3-Minute Demo Plan: AI Live Call Insights Solution

Opening Hook (15 seconds)

"Customer support agents receive over 50 calls daily but struggle to access relevant information instantly. Our AI-powered solution transforms live conversations into actionable insights in real-time."

Problem Statement (20 seconds)

- **The Challenge**: Support agents need instant access to relevant product information during live calls
- **Current Pain Points**:
  - Agents put customers on hold to search for information
  - Inconsistent responses across different agents
  - Missed opportunities to provide proactive suggestions
- **Our Solution**: Real-time AI suggestions powered by speech-to-text and RAG (Retrieval-Augmented Generation)

Architecture Overview (30 seconds)

Tech Stack Highlights

- **Frontend**: React with real-time WebSocket connections
- **Backend**: Node.js with AWS services integration
- **AI Pipeline**:
  - AWS Transcribe for speech-to-text
  - Claude Sonnet 4 for intelligent suggestion generation
  - Amazon Bedrock embeddings for semantic search
  - Local RAG system for instant knowledge retrieval

Key Innovation

- **Two-Stage AI Processing**:
  1. **Smart Trigger Detection**: AI evaluates if suggestion is needed
  2. **Contextual Suggestion Generation**: RAG-powered recommendations

Live Demo (75 seconds)

Demo Scenario: Customer Support Call

Setup: "Let me show you how our solution works during a live customer call about SLA issues"

Demo Flow:

1. **Start Call Simulation** (15 seconds)
   - Show real-time transcription appearing
   - Display speaker identification (Customer vs Agent)
   - Show conversation summary updating live
2. **Customer Pain Point** (20 seconds)
   - Customer says: "*Your tool completely missed 2 SLAs last month. We lost a huge contract.*"
   - **Highlight**: AI instantly detects this needs attention
   - **Show**: LLM evaluation JSON response in real-time
3. **AI Suggestion Generation** (25 seconds)
   - **Demonstrate**: RAG system searching knowledge base
   - **Show**: Suggestion card appears instantly:

     ```
     Title: Calm SLA Escalation Response
     Content:
     • Acknowledge SLA breach, no deflection
     • Use steady tone: "I get how serious this is"
     • Offer SLA report review + escalation path
     ```

4. **Agent Response** (15 seconds)
   - Show how agent uses the suggestion
   - Display conversation continuing with improved response

Key Features to Highlight:

- **Real-time Processing**: No delays in suggestion generation
- **Context-Aware**: Uses full conversation history
- **Actionable Insights**: Specific, implementable suggestions
- **Scalable**: Works with any knowledge base

Unique Value Propositions (30 seconds)

What Makes Us Different:

1. **Intelligent Filtering**: Only shows suggestions when truly needed (not overwhelming)
2. **Conversation Context**: Uses full conversation history, not just current message
3. **Dual AI Pipeline**: Evaluation + Generation for optimal relevance
4. **Real-time RAG**: Instant knowledge base search during live calls
5. **Speaker Intelligence**: Automatically identifies customer vs agent roles

Business Impact:

- **Reduced Call Times**: Agents get instant access to relevant information
- **Consistent Quality**: Every agent has access to the same knowledge base
- **Improved CSAT**: Customers receive accurate, contextual responses
- **Scalable Training**: New agents perform like experienced ones

Technical Innovation Highlights (20 seconds)

Advanced Features:

- **Semantic Search**: Vector embeddings for intelligent information retrieval
- **Multi-turn Awareness**: Understands conversation flow and context
- **Adaptive Suggestions**: Different suggestion types based on conversation stage
- **Real-time Processing**: Sub-second response times
- **AWS Integration**: Leverages AWS Transcribe, Bedrock, and Claude

Closing & Call to Action (10 seconds)

"Our solution transforms every support call into an opportunity for exceptional customer experience. We're not just transcribing speech—we're creating intelligent, context-aware assistance that scales human expertise."

Demo Preparation Checklist

Before Demo:

☐ Prepare sample audio files for different scenarios
☐ Ensure knowledge base is loaded with relevant content
☐ Test WebSocket connections
☐ Prepare backup slides in case of technical issues
☐ Practice timing for each section

Backup Scenarios:

- **Technical Issues**: Have recorded demo video ready
- **Network Problems**: Offline slides with screenshots
- **Audio Issues**: Text-based demo with manual input

Demo Best Practices:

1. **Start with Impact**: Lead with business value, not technical details
2. **Show, Don't Tell**: Live demo is more powerful than slides
3. **Handle Edge Cases**: Prepare for questions about accuracy, latency

4. **Emphasize Scalability**: Show how it works across different industries
5. **End with Vision**: Paint picture of future possibilities

---

Judging Criteria Alignment

Innovation (25%)

- Novel two-stage AI pipeline
- Real-time RAG implementation
- Intelligent suggestion filtering

Technical Implementation (25%)

- Production-ready architecture
- AWS services integration
- Scalable WebSocket design

Business Value (25%)

- Clear ROI for customer support
- Measurable impact on call efficiency
- Scalable across industries

Presentation (25%)

- Clear problem articulation
- Compelling live demo
- Professional delivery

Potential Q&A Preparation

Q: How accurate is the speech-to-text? A: We use AWS Transcribe with speaker diarization, achieving 95%+ accuracy for clear audio. The system handles multiple speakers and accents.

Q: What's the latency for suggestions? A: Sub-second response time due to precomputed embeddings and efficient RAG pipeline.

Q: How does it handle different industries? A: Modular knowledge base design allows easy customization for any domain - just replace the vector database.

Q: Security and compliance? A: All data processing uses AWS services with enterprise-grade security. No conversation data is stored permanently.

Q: Scalability? A: WebSocket architecture supports concurrent connections, and AWS services auto-scale based on demand.

Success Metrics to Mention

- **Response Time**: < 1 second for suggestion generation
- **Accuracy**: 90%+ relevant suggestions based on context
- **Scalability**: Handles 100+ concurrent calls
- **Integration**: Works with existing call center infrastructure