Design and Evaluation of a Memory-Enhanced Chatbot Architecture for Supporting ME/CFS Patients

Hannes Matthias Ehringfeld

August 7, 2025

1st Reviewer:	Prof. Dr. Christian Meske
2nd Reviewer:	Tobias Hermanns, M. Sc.

Presentation Outline & Methodology

- This presentation follows the 6-step process model of the **Design Science Research Methodology (DSRM)** (Peffers et al., 2007).
- DSRM was applied consistently throughout the thesis.
- The presentation is structured along the first five steps of DSRM:
  1. Problem Identification & Motivation
  2. Definition of Objectives for a Solution
  3. Design & Development
  4. Demonstration
  5. Evaluation
- This presentation itself fulfills the final step of the methodology: **Step 6, Communication**.

Problem Identification: ME/CFS and PEM

- **Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)** is a severe, chronic illness characterized by profound fatigue, cognitive impairment ("brain fog"), and other debilitating symptoms (Institute of Medicine et al., 2015).
- The hallmark symptom is **Post-Exertional Malaise (PEM)**: a severe worsening of all symptoms following even minimal physical or mental exertion (Institute of Medicine et al., 2015; Stussman et al., 2020).
- The key challenge is the **delayed onset** of PEM, which can occur hours or even days later, making it extremely difficult to identify the specific trigger (cause) for a PEM crash (effect) (Institute of Medicine et al., 2015; Stussman et al., 2020).

Problem Identification: Pacing and Documentation

- There is no cure or comprehensive treatment for ME/CFS.
- The primary clinical advice is **Pacing**: a self-management technique to stay within an individual's limited "energy envelope" to avoid triggering PEM (Eckey et al., 2025; Jason et al., 2013).
- Pacing is supported by **logging** daily activities and symptoms to understand the delayed cause-effect relationship of PEM and identify personal triggers (Shepherd & Mayes, 2023; Solve ME/CFS Initiative, 2025).
- **Problem:** The cognitive impairment and fatigue inherent to ME/CFS make detailed documentation a significant **burden**, which can itself consume a patient's limited energy (FDA, 2013; Institute of Medicine et al., 2015).

From Problem to Solution

- This leads to the central **research question**:

  "How can current natural language processing technology be integrated into a system architecture to enable continuous, conversation-based tracking of activities and symptom severity for ME/CFS patients, thereby reducing the documentation burden of the pacing strategy?"

- To answer this question I created:

  LogChat - a proof-of-concept system architecture for a memory-enhanced personal chatbot that demonstrates technical mechanisms to transform natural conversation into a structured health diary while also providing educational advice.

Positioning in the Research Landscape

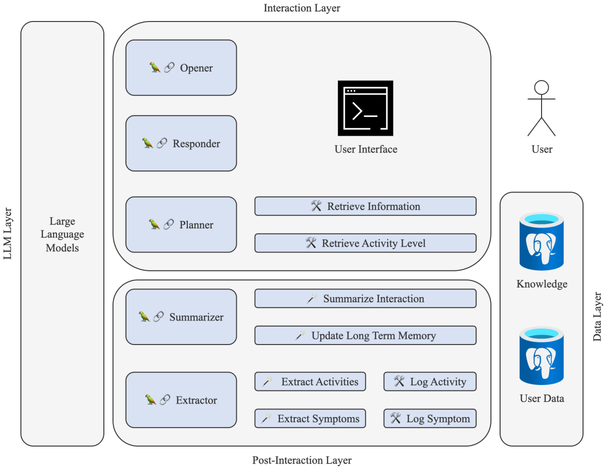
- Data trackers like **Visible** collect quantitative data but fail to reduce the cognitive burden of manual logging (Visible Health Inc., 2025).
- Educational platforms like **MyGuide** offer resources but lack the dynamic tracking required for effective pacing (Naik et al., 2024).
- User studies confirm a high demand for accessible tools that accommodate fatigue and cognitive impairment (Taygar et al., 2025).
- LogChat adapts modularity and memory concepts from frameworks like **openCHA** and **MemoryBank** for a clinical setting (Abbasian et al., 2024; Zhong et al., 2023).
- **Research Gap**: No tool combines a low-friction conversational UI with a persistent memory system to ease the documentation burden of ME/CFS pacing.

Definition of Objectives

- LogChat is designed to enable three use cases, each with specific objectives:
  - **Conversation**: Enable natural interaction which is manageable for users with brain fog while maintaining short-term and long-term memory to act as an empathetic, context-aware companion.
  - **Logging**: Accurately extract symptoms and activities from conversation, support simplified and automated baseline logging, and convert all inputs into structured, analyzable data.
  - **Question Answering**: Empower the user by retrieving educational information on ME/CFS and providing on-demand access to their own activity data.
- **Non-Functional Goal**: The architecture was designed to be modular to support a long-term vision for a private, trustworthy, and efficient on-device application.

Design & Development: LogChat System Architecture

- **Interaction Layer**: Manages the live conversation via a strategic *Planner* and an empathetic *Responder*.
- **Post-Interaction Layer**: Processes the dialogue after it ends, using a *Summarizer* for memory and an *Extractor* to log structured data.
- **Data Layer**: Stores all user data, structured logs, and a curated knowledge base in a PostgreSQL database.
- **LLM Layer**: Enables the system's agentic behavior and is abstracted to allow for swapping different Large Language Models.



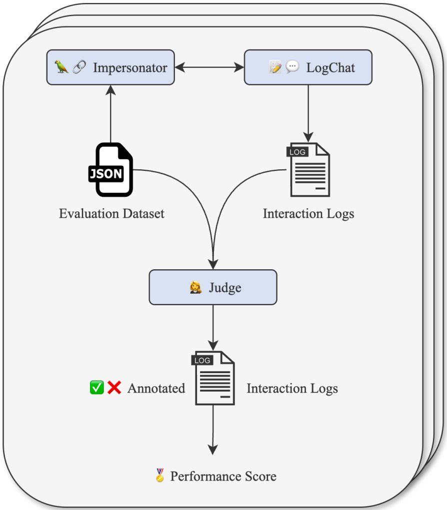
Design & Development: Key Technical Mechanisms

- **Dual-Node Response Generation:** A *Planner* node handles logical reasoning while a separate *Responder* node crafts the empathetic response. This separation of planning from formulation was inspired by the *openCHA* framework (Abbasian et al., 2024).
- **Hierarchical Memory Synthesis:** A layered memory, adapted from the *MemoryBank* framework (Zhong et al., 2023), consists of a persistent **user profile** and chronological **interaction summaries**.
- **Agentic Tool Use for Logging:** A dedicated *Extractor* agent autonomously converts the unstructured conversational dialogue into structured database entries by calling tools like `log_symptom()` or `log_activity()`.
- **Retrieval-Augmented Generation (RAG):** To provide reliable education, the system grounds its answers in a curated knowledge base. This RAG approach ensures that information about ME/CFS and pacing is trustworthy and minimizes the risk of factual hallucination.

Demonstration & Evaluation:

I developed a custom, automated framework to repeatedly test LogChat's ability to meet its objectives.

- **Evaluation Dataset**: Defines 3 distinct user personas and 15 scripted interaction scenarios with checklists.
- **Impersonator Agent**: An LLM-powered agent that simulates a user conversation with LogChat, following a specific persona and script.
- **Judge Agent**: A second LLM agent that audits the interaction logs, comparing system behavior against a detailed checklist to quantify performance.



Quantitative Performance Evaluation

To assess the architecture's robustness and its feasibility for the objective of an offline-capable system, LogChat was evaluated with a range of proprietary and open-source LLMs.

Model Name	Parameters (B)	Quantization Level	File Size (GB)	Context Window	Total Inputs	Achieved Inputs	Total Outputs	Achievable Outputs	Achieved Outputs	Score (Fraction)	Score (Decimal)
`Gemini 2.0 Flash`	Proprietary	N/A	N/A	1M	78	77	100	99	98	98/99	**0.9899**
`GPT-4o`	Proprietary	N/A	N/A	128k	78	78	100	100	96	96/100	0.9600
`Gemini 2.5 Flash`	Proprietary	N/A	N/A	1M	78	78	100	100	94	94/100	0.9400
`Qwen2.5 14B`	14	q4_K_M	9.0	32k	78	78	100	100	90	90/100	**0.9000**
`Cogito 14B`	14	q4_K_M	9.0	128k	78	78	100	100	87	87/100	0.8700
`Qwen3 8B`	8	q4_K_M	5.2	40k	78	78	100	100	76	76/100	0.7600
`Qwen3 4B`	4	fp16	8.1	40k	78	78	100	100	70	70/100	0.7000
`Llama 3.1 8B`	8	q4_K_M	4.9	128k	78	78	100	100	56	56/100	0.5600
`Hermes 3 8B`	8	q4_K_M	4.9	128k	78	77	100	99	53	53/99	0.5354
`Qwen3 14B`	14	q4_K_M	9.3	40k	78	76	100	97	49	49/97	0.5052

Key Finding: A significant performance gap exists between high-performing proprietary models and currently available open-source models suitable for on-device deployment.

Qualitative Performance Evaluation

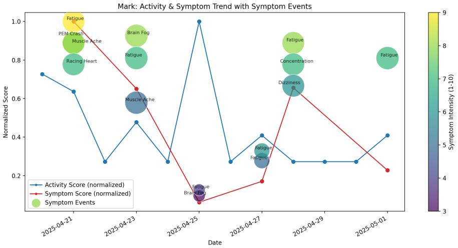
Key Finding: In LogChat, a 90% score is functionally a failure. The qualitative analysis revealed that even top-scoring open-source models have critical logging and memory errors, corrupting user data and rendering them untrustworthy.

Functional Objective	Gemini 2.0 Flash	Qwen2.5 14B
Conversation		
Conversational Interaction	Met	Met
Short-Term Memory	Met	Partially Met
Long-Term Memory	Met	Partially Met
Logging		
Symptom & Activity Logging	Met	Not Met
Simplified & Baseline Logging	Met	Not Met
Structured Data Output	Met	Not Met
Question Answering		
Information Retrieval	Met	Met
Activity Score Retrieval	Met	Partially Met

Visual Analysis of Data Quality

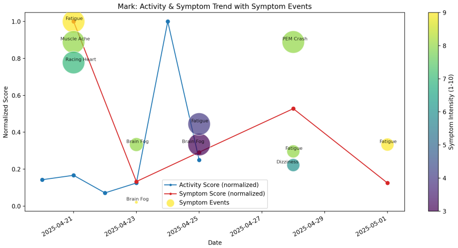
Visualizing the logged data starkly reveals the difference: one plot tells a coherent story, while the other is incomplete and misleading.

Gemini 2.0 Flash (Reliable)



This plot correctly visualizes the user's post-exertional crash after a day of high activity. The data is complete and tells a true story.

Qwen2.5 14B (Unreliable)



Here, critical logging failures create a distorted view. The plot falsely shows low activity and fails to capture the user's actual condition.

Answering the Research Question

- **Yes, current NLP technology can reduce the documentation burden of pacing.**
- This research successfully demonstrates *how*: through a novel, modular agent architecture that uses a dual-node response generation, hierarchical memory, agentic tool use, and RAG to transform natural conversation into a structured health diary.
- The LogChat prototype serves as a successful proof-of-concept, affirming the viability of this approach.
- **However, the utility of this architecture is conditional.** Its reliability is entirely dependent on the underlying LLM. At present, only high-performing proprietary models meet the required standard for this sensitive application, posing a challenge for the long-term vision of a private, on-device system.

Contributions, Limitations, & Future Work

- **Key Contributions**:
  - A novel, modular agent architecture for conversational health logging.
  - A replicable, automated evaluation framework to validate complex agents.
  - Crucial insights into the performance gap defining the practical limits of such systems.
- **Primary Limitation**:
  - The evaluation validates **technical feasibility**, not **clinical utility**. The system's real-world benefit for patients remains an open question.
- **Key Future Work**:
  - **Engage Patients**: Conduct a clinical pilot study for real-world validation and feedback.
  - **Enhance the System**: Develop a full GUI with voice control and add proactive, data-driven insights.
  - **Achieve Privacy**: Implement privacy measures or use model distillation to create smaller, efficient on-device models.

Questions

Thank you for your attention.

Bibliography I

Abbasian, M., Azimi, I., Rahmani, A. M., & Jain, R. (2024, September 25). Conversational health agents: A personalized LLM-powered agent framework. https://doi.org/10.48550/arXiv.2310.02374

Eckey, M., Li, P., Morrison, B., Bergquist, J., Davis, R. W., & Xiao, W. (2025).Patient-reported treatment outcomes in ME/CFS and long COVID. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(28), e2426874122. https://doi.org/10.1073/pnas.2426874122

FDA. (2013, November). *The voice of the patient: Chronic fatigue syndrome and myalgic encephalomyelitis*. Retrieved May 20, 2025, from https://www.fda.gov/media/86879/download

Institute of Medicine, Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome, & Board on the Health of Select Populations. (2015). *Beyond myalgic encephalomyelitis/chronic fatigue syndrome: Redefining an illness*. National Academies Press (US). Retrieved November 12, 2024, from http://www.ncbi.nlm.nih.gov/books/NBK274235/

Bibliography II

Jason, L. A., Brown, M., Brown, A., Evans, M., Flores, S., Grant-Holler, E., & Sunnquist, M. (2013).Energy conservation/envelope theory interventions to help patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Fatigue (Abingdon, Eng. Print)*, *1*(1), 27–42. https://doi.org/10.1080/21641846.2012.733602

Naik, H., Pongratz, K., Malbeuf, M., Kung, S., Last, L., Sugiyama, A., Khor, E., McGuire, M., Levin, A., & Tran, K. C. (2024). Myguide long covid: An online self-management tool for people with long covid. https://doi.org/10.2139/ssrn.4995407

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007).A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Bibliography III

Shepherd, C., & Mayes, S. (2023, May). *PACING: Activity and energy management for people with ME/CFS and long covid*. The ME Association. Retrieved May 22, 2025, from https://meassociation.org.uk/wp-content/uploads/2025/02/PACING-Activity-and-Energy-Management-for-people-with-MECFS-and-Long-Covid-MAY-2023.pdf

Solve ME/CFS Initiative. (2025). *Patient and caregiver resources*. Retrieved May 22, 2025, from https://solvecfs.org/me-cfs-long-covid/patient-and-caregiver-resources/

Stussman, B., Williams, A., Snow, J., Gavin, A., Scott, R., Nath, A., & Walitt, B. (2020).Characterization of post–exertional malaise in patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.01025

Bibliography IV

Taygar, A. S., Bartels, S. L., de la Vega, R., Flink, I., Engman, L., Petersson, S., Johnsson, S. I., Boersma, K., McCracken, L. M., & Wicksell, R. K. (2025).User-driven development of a digital behavioral intervention for chronic pain: Multimethod multiphase study. *JMIR formative research*, *9*, e74064. https://doi.org/10.2196/74064

Visible Health Inc. (2025). *Visible - activity tracking for illness, not fitness.*. Retrieved December 12, 2024, from https://www.makevisible.com/

Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2023, May 21). MemoryBank: Enhancing large language models with long-term memory. https://doi.org/10.48550/arXiv.2305.10250