Title:

A Three-Tier Personality Framework for Behaviorally Consistent Large Language Models: Integrating Psychological Theory into LLM Agent Design

---

1. Background and Motivation

Large Language Models (LLMs) are increasingly being deployed in interactive and multi-role scenarios, where coherence, personality, and social alignment are key to user trust and engagement. Despite recent progress in persona-conditioned generation, current techniques—such as prompt injection, fine-tuning, or memory augmentation—often fall short in maintaining consistency across contexts and over time. These approaches tend to treat "personality" as a surface-level token control or style shift, lacking a principled and structured framework that reflects how personality governs behavior in humans.

In contrast, psychological science offers well-established models that distinguish between relatively stable personality traits (e.g., the Big Five), contextual adaptations (e.g., values, goals, and motivations), and moment-to-moment behavioral expressions (e.g., emotions, tone, stance). Frameworks such as McAdams' three-layer personality model, the Five Factor Theory (FFT), and the TESSERA framework describe personality as a layered and dynamic system. However, these insights have yet to be systematically integrated into LLM design.

This project aims to fill that gap by proposing a three-tier personality modeling framework for LLMs, grounded in psychological theory: - **Stable Layer**: enduring dispositions such as personality traits and identity; - **Growth Layer**: experience-based adaptations like goals, motivations, and role-specific knowledge; - **Performance Layer**: expressive behaviors including tone, stance, affect, and verbal style, modulated by situational context.

---

2. Research Objectives and Contributions

This research aims to establish a structured and psychologically grounded framework for modeling, injecting, and evaluating personality in LLMs. The key contributions include:

1. **A Three-Tier Personality Framework for LLMs**: A Stable–Growth–Performance architecture supporting enduring consistency and flexible expression.

2. **Mapping Personality Layers to Model Components**:

3. Stable: trait embeddings/persona profiles;
4. Growth: dynamic memory modules/latent trackers;

5. Performance: prompt-time modulation/vector steering.

6. **Training and Conditioning Strategies**: SFT, contrastive loss, memory injection, and LoRA steering to inject and control layered personality.

7. **Evaluation Metrics**:

8. Consistency across contexts;
9. Expressivity in behavior;
10. Adaptability in role-switching.
11. **Prototype Agent**: A demo LLM agent exhibiting layered personality expression, adaptable over long-term and cross-scenario interactions.

---

3. Methodology and Technical Roadmap

3.1 Layer Specification

- **Stable Layer**: Big Five traits, identity archetypes; fixed vectors or long-term memory entries.
- **Growth Layer**: Acquired values, goals, roles; adaptive modules updated via feedback or exposure.
- **Performance Layer**: Affective states, stance, style; controlled via prompts or dynamic decoding.

3.2 Data Strategy

- Collect multi-context persona-behavior triplets from:
- Twitter, podcasts, interviews;
- Persona-Chat, Character-DIAL, RealToxicityPrompts;
- Public biographies and annotated corpora.
- Use GPT-4/5 for tagging traits, stances, and style labels.

3.3 Model Training

- Base Models: gpt-oss-20b, Qwen3-8b, LLaMA3.1-8b-instruct.
- Techniques:
- Persona-aligned SFT;
- Memory-based growth conditioning;
- Contrastive persona-behavior supervision;
- Steering module for performance layer.

3.4 Evaluation

- **Consistency**: Persona retention score, trait drift tracking.
- **Expressivity**: Human-style judgment, classification accuracy.
- **Adaptability**: Role-change plausibility, response plausibility.

3.5 Implementation Resources

- **Compute**: A100 GPU x 4–8; ~1000–2000 GPU hours.
- **LLM API Use**: GPT-4/5 ~15M tokens for preprocessing.
- **Personnel**: 1 NLP researcher (full-time), 1 data engineer, 1 psychology advisor.

---

4. Related Work

Recent studies on LLM personality modeling include: - **Prompt-based conditioning**: e.g., Wang et al. (2023), Liu et al. (2023); - **Emergent persona tracking**: e.g., Zhao et al. (2025); - **Behavioral injection**:

e.g., Li et al. (2025); - **Memory/vector personalization**: PersonalGPT (Cao et al., 2023), PersonaAdapt (Chen et al., 2024); - **Psychological alignment**: Caron & Srivastava (2022), Mills & Emmons (2023).

None of these methods integrates multi-layer personality theory or behaviorally grounded modeling across contexts and time.

---

5. Theoretical Foundations

Our framework is rooted in: - **McAdams' Three-Level Personality Theory**; - **Five-Factor Theory (FFT)**; - **Whole Trait Theory (WTT)**; - **TESSERA Framework**.

These models support the stratification of traits, experiences, and behavioral expression, directly informing our LLM component design.

---

6. Significance and Future Directions

This work offers: - A principled foundation for interpretable and controllable personality modeling in LLMs; - A methodology for layered persona injection and adaptation; - A bridge between psychological theory and machine learning practice.

Future extensions include personality development modeling, real-time emotional feedback integration, and multi-agent personality interaction simulations.