triage

Stage 2: Model Routing

Call MedGemma?

Yes -> MedGemma Recommendation -> Control Initial Response

No -> Control Initial Response

Stage 3: Conversation Loop

Follow-up Questions?

Yes -> Control Follow-up Loop -> back to Follow-up Questions?

Model Strategy

- **Qwen (`Qwen/Qwen2.5-3B-Instruct`)**
 - primary local text model for routing and follow-up answers
 - lightweight enough for faster demo runs
- **MedGemma (`google/medgemma-1.5-4b-it`)**
 - optional medical multimodal reasoning/refinement
 - intended source of recommendation refinement when available
- **Safety behavior**
 - hard clinical guardrails override to emergency when red flags detected
 - MedGemma is auto-called for non-routine triage (urgent/emergency), unless disabled

Environment-Aware Triage

- standard
 - default referral behavior
- remote_village
 - nurse-led stabilization, transport coordination, limited imaging assumptions
- limited_access_region
 - cost-aware, staged diagnostics, feasible referral paths

Current POC Scope

Strengths

- End-to-end local workflow orchestration
- Context-aware constraints
- Structured triage outputs for demo usability

Next Steps

- Tighten structured MedGemma JSON output reliability
- Improve follow-up memory/personalization
- Add lightweight UI for nurse/patient workflow