Design Brief for Claude: InkLink Local AI Assistant Ecosystem (v3 - Limitless Enhanced)

Project: Augmenting InkLink with Always-On Local AI Assistants, including Personalized Models from Limitless Pendant Data
Target User: Senior Engineering Manager using ProtonMail & Proton Calendar, and Limitless Pendant
Primary Cloud AI: Claude Max (for OCR & high-level orchestration)
Local AI Platform: Ollama on User's Ubuntu Desktop
Communication: Multi-Connection Protocol (MCP)
Output Interface: reMarkable Tablet via dynamically generated .rm templates

1. Introduction & Vision (Updated)

This document outlines the development plan for extending the **remarkable-ink-link (InkLink)** project with a sophisticated system of **always-on, local AI assistant agents**. These agents will run on the user's powerful Ubuntu desktop, leveraging local Large Language Models (LLMs) via Ollama.

Core Vision (Extended):

- **Claude Max** remains the primary engine for high-fidelity handwriting recognition and high-level orchestration.
- A suite of **specialized local AI agents** operate asynchronously, performing tasks like data pre/post-processing, proactive knowledge discovery, and PIM integration (ProtonMail/Calendar).
- A **hyper-personalized local AI model, fine-tuned on the user's Limitless Pendant audio transcripts**, will provide an unparalleled understanding of the user's speech, context, habits, and knowledge. This model will work in concert with other local agents and Claude Max.
- Communication occurs via **MCP**.
- Outputs are delivered via **reMarkable templates**.

Key Benefits (Extended):

- **Unprecedented Personalization:** The Limitless-trained model will offer insights and assistance based on the user's actual spoken words and daily context.
- Enhanced Privacy: All data, including sensitive Limitless transcripts and the fine-tuned model, remains entirely local.
- Contextual Ambient Intelligence: The system gains a "memory" of the user's spoken life, enabling more relevant and proactive assistance.

2. Core System Components (Limitless Data

Integration)

- **Ubuntu Hub:**
    - **Limitless Data Ingestion Pipeline:** Securely ingests and stores Limitless Pendant transcripts locally (building on existing LimitlessLifeLogService). Data needs to be preprocessed for fine-tuning.
    - **Local Model Fine-Tuning Environment:** Tools and scripts for preparing Limitless data and fine-tuning an Ollama-compatible LLM.
    - **Personalized LLM (Served by Ollama):** The fine-tuned model itself, accessible to local agents.
- **Local AI Agents (Python):** Can now query the personalized Limitless-trained LLM for specific tasks.
- *(Other components like Claude Max, reMarkable, MCP Router remain as previously defined)*

3. Detailed Design Tasks for Claude's Input (Extended)

3.1. - 3.3. (Local Agent Framework, MCP, reMarkable Templates - As per v2 Brief)

(No changes to these sections from the previous brief, Claude should refer to v2 for these details)

3.4. ProtonMail & Proton Calendar Integration (As per v2 Brief)

(No changes to this section from the previous brief, Claude should refer to v2 for these details)

NEW SECTION: 3.5. Limitless Pendant Data & Personalized Model Fine-Tuning

- **Objective:** Design a pipeline for securely ingesting Limitless Pendant transcripts, preparing them for fine-tuning, training a personalized local LLM, and integrating this model into the local agent ecosystem.
- **Tasks for Claude's Input:**
    1. **Limitless Data Ingestion & Preprocessing Strategy:**
        - How should raw transcripts from the Limitless Pendant (presumably accessed via LimitlessLifeLogService and LimitlessAdapter) be securely stored locally?
        - What preprocessing steps are needed for fine-tuning? (e.g., cleaning, speaker diarization if available/needed, formatting into instruction/completion pairs or conversational format, PII scrubbing if desired even for local use).
        - How to handle the continuous nature of the data for periodic re-tuning?

2. **Local LLM Fine-Tuning Workflow:**
   - **Model Selection for Fine-Tuning:** Recommend base Ollama models suitable for fine-tuning with conversational text data (e.g., llama3:8b, mistral:7b, phi3:mini/medium). Consider models known for good instruction-following if creating a task-oriented personalized assistant.
   - **Fine-Tuning Tools & Techniques:**
     - Outline a process using local tools (e.g., ollama create with a custom Modelfile incorporating training data, or Python libraries like transformers with PEFT/LoRA, Axolotl, llama.cpp training scripts) compatible with the user's RTX 4090.
     - Suggest data formatting for fine-tuning (e.g., question-answer pairs, instruction-response, summarization tasks based on transcripts).
     - Advise on managing training datasets derived from Limitless transcripts.
3. **Personalized Model Deployment & Access:**
   - Once fine-tuned, how will this new personalized model (e.g., ollama serve my-limitless-llm) be managed and accessed by other local agents via the OllamaAdapter?
   - How to version and update the personalized model as more Limitless data becomes available?
4. **"Helpful Purposes" - Use Cases for the Personalized Model:**
   - **Personalized Summarization:** "Summarize my key discussions from yesterday based on my Limitless data."
   - **Action Item Recall:** "What action items did I mention I needed to do when talking to Sarah on Tuesday?"
   - **Contextual Reminders:** If the user said "I need to remember to buy milk after my 10 AM meeting," the agent could generate a reminder.
   - **Personal Q&A:** "What were my main concerns about Project X when I discussed it last week?"
   - **Spoken Thought Linking:** Connect spoken ideas from Limitless data to written notes on the reMarkable or content in ProtonMail/Calendar.
   - **Early Idea Capture:** Transcribe and flag potential ideas or tasks mentioned verbally before they're formally written down.
   - **Personalized Content Generation:** Draft emails or notes in the user's typical speaking style.
5. **Privacy & Security of Limitless Data and Fine-Tuned Model:**
   - Reiterate that all Limitless transcripts, derived training data, and the fine-tuned model remain strictly on the user's local machine.
   - Consider if any additional local PII scrubbing is desired *before* fine-tuning, even for a local model.

3.6. Agent Orchestration & Data Flow Examples (Updated for

Limitless)

- Please update one or two data flow examples (e.g., "Daily Briefing") to show how the **Personalized Limitless LLM** could be queried by other local agents to provide richer, more personalized context.
    - **Example:** The DailyBriefingAgent could query the PersonalizedLimitlessLLM with: "Based on my spoken conversations in the last 24 hours, are there any emergent tasks or concerns I mentioned related to today's meeting with 'Project Phoenix Team'?"

3.7. Error Handling, Resilience, and Monitoring (As per v2 Brief, with additions for fine-tuning pipeline)

- Add considerations for monitoring the fine-tuning pipeline and validating the performance of the personalized model.

4. Initial Agent Ideas (Consider interaction with Personalized Limitless LLM)

1. **Daily Briefing & Agenda Prep Agent:** (Can now query Personalized Limitless LLM for spoken context related to meetings/tasks).
2. **Continuous Knowledge Curator & Synthesizer Agent:** (Can link written notes with spoken concepts from Limitless data via embeddings and the Personalized LLM).
3. **Proactive Project & Goal Tracker Agent:** (Can pick up on spoken commitments or progress updates from Limitless data).
4. **NEW: LimitlessContextualInsightAgent:**
    - **Functionality:** Specifically designed to interact with the fine-tuned Limitless model.
    - Provides MCP services like get_spoken_summary(time_period), recall_spoken_action_items(keywords, time_period), find_spoken_context(topic).
    - Other agents would query this agent to get insights derived from the Limitless data.

5. Deliverables Expected from Claude (Updated)

1. Architectural diagrams (incorporating Limitless data pipeline and personalized model).
2. Class structures (LocalAgent, OllamaAdapter).
3. MCP message schemas (including for Limitless-derived insights).
4. reMarkable template specifications.
5. Sequence diagrams for Proton and **Limitless integration workflows**.
6. Strategies for error handling, state persistence, security, and **Limitless data management/fine-tuning pipeline monitoring**.
7. Ollama model recommendations (base models for agents, and **base models + strategies for fine-tuning with Limitless data**).

This updated brief directs the design effort to fully incorporate the powerful potential of the user's Limitless Pendant data, creating a truly unique and deeply personalized AI assistant ecosystem within InkLink.