WrenAI Offline Feasibility Assessment Report

Date: August 13, 2025
Version Tested: WrenAI 0.27.0 (Latest Available)

Executive Summary:

After extensive testing and troubleshooting, **WrenAI cannot operate completely offline** due to fundamental architectural limitations in the current version. While the LLM component can run locally via Ollama, the embedding system requires external API connectivity, making it unsuitable for fully air-gapped environments.

Technical Findings

1. Embedding Provider Limitations

Investigation Results:

```
# Only available embedding provider in WrenAI 0.27.1:
docker exec -it wrenai-wren-ai-service-1 ls /src/providers/embedder/
__init__.py  __pycache__  litellm.py
```

Critical Issues Discovered:

- **Single Provider Support**: Only litellm_embedder is available
- **API Incompatibility**: Ollama embeddings use /api/embeddings endpoint, but LiteLLM expects OpenAI's /v1/embeddings format
- **Provider Limitations**: Attempted offline providers (huggingface_pipeline, sentence_transformers) return KeyError - not supported in this version

2. Configuration Attempts Made

Attempted Configurations:

1. ✅ **Local Ollama LLM**: Successfully configured with llama3.1:8b
2. ❌ **Ollama Embeddings**: API endpoint incompatibility

3. ❌ **HuggingFace Embeddings**: Provider not supported
4. ❌ **Sentence Transformers**: Provider not supported
5. ✅ **OpenAI Embeddings**: Only working solution (requires internet)

3. Embeddings Dependency Analysis

Critical Architectural Requirement:
Embeddings **cannot be removed or disabled** from WrenAI as they are fundamental to the system's operation.

Core Functionality Dependencies:

- **Database Schema Understanding**: Converting table/column descriptions into vectors for semantic search
- **Question-Answer Matching**: Finding similar historical questions using vector similarity
- **Context Retrieval**: Matching user questions to relevant database schemas
- **Intent Classification**: Understanding what the user is asking for
- **SQL Generation**: Retrieving relevant examples and patterns

Current Operational Status

What Works Offline:

- ✅ Ollama LLM (llama3.1:8b) - Local inference
- ✅ Qdrant Vector Database - Local storage
- ✅ WrenAI UI and Engine Services - Local processing
- ✅ Docker Container Infrastructure - Local deployment

What Requires Internet:

- ❌ **Embedding Generation** - Must use OpenAI API
- ❌ **Vector Similarity Search** - Depends on OpenAI embeddings
- ❌ **Semantic Schema Understanding** - Cannot function without embeddings
- ❌ **Natural Language Processing** - Core NL2SQL functionality disabled

Conclusion

WrenAI 0.27.0 is fundamentally incompatible with offline deployment due to:

1. **Mandatory Embedding Dependency**: Core architecture requires embeddings that cannot be removed
2. **Single Provider Limitation**: Only litellm_embedder supported, which requires OpenAI API
3. **API Incompatibility**: Local embedding solutions (Ollama) use incompatible endpoints
4. **Architectural Design**: Vector-first approach makes offline operation impossible
5. **No Workaround Available**: Current system cannot be modified for offline use
6. **Ongoing Internet Requirement**: Continuous API connectivity needed for operation

Final Assessment: WrenAI offline deployment is **technically impossible** with the current architecture and version limitations.