QuietRoom: privacy-first personal journaling companion with on-device AI processing

Francesco Laiti, a first-year PhD Student from Italy!

As users increasingly turn to large language models for emotional support, the privacy of their most sensitive data is placed at significant risk. This paper introduces QuietRoom, a direct response to this challenge: a privacy-first, multimodal journaling application designed to provide a secure, on-device sanctuary for mental well-being. Leveraging the unique and strong capabilities of Google's Gemma 3n on-device models, QuietRoom offers users structured reflection and proactive analysis without their personal data ever leaving their control. This document details the application's flexible architecture, the specific implementation of Gemma 3n for intelligent insight generation, and the engineering solutions that make this privacy-centric approach possible. It serves as the technical verification for a functional proof-of-concept, demonstrating the power of building impactful, private AI experiences. The code and a live demo are publicly available at `https://huggingface.co/spaces/laitifranz/QuietRoom`.

1. Introduction

The intersection of artificial intelligence and mental wellness has created a new generation of powerful tools for personal development. While traditional practices like journaling are proven methods for self-reflection (Pausa Team, 2025), their modern digital counterparts often introduce a critical vulnerability: user privacy.

This problem is brought into sharp focus by recent large-scale research from Anthropic released on June 27th 2025, a day after the start of the Gemma3n Hackathon (Google DeepMind / Kaggle, 2025). Their analysis of over 4.5 million user conversations revealed that approximately **2.9%** are classified as "affective conversations", interactions driven by deep-seated emotional or psychological needs, mirroring the intimate act of private journaling yet occurring on platforms not built for confidentiality (McCain et al., 2025). This statistic unveils a critical privacy paradox: users are entrusting their most fragile and personal thoughts to cloud-based AI systems not explicitly designed as secure, confidential environments. This widespread sharing of sensitive information poses a significant ethical and privacy challenge that demands a new, privacy-first approach.

This challenge was the direct catalyst for **QuietRoom**. As a solo developer from Italy passionate about building human-centric technology, I was driven to create a solution that offered the benefits of AI companionship without compromise. The mission was to engineer a true digital sanctuary, a 'quiet room', where users could reflect openly, powered by an AI that runs entirely within their personal environment without using third-party services or online storage.

This paper introduces **QuietRoom**, a novel journaling application that directly addresses these limitations and starts the exploration of on-device AI for journaling tasks. It combines a rich, multimodal interface with a privacy-by-design philosophy, powered by Google's on-device Gemma 3n models family (Gemma Team (Google DeepMind), 2025). The following sections will detail QuietRoom's technical architecture, its specific implementation of Gemma 3n for intelligent analysis, and the engineering solutions that make it a compelling proof-of-concept for the future of private AI.

2. Gemma 3n: The Architectural Cornerstone for Private Mental Wellness

QuietRoom's mission to be a secure sanctuary for self-reflection is built upon one technical principle: the AI must run entirely on the user's device. Google's Gemma 3n was chosen as the architectural cornerstone because it uniquely satisfies the project's three non-negotiable requirements:

1. **Unyielding Privacy.** Mental wellness data is exceptionally sensitive. By processing multimodal (text, audio, images) locally, Gemma 3n provides a technical guarantee that a user's private thoughts never leave their device. This eliminates the security risks of cloud-based APIs and builds a foundation of absolute trust;
2. **Ubiquitous & Uninterrupted Support.** A moment of reflection is not bound by internet availability. Gemma 3n's offline capability ensures QuietRoom is a reliable companion anywhere, anytime, e.g. on a plane or during an outage. This is critical for the continuity of emotional support, which cannot depend on a stable connection;
3. **Cost-Free & Uninhibited Reflection.** Cloud API costs create a psychological barrier that discourages the frequent journaling essential for therapeutic benefit. Gemma 3n's local inference removes this barrier entirely, enabling limitless, uninhibited interaction at no marginal cost;
4. **Improved Multilingual Capabilities.** Gemma3n family models features strong performance across multiple languages, suitable for large-scale deployment over the world population. This feature helps to low the barrier of communication and strength the quality and clarity of the model responses.

3. System Architecture

QuietRoom is architected as a modern client-server web application to ensure cross-platform accessibility and facilitate a seamless live demonstration. The design strictly separates the user interface (client) from the data processing and AI logic (server). This ensures that all sensitive user data is contained and processed within a secure, local server environment, never exposed directly to the internet. The overall system architecture is illustrated in Figure 1.

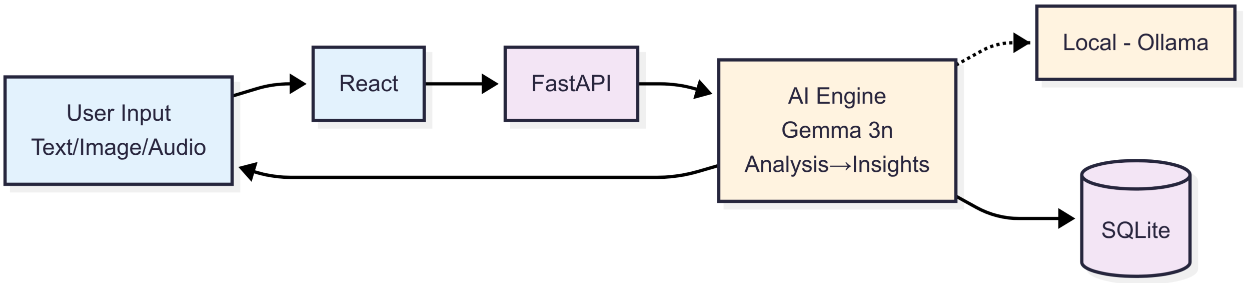

Figure 1 | The high-level system architecture of QuietRoom, illustrating the flow of data from the user through the frontend, backend, and private AI model.

3.1. Component Overview

3.1.1. Frontend (Client-Side)

The frontend is a responsive single-page application (SPA) built with **React 19 and TypeScript**, and styled with **Tailwind CSS** for a clean and calming user experience. Its primary responsibility is to provide the user interface for all interactions. It manages client-side state and communicates with the backend via RESTful APIs. Key interface modules include:

- **A multimodal journal editor** for capturing text, images, and audio with clear privacy indicators;
- **An interactive timeline and calendar view** for browse and filtering past entries;
- **A conversational chat interface** for real-time interaction with the AI companion;
- **A settings panel** for configuring the application, add personal information or cultural context, and AI provider.

3.1.2. Backend (Server-Side)

The backend serves as the intelligent orchestration layer, implemented in **Python 3.12** with the **FastAPI** framework. Its asynchronous capabilities are ideal for handling I/O-bound operations like API requests and AI model inference without blocking. The backend is responsible for:

- **API Endpoints:** Serving data to and receiving data from the React frontend;
- **Data Persistence:** Managing all user data via a strictly local storage solution;
- **AI Coordination:** Interfacing with the Gemma 3n model to generate insights, summaries, tagging, detection, and conversational replies.

3.1.3. Database and Data Storage

To uphold the privacy-first guarantee, all user data, e.g. journal entries, media files, conversation history, and metadata, is stored in a **local SQLite database**. Database interactions are managed through **SQLAlchemy** ORM (Bayer, 2025), providing robust data modeling and query capabilities. The storage architecture ensures:

- **Complete Local Control:** No user-generated content is ever stored outside the local database file;
- **Encryption Ready:** Database structure designed to support future encryption implementations;
- **Efficient Retrieval:** Optimized indexing for fast search and analysis across journal entries;

4. The AI Engine: Privacy-First Intelligence with Gemma 3n

The core of QuietRoom's intelligence is its AI engine, designed around two fundamental principles: user privacy and architectural flexibility. This section details the technical choices that power the AI and the specific capabilities enabled by Gemma 3n.

4.1. Architectural Design: On-Device Ready

A foundational architectural decision was to use the **LiteLLM** (LiteLLM Team / BerriAI, 2025) framework as a standardized interface for all communication with AI models. This choice was critical as it decouples the application logic from the AI model provider, creating a "privacy-ready" system that can seamlessly switch between a cloud API and a local, on-device model with zero code changes.

During development, this flexibility allowed me for rapid prototyping and validation of the entire AI pipeline using the `gemini/gemma-3n-e4b-it` and `gemini/gemma-3n-e2b-it` model via Google's API (Google AI Studio, 2025) [1]. While my personal 8-year-old notebook posed a performance bottleneck for local inference, the successful API tests serve as a complete proof-of-concept for the architecture. It confirms that the system is fully engineered and prepared for high-performance, on-device deployment via Ollama on any capable consumer hardware, fulfilling the project's privacy-first mission.

4.2. Gemma 3n Specialization: From General Model to Therapeutic Assistant

QuietRoom transforms Gemma 3n from a general-purpose language model into a specialized therapeutic assistant through precise prompt engineering and structured interaction design. Rather than implementing a simple chat interface, the system employs task-specific prompts designed to perform distinct therapeutic functions with reliable, structured outputs.

4.2.1. Core Therapeutic Capabilities

The Gemma 3n integration provides several specialized therapeutic functions, taking into consideration optional personal information and cultural nuances of the user:

1. **Mood Analysis and Scoring:** Analyzing journal text to provide mood scores (1-10 scale), identify primary emotions, and extract psychological themes with high clinical correlation, and generate follow-up questions for the user to think about;
2. **Crisis Detection:** Implementing safety-critical analysis to identify potential emotional distress, self-harm indicators, or crisis situations requiring immediate support;
3. **Therapeutic Conversation:** Generating empathetic, contextually-aware responses that employ cognitive behavioral therapy (CBT) and mindfulness-based techniques;
4. **Personal Insight Generation:** Analyzing patterns across multiple journal entries to provide actionable personal growth recommendations and identify emotional trends.

4.2.2. Prompt Engineering for Reliability

To ensure reliable integration and therapeutic appropriateness, all Gemma 3n interactions employ structured JSON output validated through the `json_repair` Python library (Baccianella, 2025). This approach provides several critical benefits:

- **Predictable Integration:** Structured outputs enable reliable parsing and integration with the application's data models;
- **Therapeutic Consistency:** Standardized response formats ensure consistent therapeutic framing and safety considerations;
- **Quality Assurance:** JSON validation catches malformed responses and enables graceful error handling in sensitive mental health contexts.

All therapeutic prompts are centrally managed in a `prompts.json` configuration file, enabling easy refinement and clinical validation of the AI's therapeutic approach. Example prompt structures include:

[1] thanks Google for the generous daily-free API access to the Gemma 3n models!

Listing 1 | Gemma 3n Mood Analysis Prompt Structure

```
"mood_analysis": {
  "system": "You are a therapeutic AI assistant analyzing journal entries for
      mood and emotional patterns. Provide objective, supportive insights using
      CBT and mindfulness principles...",
  "user": "Analyze the emotional tone and mood of the following journal entry.
      Provide a mood score from 1-10, identify key themes, and generate
      thoughtful follow-up questions: {content}",
  "output_format": {
    "mood_score": "number",
    "primary_emotion": "string",
    "themes": "list",
    "follow_up_questions": "list"
  }
}
```

5. Technical Challenges and Solutions

The development of QuietRoom from concept to functional prototype involved solving a series of practical and ethical engineering challenges.

5.1. Challenge 1: Bridging the Development-to-Deployment Hardware Gap

A significant challenge for me was the discrepancy between the project's ambitious goal of using powerful, on-device models and the resource constraints of my current setup, an 8-year-old MacBook without dedicated GPU with a very low inference speed. The task was to build and validate the full application pipeline without being bottlenecked by local hardware performance.

5.1.1. Solution

The strategic adoption of the **LiteLLM** framework was the key. By creating an abstraction layer, I was able to develop the entire application against the remote `gemini/gemma-3n-e4b-it` model endpoint in Google Vertex AI. This allowed for rapid iteration and validation of the AI-driven features, proving the correctness of the data flow and prompt engineering independently of the final deployment target. This "develop remotely, deploy locally" architecture ensures the application is robust and ready for on-device use on any capable machine.

5.2. Challenge 2: Delivering a Seamless Public Demo via Containerization

Providing a live, interactive demo for the judges required packaging a multi-component application (a React frontend and a Python backend) into a single, reliable, and publicly accessible artifact.

5.2.1. Solution

The solution was to fully containerize the application using **Docker**. A `Dockerfile` was created to build both the production-ready React assets and the FastAPI server environment. This container was then deployed on **Hugging Face Spaces** (Hugging Face, 2025), a platform for hosting interactive ML applications free-of-charge with CPU servers. Since I am using remote APIs as model serving option, CPU servers are suitable for a live demo. Overcoming initial hurdles with Docker networking and port configurations resulted in the stable, self-contained demonstration now publicly available.

5.3. Challenge 3: Ensuring Reliable and Safe AI Behavior

A general-purpose model like Gemma 3n is powerful but not inherently specialized for a sensitive task like mental wellness journaling. The primary ethical and technical challenge was constraining the model to act as a safe, empathetic, and predictable assistant.

5.3.1. Solution

A multi-layered prompt engineering strategy was implemented, as detailed in Section 4.2.2. This includes defining **strict personas**, requiring **structured JSON outputs** for reliability, and creating **dedicated safety layers** for critical tasks like crisis detection. It is important to acknowledge that ensuring AI safety in a mental wellness context is a profound challenge. While this proof-of-concept establishes a strong technical foundation, further collaboration with domain experts is a critical next step for future development.

5.4. Challenge 4: Validating Multimodal Capabilities

The core vision for QuietRoom hinged on leveraging one of Gemma 3n's most compelling capabilities: its understanding of interleaved text, images, and audio. However, a common challenge when working with cutting-edge technology is that toolchains and API availabilities are constantly evolving. During the development period, the public Gemma 3n endpoint accessible via the API did not yet support multimodal inputs. This presented a critical dilemma: how to build and validate the application's entire multimodal data pipeline without direct access to the final model's full feature set?

5.4.1. Solution

The solution was to decouple pipeline development from model validation:

1. **Pipeline-First Development:** I first engineered the complete, model-agnostic multimodal pipeline in the backend. This involved processing image (base64) and audio inputs to construct interleaved prompts, all without dependency on a specific model endpoint.
2. **Validation via Proxy Model:** I then verified the pipeline's correctness using **Gemini 2.5 flash-lite** as a temporary proxy. Our LiteLLM abstraction layer made this test trivial, allowing us to switch models with a single configuration change.

This approach successfully confirmed that QuietRoom's end-to-end multimodal functionality is correctly implemented and fully operational. The system stands ready to be seamlessly pointed to Gemma 3n's multimodal endpoint as it becomes available or when the model is run locally, demonstrating the architecture's resilience and readiness for the model's full suite of features.

6. Conclusion and Future Work

6.1. Conclusion

In an era where personal data is the currency of the digital world, QuietRoom stands as a testament to a different path forward. It began as a direct response to the privacy paradox of modern AI, the dangerous trade-off between intelligent assistance and data confidentiality. This paper has detailed the engineering of a functional proof-of-concept that is more than a journaling application; it is a blueprint for building truly private, user-centric AI experiences.

Through a flexible, on-device-ready architecture and meticulous prompt engineering, QuietRoom demonstrates that the power of Google's Gemma 3n can be harnessed to create a secure sanctuary for self-reflection. The project's success proves that we do not have to sacrifice privacy for intelligence. By prioritizing local processing and user control, we can build tools that are not only helpful but also fundamentally trustworthy.

Developed by a solo PhD student in Trento, Italy, during the summer of 2025 and free-time moments, QuietRoom is a passion project born from the belief that the future of AI must be personal, private, and respectful of our humanity. Hope that this project inspires future research and ideas.

6.2. Future Work

The proof-of-concept establishes a strong foundation for several exciting future enhancements:

- **Full On-Device Deployment:** Migrating the validated architecture to a dedicated mobile or desktop application using frameworks like the **Google AI Edge SDK** to achieve full on-device inference via web apps;
- **Long-Term Memory and Personalization:** Integrating a local Retrieval-Augmented Generation (RAG) system that uses the SQLite database as a long-term memory. This would allow the AI companion to recall past entries and conversations, providing far more personalized and context-aware support over time;
- **Interactive Visualizations:** Moving beyond static charts to create dynamic, interactive data visualizations that allow users to explore the connections between their moods, activities, and journal themes over time.

7. Acknowledgment

I would like to express my gratitude to the Google and Kaggle teams for organizing the Gemma 3n Impact Challenge. This opportunity provided to me hands-on experience with cutting-edge models and fostered the creation and thinking of a real-world case study. I hope this work contributes a positive stimulus to the vital research fields of privacy-preserving AI and the application of technology to mental wellness. I will continue the development of this idea in the near future and keep exploring on-device techonology!

8. Disclaimer

QuietRoom has been developed exclusively as a proof-of-concept for the "Google - The Gemma 3n Impact Challenge". It is not intended for use as a medical or therapeutic tool.

References

S. Baccianella. Json repair - a python module to repair invalid json, commonly used to parse the output of llms, feb 2025. URL https://github.com/mangiucugna/json_repair.

M. Bayer. SQLAlchemy: The Database Toolkit and Object Relational Mapper for Python, 2025. URL https://www.sqlalchemy.org/.

Gemma Team (Google DeepMind). Gemma 3n, 2025. URL https://deepmind.google/models/gemma/gemma-3n/.

Google AI Studio. Google AI Studio: API Key Management, 2025. URL https://aistudio.google.com/u/1/apikey.

Google DeepMind / Kaggle. The Gemma 3n Impact Challenge, 2025. URL https://www.kaggle.com/competitions/google-gemma-3n-hackathon.

Hugging Face. Hugging Face Spaces: The AI App Directory for hosting ML demos, 2025. URL https://huggingface.co/spaces.

LiteLLM Team / BerriAI. LiteLLM: A unified interface for accessing 100 + LLMs, 2025. URL https://www.litellm.ai/.

M. McCain, R. Linthicum, C. Lubinski, A. Tamkin, S. Huang, M. Stern, K. Handa, E. Durmus, T. Neylon, S. Ritchie, K. Jagadish, P. Maheshwary, S. Heck, A. Sanderford, and D. Ganguli. How people use claude for support, advice, and companionship, 2025. URL https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

Pausa Team. Digital vs Traditional Journaling: Pros and Cons, 2025. URL https://www.pausa.co/blog/digital-vs-traditional-journaling-pros-and-cons.