Baker Group Local Private LLM Implementation

Baker Group Local Private LLM Implementation

AI Inference Solution Architecture & Statement of Work (Tentative)

Client: Baker Group

Date: September 10, 2025

Project: Local Private LLM Solution for Asset/Liability Management

---

Executive Summary

Proactive Technology Management (PTM) proposes a comprehensive **Local Private LLM Inference Solution** for Baker Group, designed to deliver AI-powered document processing, meeting summarization, and key-value extraction capabilities while maintaining strict regulatory compliance for community financial institutions. This solution leverages on-premises inference infrastructure with a modern web interface, ensuring data sovereignty and regulatory adherence.

The proposed architecture implements PTM's **Fusion Development principles**, harmonizing local AI processing (Mind), automated workflows (Body), and structured data management (Soul) within a compliant, air-gapped environment. This approach delivers measurable ROI through process automation while meeting the stringent regulatory requirements of the Asset/Liability Management sector.

Problem Statement & Business Drivers

Baker Group, as a leader in Asset/Liability Management for Community Financial Institutions, faces critical challenges in maintaining regulatory compliance while leveraging modern AI capabilities:

Regulatory Compliance Requirements

- Financial institutions must maintain data sovereignty and prevent sensitive information from leaving their infrastructure
- Strict audit trails and governance controls required for all AI-driven processes
- Need for transparent, explainable AI outputs for regulatory reporting

Operational Efficiency Drivers

- Manual document processing and meeting summarization consuming significant staff time
- Inconsistent key-value extraction from financial documents impacting decision-making speed
- Need for standardized AI-powered workflows while maintaining human oversight capabilities

Strategic Objectives

- Implement AI capabilities without compromising regulatory compliance
- Reduce manual processing time by 60-80% for document-intensive workflows
- Establish foundation for expanded AI automation across Baker Group's service offerings

---

Solution Architecture Overview

Core Model Stack for GTX 5090 (32GB VRAM)

Primary Text Processing

- **gpt-oss-20b**: Core text processing, reasoning, and transcript analysis via

Ollama

- **Resource Allocation**: ~14GB VRAM

Multimodal Document Processing

- **llama3.2-vision-11b**: Meta's multimodal model for PDF document analysis and visual reasoning
- **Alternative**: **gemma3 (vision variant)**: Google's multimodal model for visual document processing
- **Resource Allocation**: ~12GB VRAM

Speech Processing Architecture

- **Device-Based Audio Capture**: iPhone with Siri audio services integration
- **Just Press Record App**: Professional audio recording with cloud sync capabilities
- **Audio File Transfer**: Secure file transfer to local Windows inference server

- **Available VRAM Buffer**: ~6GB for inference operations and model switching

System Context - C4 Level 1

The solution implements a **local inference architecture** centered around Ollama serving specialized models, integrated with a modern web interface and iOS device audio capture:

Core Components:

- **Ollama Inference Server**: Local LLM hosting on Windows with GTX 5090 (32GB VRAM)
- **Web Interface**: React/TypeScript SPA via lovable.dev, hosted on IIS
- **PydanticAI Agent Orchestra**: Structured AI agents leveraging the two core models
- **iOS Audio Integration**: iPhone-based recording via Just Press Record app

- **Output Validation Layer**: Pydantic-enforced structured outputs

Container Architecture - C4 Level 2

1. Model Inference Layer

text

```
Ollama Server Stack (32GB VRAM):├── gpt-oss-20b (Text LLM) - Transcript processing and reasoning└── llama3.2-vision-11b (Multimodal) - PDF/Document visual analysis
```

2. iOS Audio Integration Layer

- **Just Press Record App**: Professional iPhone audio recording
- **Audio File Sync**: Secure transfer mechanism (AirDrop, or direct upload)
- **Audio Processing Pipeline**: Server-based audio file ingestion and routing to gpt-oss-20b
- **Mobile-Optimized Interface**: Responsive web interface for iOS Safari

3. Document Processing Pipeline

- **PDF Ingestion**: PyMuPDF + visual processing via llama3.2-vision
- **Image Preprocessing**: Base64 encoding for visual elements
- **Layout Preservation**: Maintains document structure and formatting context
- **OCR Integration**: Enhanced text extraction from scanned documents

4. Agent Orchestration Layer

python

```
class MeetingSummarizationAgent:    """Processes audio transcripts using gpt-oss-20b"""    model: "gpt-oss-20b"    input_schema: AudioTranscriptInput    output_schema: StructuredMeetingSummaryclass DocumentAnalysisAgent:    """Processes visual documents using llama3.2-vision"""    model: "llama3.2-vision-11b"    input_formats: [PDF, Images, Scanned_Documents]    output_schema: FinancialDocumentExtractionclass ComplianceValidationAgent:    """Validates outputs for regulatory compliance"""    text_model: "gpt-oss-20b"    vision_model: "llama3.2-vision-11b"    validation_frameworks: [SOX, FFIEC, NCUA]
```

Technical Implementation Specifications

Core Technology Stack

Infrastructure

- **OS**: Windows Server or Windows 11 Pro
- **GPU**: NVIDIA GTX 5090 with 32GB VRAM and CUDA support
- **Memory**: 64GB RAM for optimal multimodal processing
- **Storage**: 2TB NVMe SSD for model storage and data processing

AI & Inference

- **LLM Runtime**: Ollama 0.3+ with gpt-oss-20b and llama3.2-vision-11b models
- **Agent Framework**: PydanticAI for type-safe agent orchestration
- **Model Configuration**: Optimized temperature and context settings for financial tasks

Frontend & Web Server

- **Development Framework**: lovable.dev (TypeScript/Vite/React)
- **Web Server**: IIS 10+ with SSL/TLS encryption
- **Mobile Optimization**: iOS Safari-optimized responsive interface

iOS Integration

- **Recording App**: Just Press Record for professional audio capture
- **File Transfer**: Multiple secure transfer options (AirDrop, iCloud, direct upload)
- **Audio Formats**: M4A, WAV, MP3, AAC support

Security & Compliance

- **Network Isolation**: Air-gapped or restricted network access
- **Data Encryption**: AES-256 encryption for data at rest
- **Access Control**: Role-based authentication and authorization
- **Audit Trail**: Comprehensive logging for regulatory compliance

Model Resource Management

32GB VRAM Allocation

- **gpt-oss-20b**: ~14GB VRAM (optimized for transcript processing)
- **llama3.2-vision-11b**: ~12GB VRAM (document visual analysis)
- **Processing Buffer**: ~6GB reserved for inference operations

Model Loading Strategy

- **Hot Loading**: Both models can remain loaded simultaneously
- **Instant Switching**: No model swap delays between transcript and document processing
- **Optimized Inference**: Parallel processing capabilities for complex workflows

Specialized Processing Capabilities

Transcript Processing with gpt-oss-20b

- **Meeting Summarization**: Structured summaries with key decisions and action items
- **Financial Context Understanding**: Asset/Liability Management terminology expertise
- **Compliance Annotation**: Regulatory requirement identification and flagging
- **Audio File Integration**: Direct processing of iPhone-recorded audio files

Document Analysis with llama3.2-vision-11b

- **Financial Table Extraction**: Automated detection and structured data extraction
- **Chart Recognition**: Understanding of financial charts, graphs, and visual data
- **Layout Analysis**: Multi-column layouts, headers, and footnote processing

- **Signature Detection**: Processing of authorization sections and compliance signatures

---

Implementation Timeline - 8 Weeks

Phase 1: Infrastructure & Model Deployment (Weeks 1-2)

Milestone 1.1: Hardware Setup & Model Installation

- Windows server configuration and GPU optimization
- Ollama installation with gpt-oss-20b and llama3.2-vision-11b deployment
- Performance benchmarking and 32GB VRAM optimization
- **Deliverable**: Dual-model inference system operational

Milestone 1.2: iOS Integration Framework

- Just Press Record workflow design and testing
- Secure audio file transfer mechanisms setup
- Server-based audio file ingestion pipeline
- **Deliverable**: Complete iOS-to-server audio workflow

Phase 2: Web Interface & Processing Pipelines (Weeks 3-4)

Milestone 2.1: Mobile-Optimized Web Application

- lovable.dev React application with iOS Safari optimization
- Audio file upload interface with multiple format support
- PDF document upload with preview capabilities
- **Deliverable**: Functional web interface with multimodal inputs

Milestone 2.2: Core Processing Integration

- gpt-oss-20b integration for transcript processing
- llama3.2-vision-11b integration for document analysis
- Combined workflow processing capabilities
- **Deliverable**: Complete text and visual processing system

Phase 3: Agent Implementation & Workflows (Weeks 5-6)

Milestone 3.1: Specialized AI Agents

- Meeting Summarization Agent with gpt-oss-20b
- Document Analysis Agent with llama3.2-vision-11b
- Compliance Validation Agent using both models
- **Deliverable**: Full agent suite with structured outputs

Milestone 3.2: Workflow Integration

- Audio-to-transcript-to-summary workflows
- PDF-to-structured-data extraction workflows
- Compliance validation and audit trail generation
- **Deliverable**: Production-ready processing workflows

Phase 4: Testing, Security & Deployment (Weeks 7-8)

Milestone 4.1: Integration Testing & Security Assessment

- End-to-end testing with real financial documents and audio
- iPhone integration testing and optimization
- Security penetration testing and vulnerability assessment
- **Deliverable**: Security-validated, performance-optimized system

Milestone 4.2: Training & Production Go-Live

- Baker Group staff training on all capabilities
- Just Press Record app setup and user guides
- Production deployment with monitoring and support
- **Deliverable**: Live production system with full capabilities

---

Success Metrics & ROI Framework

Key Performance Indicators

Processing Efficiency Metrics

- **Audio Processing Speed**: <2 minutes from iPhone upload to structured summary
- **Document Analysis Time**: <30 seconds for complex multi-page financial reports
- **Accuracy Targets**: >95% for document extraction, >90% for meeting summarization
- **System Uptime**: >99.5% availability

Business Impact Measurements

- **Staff Time Savings**: 30-40 hours per week across document processing workflows
- **Error Reduction**: >85% decrease in manual data entry errors
- **Decision Speed**: 50% faster turnaround on financial analysis tasks
- **Compliance Efficiency**: >90% automated compliance flag identification

Expected ROI Timeline

- **Month 1-2**: Break-even through reduced manual processing time
- **Month 3-6**: Positive ROI through improved operational efficiency
- **Month 6-12**: Full ROI realization through expanded capabilities and error reduction

---

Risk Management & Compliance Framework

Security & Regulatory Controls

Data Protection

- All processing occurs on-premises with no external data transmission
- AES-256 encryption for data at rest and in transit
- Role-based access control with multi-factor authentication
- Regular security assessments and compliance audits

Audit & Governance

- Comprehensive logging of all AI interactions and decisions
- Immutable audit trails for regulatory reporting
- Version control for all prompts and model configurations
- Quarterly compliance reviews and system updates

Risk Mitigation Strategies

- Human oversight required for high-stakes financial decisions
- Confidence scoring and uncertainty flagging for all outputs
- Fallback procedures for system failures or maintenance
- Regular model performance monitoring and validation

Investment Summary

Professional Services Investment

- **Strategy & Architecture**: 32-48 hours senior consulting
- **Implementation & Development**: 120-160 hours technical delivery
- **iOS Integration & Testing**: 24-32 hours mobile optimization
- **Training & Change Management**: 16-24 hours organizational support

Technology & Infrastructure

- **Hardware**: Client-provided GTX 5090 system with 32GB VRAM
- **Software Licensing**: Included in development scope
- **iOS App**: Just Press Record (one-time purchase per device)
- **Ongoing Support**: monthly maintenance contract

Timeline & Deliverables

- **Total Duration**: 8 weeks from project initiation
- **Go-Live Date**: Week 8 with full production capability
- **Post-Launch Support**: 30-day warranty period with ongoing support options

Next Steps & Engagement Process

1. **Hardware Validation**: Confirm GTX 5090 system meets 32GB VRAM specifications
2. **iOS Device Assessment**: Validate iPhone compatibility and Just Press Record app deployment
3. **Security Requirements Review**: Finalize compliance requirements and audit procedures
4. **Statement of Work Approval**: Detailed project scope and timeline confirmation
5. **Project Kickoff**: Immediate initiation of 8-week implementation schedule

This comprehensive solution delivers advanced AI capabilities through local inference processing while maintaining the strict regulatory compliance essential for community financial institutions. The combination of powerful on-premises processing with intuitive iPhone integration provides Baker Group with a competitive advantage in Asset/Liability Management services, delivering measurable efficiency gains and enhanced decision-making capabilities.

Proactive Technology Management's Fusion Development methodology ensures rapid time-to-value while building sustainable, scalable AI capabilities that respect and enhance regulatory compliance requirements. This solution positions Baker Group at the forefront of AI-enabled financial services while maintaining the security and compliance standards critical to their industry leadership.