Project Documentation: Offline AI Model with Voice Interaction

Version: 1.0

Last Updated: 2025-05-26

Team: [Your Team Name]

1. Project Overview

Objective

Develop an offline, voice-based AI assistant for students in underserved regions (e.g., Namibia) to provide equitable access to educational resources without internet dependency.

Key Features

- Voice-first interaction (IVR calls + SMS fallback)
- Offline AI model (no cloud API costs)
- Multilingual support (English, Afrikaans, Oshiwambo)
- Low-cost hardware compatibility (Raspberry Pi, basic phones)

2. Technical Stack

Component	Technology	Purpose
Backend Framework	FastAPI + Uvicorn	Handle voice/SMS requests
AI Model	Phi-2 (quantized GGUF)	Offline Q&A generation
Speech-to-Text	Whisper.cpp	Convert voice queries to text
Text-to-Speech	Coqui TTS	Convert AI responses to voice
Database	Firebase/SQLite	Session management
Telecom Integration	Africa's Talking API	SMS/voice call routing

3. Implementation Timeline

Project Documentation: Offline AI Model with Voice Interaction

Phase 1: Research & Planning (Weeks 1-2)

- Benchmark Phi-2 vs. Gemma-2B for offline performance.

- Partner with telecom providers (e.g., MTC Namibia) for IVR integration.

- Gather educational content (curricula, FAQs) for model fine-tuning.

Deliverables: Technical design document, Signed agreements with vendors.

Phase 2: Core AI & Voice POC (Weeks 3-6)

- Build offline pipeline: Whisper.cpp (STT) Phi-2 (NLP) Coqui TTS.

- Test on Raspberry Pi 5 (latency < 2s).

Deliverables: Working POC, Accuracy report vs. Mistral-7B.

Phase 3: Backend & Telecom Integration (Weeks 7-10)

- Extend FastAPI with /voice endpoints for IVR.

- Integrate Africa's Talking API for SMS fallback.

Deliverables: Unified backend, Performance metrics.

Phase 4: Pilot Deployment (Weeks 11-14)

- Deploy to 5 schools in Namibia.

- Train 20+ educators, Collect feedback.

Deliverables: Pilot report, Dialect-optimized Whisper.cpp.

Phase 5: Scaling & Optimization (Weeks 15-20)

- Add Oshiwambo/Afrikaans support.

- Quantize models for 8-bit.

- National rollout planning.

Deliverables: Production system, Stakeholder documentation.

4. Risk Management

Risk	Mitigation Strategy
Poor STT accuracy	Fine-tune Whisper.cpp on local data
High hardware costs	Use Raspberry Pi clusters
Telecom API latency	Cache frequent queries

5. Success Metrics

- Latency < 3s per query
- 95%+ accuracy on Q&A
- 1,000+ pilot users
- 80%+ educator satisfaction

6. Appendices

A. Glossary

- IVR: Interactive Voice Response
- GGUF: Quantized model format

B. References

- Whisper.cpp GitHub: https://github.com/ggerganov/whisper.cpp
- Phi-2 Model Card: https://huggingface.co/microsoft/phi-2

Approval:

[Project Lead Name] | Signature: _____ | Date: 2025-05-26

Next Steps:

1. Share this doc with stakeholders.

2. Kick off Phase 1 (Research & Planning).