Title: Developing Computational Architectures for Emotionally Consistent AI Companions

Supervisors: Armando Vieira PhD, Kuldar Taveter PhD
Key Words: Affective Computing, Conversational AI, Long-term Interaction, Computational Architecture, Emotion Modeling, User Engagement.

1. Background and Motivation

AI companions and chatbots are increasingly deployed in long-term, high-stakes contexts such as mental health support (e.g., Woebot, Wysa), educational tutoring, and elderly care. A critical factor for their success and user trust is the perception of a stable, coherent "personality" and emotional response. Current chatbot architectures, often based on large language models (LLMs), excel at generating contextually appropriate single-turn responses but lack a persistent internal state. This leads to **emotional inconsistency**—where the AI's expressed emotions, opinions, and recollections may change erratically across conversations.

The "schizophrenia of AI" disrupts user engagement, breaks the illusion of social connection, and poses significant risks in therapeutic settings where stability is paramount. This project aims to address this gap by moving beyond single-turn response generation to design a computational architecture that maintains and evolves a consistent emotional state over time. Inspired by the modern neuroscientific Theory of Constructed Emotion, our approach moves beyond static emotion categories to model a dynamic, context-dependent emotional state, creating a more believable and reliable AI companion.

2. Research Objectives

The primary objective is to design, implement, and evaluate a novel computational architecture for an AI companion that demonstrates significantly improved emotional consistency over long-term interactions.

Specific objectives include:
1. **Literature Review:** Conduct a comprehensive review of existing computational emotion models, with a focus on modern theories like the Theory of Constructed Emotion, memory architectures for chatbots, and evaluation metrics for long-term human-AI interaction.
2. **Architecture Design**: Propose a hybrid architecture that integrates: * A dynamic emotion model whose state persists across sessions and evolves based on user interaction. This model will be grounded in the principles of the Theory of Constructed Emotion, representing affective state as a dynamic, multi-dimensional construct that is continuously updated based on interoceptive (internal) and exteroceptive (user interaction) cues.
3. **Implementation:** Develop a functional prototype implementing the proposed architecture.
4. **Evaluation:** Design and run a longitudinal user study (e.g., over 2-4 weeks) to quantitatively and qualitatively assess the prototype against a baseline LLM on metrics of consistency, engagement, and user perception.

3. Methodology

- **Phase 1: Modeling & Design:** The student will explore models for representing emotion. The core innovation will be the design of a computational framework based on the

Theory of Constructed Emotion. Instead of classifying emotions into discrete categories (e.g., joy, anger), the AI's affective state will be represented as an evolving set of core affective dimensions (e.g., Valence, Arousal) and a conceptual map. "Update rules" will be designed to simulate how the AI's emotional state is constructed based on the context of the user's input (exteroceptive cues), its own prior state (interoceptive cues), and the retrieved memories that give the interaction meaning.

- **Phase 2: Memory Integration:** A database will be designed to store summaries of past interactions tagged with emotional significance. This memory will be retrieved to inform current responses (e.g., "Last week you mentioned feeling anxious about exams. How did that go?"), ensuring conversational continuity.
- **Phase 3: System Development:** The prototype will be built, likely using a framework like LangChain or LlamaIndex to orchestrate the memory, emotion model, and an LLM API (e.g., OpenAI GPT, Llama 2/3). The LLM will be carefully prompted or fine-tuned to express the designated emotional state.
- **Phase 4: Evaluation:** A within-subjects study will be conducted where participants interact with both the proposed AI and a baseline AI. Methods will include:
    - **Quantitative:** User engagement metrics (session length, frequency), standardized questionnaires (e.g., Godspeed Questionnaire Series for perceived consistency and anthropomorphism).
    - **Qualitative:** Thematic analysis of user interviews and feedback to identify moments where consistency was perceived or broken.
    - **Technical:** Expert-led consistency checks (e.g., presenting the AI with the same scenario at different times to check for drift in response).

4. Expected Outcome and Contribution

This project will contribute: * A novel, open-source computational architecture for emotionally consistent AI. * Empirical evidence on the importance of emotional consistency for long-term user engagement. * A validated framework for evaluating emotional consistency in conversational agents. * A research paper suitable for publication in a relevant conference (e.g., ACM CHI, ACL, IVA).

5. Candidate Profile

We are looking for a highly motivated student with:

- A strong background in Computer Science, Data Science, or a related field.
- Proficiency in Python programming and experience with AI/ML libraries (e.g., PyTorch, TensorFlow, Hugging Face).
- A solid understanding of (or strong interest in learning about) NLP, LLMs, and Affective Computing.
- Excellent analytical and writing skills.
- Knowledge of human-subject research ethics is an advantage.

References

1. Theory of constructed emotion meets RE: An industrial case study [link]
2. **Modern Emotion Theory**: Barrett, L. F. (2017). How Emotions are Made: The Secret Life of the Brain. Houghton Mifflin Harcourt.

3. **Memory in Chatbots:** Xu, J., Szlam, A., & Weston, J. (2022). "Beyond Goldfish Memory: Long-Term Open-Domain Conversation." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
4. **The Problem of Consistency:** Ji, T., Fu, J., Xu, W., Mao, K., & Zeng, M. (2023). "Towards Emotional Consistency in Dialogue System." *Findings of the Association for Computational Linguistics: EACL 2023*.
5. **Evaluation & Long-term Engagement:** Araujo, T. (2018). "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions." *Computers in Human Behavior, 85*, 183-189.
6. **Therapeutic Context:** Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial." *JMIR Mental Health, 4*(2), e19.

How to Apply: Interested candidates should email armando.vieira@ut.ee with their CV and a brief statement of interest, highlighting their relevant experience and motivation for this project.