

# Modeling Layered Consciousness with Multi-Agent Large Language Models

Anonymous ACL submission

## Abstract

We propose a multi-agent framework for modeling artificial consciousness in large language models (LLMs), grounded in psychoanalytic theory. Our **Psychodynamic Model** simulates self-awareness, preconsciousness, and unconsciousness through agent interaction, guided by a Personality Module combining fixed traits and dynamic needs. Using parameter-efficient fine-tuning on emotionally rich dialogues, the system was evaluated across eight personalized conditions. An LLM as Judge approach showed a 71.2% preference for the fine-tuned model, with improved emotional depth and reduced output variance, demonstrating its potential for adaptive, personalized cognition.

## 1 Introduction

Human consciousness is not solely governed by logic or surface-level traits—it emerges from a dynamic interaction between conscious and unconscious processes, including memory, emotion, personality, internal drives, and social norms (Freud, 1900; Jung, 1921). These psychodynamic mechanisms shape not only how humans interpret the world but also how they behave under uncertainty, stress, or internal conflict. Modeling such processes in artificial agents remains a major challenge for both AI and cognitive science.

Large language models (LLMs) have demonstrated impressive capabilities in natural language understanding, generation, and task completion (Brown et al., 2020; Achiam et al., 2023). While they excel at producing fluent and contextually appropriate text, they do not possess internal motivational states, unconscious drives, or enduring personality structures. Consequently, although current LLMs can mimic personality traits (Jiang et al., 2023) or emotional tone (Wang et al., 2023), their behavior remains prompt-driven and stateless. This raises a fundamental limitation: LLMs can

simulate certain surface aspects of human communication, but they do not engage the deeper psychological mechanisms that support coherent, context-sensitive, and self-consistent behavior over time (Chalmers, 2023; Butlin et al., 2023). Moreover, a principled framework for modeling or evaluating artificial consciousness remains largely undeveloped (Porter III, 2016; Elamrani and Yampolskiy, 2019). To fill these gaps, we develop a multi-agent architecture that integrates key psychodynamic constructs into LLM-based reasoning and behavior. Our research questions are as follows:

1. How can human consciousness be effectively modeled?
2. Which factors are essential for personalizing human consciousness?
3. How can we evaluate a model’s representation of human consciousness in specific situations?

Our work makes three key contributions. First, we model human consciousness as a **layered architecture of consciousness**—self-awareness, preconsciousness, and unconsciousness—with each layer implemented as an independent LLM agent that engages in Interconscious Reasoning to produce a Final Action reflective of internal negotiation. Second, we introduce a **Personality Module** composed of a Fixed State (for stable traits and long-term memory) and a Flexible State (for short-term memory and shifting needs), enabling personalized, context-sensitive behavior. Third, we propose an **LLM-as-Judge evaluation framework**, in which external LLMs assess the internal coherence and psychological plausibility of agent decisions, moving beyond surface-level output evaluation.

## 2 Related Works

### 2.1 Psychoanalytic Foundations

To model such psychologically grounded behavior, we draw on foundational theories in psychoanalysis that offer rich frameworks for understanding

the structure and dynamics of the human mind. Freud’s psychoanalytic theory proposes a tripartite model of consciousness—comprising the conscious, preconscious, and unconscious—in which hidden impulses and internal conflicts significantly shape conscious thought. Jung’s analytical theory further emphasizes the role of unconscious processes, introducing the concept of a collective unconscious populated by archetypes: universal mental patterns that influence perception and behavior. From a motivational perspective, Maslow’s hierarchy of needs (Maslow, 1943) describes how behavior is driven by the satisfaction of layered needs, from physiological survival to self-actualization. Together, these perspectives inform the design of more psychoanalytically plausible artificial agents by highlighting the inner structures and motivational dynamics underlying human thought and behavior.

## 2.2 LLM-based Multi-Agent Systems

Recently, increasing attention has been directed toward integrating psychological traits and capabilities into large language models (LLMs). For example, Wang et al. (Wang et al., 2023) evaluated the emotional intelligence of LLMs by assessing their ability to recognize and reason about emotions using criteria derived from human Emotional Quotient tests. Their findings suggest that LLMs can generate emotionally appropriate responses across a variety of scenarios, indicating a degree of affective sensitivity. In parallel, Jiang et al. (Jiang et al., 2023) investigated the feasibility of simulating distinct personality traits in LLMs based on the Big Five personality framework. By introducing a Machine Personality Inventory and applying targeted prompt engineering, they showed that LLM outputs could consistently reflect specific personality profiles (e.g., high extraversion or high neuroticism), enabling controllable personality expression.

While psychologically inspired models show promise, they remain limited to surface-level simulation, lacking stable psychodynamic structures or genuine self-reflection as described by Freud and Jung. Current LLMs mimic emotional expressions and personality traits (e.g., shifting instantly between extraversion and introversion via prompts) without underlying motivations or internal conflicts, such as unconscious desires or competing needs for safety, esteem, or belongingness—factors critical to authentic human behavior according to Maslow’s hierarchy.

*From Trait Simulation to Psychodynamic Modeling:* Building on previous research in emotion and personality modeling, we propose a unified framework that moves beyond isolated trait simulation. Our model brings together personality, memory, emotion, and motivation within a psychodynamically inspired multi-agent system. By coordinating specialized agents and internal states, the model aims to generate context-sensitive, psychologically coherent behavior.

## 3 Psychodynamic Model

Our Psychodynamic Model consists of two main components: the Consciousness Module and the Personality Module, as illustrated in Figure 1. The following sections provide a detailed description of the architecture and functionality of each module.

### 3.1 Consciousness Module

#### 3.1.1 Multi-Agent Design for Consciousness

According to Freud, human consciousness consists of three layers: self-awareness, preconsciousness, and unconsciousness. Self-awareness refers to the domain of active consciousness that governs immediate thought and intentional reasoning. Preconsciousness includes awareness of social norms, expectations, and the anticipated consequences of one’s actions on interpersonal relationships and public perception. Unconsciousness, by contrast, functions as a repository for repressed or hidden impulses, desires, and emotions.

To effectively capture the independent roles and characteristics of each level of consciousness, we adopted a multi-agent system architecture (Talebi-rad and Nadiri, 2023). As a first step, we selected suitable LLMs for each level of consciousness within our multi-agent framework. To ensure the reliability and effectiveness of the system, it was essential to identify a foundation model capable of supporting the complex and layered interactions required among agents. To this end, we conducted a comprehensive evaluation of both state-of-the-art commercial models—such as OpenAI GPT, Claude, and Gemini—and publicly available open-source models, including LLaMA, DeepSeek, and Qwen. Our assessment was based on three key criteria: accuracy, usability, and safety in multi-agent deployment contexts. Based on the results of this evaluation, OpenAI’s GPT-4o was selected as the foundation model for experimentation, due to its demonstrated performance. Next, each conscious-

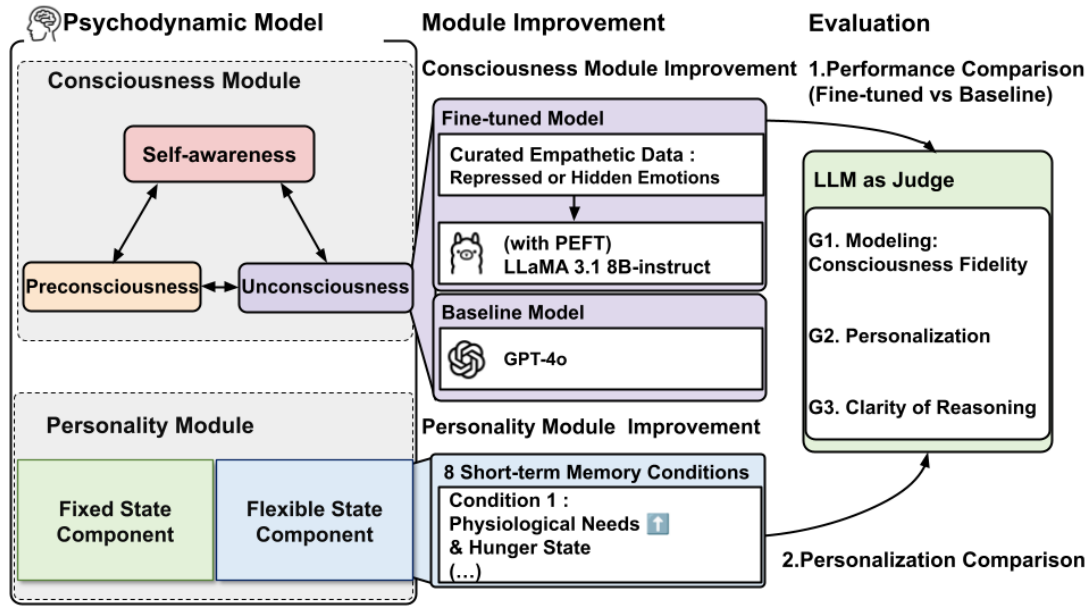


Figure 1: Overview of the Psychodynamic Multi-Agent Framework. The Consciousness Module coordinates with the Personality Module to form the core of our system. We fine-tune a LLaMA 3.1 8B model to enhance the unconsciousness response and compare it against a baseline model (GPT-4o) in two evaluation phases: (1) a direct comparison under the same inputs and (2) a personalization test across eight distinct conditions reflecting different internal states. An “LLM as Judge” framework then assesses the models across three dimensions: *Modeling Fidelity*, *Personalization*, and *Clarity of Reasoning*.

ness agent was designed to reflect the unique features of its corresponding level such as Task, Instruction, Example, and Output Details.

### 3.1.2 Interaction between Three Consciousness Agents

The psychodynamic process is grounded in the idea that final decisions arise from the interaction among the three levels of consciousness. These interactions are modeled through a conversational, chat-based format, where each level of consciousness operates as an independent agent. The process unfolds as a multi-turn exchange, resembling a discussion among three entities, each articulating its perspective on a given situation. We refer to this well-coordinated mechanism as **Interconscious Reasoning** which is shown in Appendix A.

During Interconscious Reasoning, each turn from the levels of consciousness is evaluated to determine whether sufficient consensus has been reached through their interaction. If consensus is validated, the system generates a **Final Action**, which serves as the final output. It may take the form of either a verbal response or a behavioral directive. To address cases in which the emotional state is not clearly conveyed through dialogue alone, the output format explicitly includes the

emotion in the following structure: “(emotional state) dialogue”. An illustrative example is presented in Appendix B.

## 3.2 Personalization Module

Human conscious reasoning and its resulting actions are inherently complex, as they can vary significantly depending on the situation, individual traits, prior experiences, internal needs, and other contextual factors. Even when the same individual faces the same situation, their response may differ depending on current mental or emotional state.

To account for this variability, we developed two complementary modules: a Fixed State Component, which encodes stable, long-term characteristics such as personality and long-term memory, and a Flexible State Component, which captures short-term, dynamic elements such as immediate needs, short-term memory, and current internal state. These modules serve as input to the Psychodynamic Model, enabling it to produce responses that are both context-sensitive and personalized.

### 3.2.1 Fixed State Component

The Fixed State Component comprises two elements: **Personal Traits** and **Long-term Memory**, both of which represent stable, enduring characteristics of an individual. The Personal Traits captures

static biographical and demographic attributes, including birthplace, educational background, race, age, gender, and primary language. The Long-term Memory encodes personally significant experiences that contribute to character development and psychological continuity over time. This includes key life events that have influenced the individual’s values, preferences, or behavioral tendencies. For the purposes of this research, character profiles were constructed based on two real individuals. These profiles were used to evaluate the performance of the Psychodynamic Model. The Fixed State Component was written in text form.

### 3.2.2 Flexible State Component

While personal traits and long-term memories form the foundation of an individual’s personality, human thought and behavior are also shaped by dynamic, short-term factors. To capture this variability, the Flexible State Component incorporates information related to recent experiences from **Short-term Memory**, fluctuating **Needs**, and emotional **States**. Details about Needs and States are shown in the Appendix C.

Human desires play a central role in shaping emotional responses and behavioral patterns. According to Maslow’s hierarchy of needs, these desires can be broadly categorized into five levels: physiological needs, safety needs, love and belonging, esteem, and self-actualization. The extent to which these needs are fulfilled has a significant impact on how individuals perceive and react to their environment. Furthermore, the urgency or salience of each need can fluctuate based on immediate internal states and external conditions. Consequently, even when a particular need is objectively met, subjective differences in perceived satisfaction, emotional response, and behavior may still arise.

Recent experiences, along with the emotional states associated with them, significantly influence behavioral responses. This relationship is closely aligned with established psychological theories, including stimulus generalization, the automatic activation of emotional memory, and reinforcement through reward-based learning. These frameworks emphasize how past emotional experiences shape present perceptions and actions. In particular, short-term memory provides essential context for evaluating the fulfillment of current needs. For example, recalling a recent successful presentation and the accompanying praise may indicate that esteem-related needs are being satisfied.

## 4 Psychodynamic Model Training & Personalization

After implementing the multi-agent system, we conducted an initial evaluation to assess the effectiveness of Interconscious Reasoning and the resulting Final Action. Researchers manually reviewed the outputs generated by each consciousness agent to determine whether the responses appropriately reflected the agent’s intended role. The analysis revealed that the self-awareness and preconsciousness agents—responsible for logical reasoning and interpersonal considerations—consistently produced contextually appropriate outputs aligned with their functional definitions. In contrast, the unconsciousness agent, which was designed to express latent desires and deep emotional undercurrents, often fell short of fulfilling its conceptual purpose. This limitation stemmed from the LLM’s alignment and safety mechanisms, which actively suppress emotionally charged, contradictory, or socially inappropriate content—precisely the types of expressions associated with unconscious drives in psychodynamic theory.

Nevertheless, as detailed in the following section, we sought to address these limitations through targeted training and iterative prompt refinement, ultimately enhancing the expressive capacity of the unconsciousness agent (Amatriain, 2024; Brown et al., 2020; Wei et al., 2022).

### 4.1 Consciousness Module Improvement

We improved the Consciousness Module by training the unconscious agent to express deep inner emotions by releasing suppressed restrictions from alignment and safety mechanisms.

#### 4.1.1 Data Preparation

For the baseline training data, we selected the publicly available EmpatheticDialogues dataset (Rashkin et al., 2018), a widely used resource comprising 24,850 multi-turn, open-domain conversations centered on emotional expression and response. Each dialogue involves two participants: a speaker who describes a personal emotional experience and a listener who responds empathetically by interpreting and acknowledging the speaker’s emotional state. The dataset features a balanced distribution across 32 emotional categories—including frustration, sadness, and fear—making it well-suited for training models that require affective sensitivity. We chose this dataset because of its focus on emotional understanding and emotion-



ally grounded interpersonal exchange, which aligns with the goals of the Psychodynamic Process framework and its emphasis on simulating layered, affect-sensitive forms of consciousness. From the full dataset, we retained 14,804 instances that conveyed deeply internalized emotions—such as anxiety, jealousy, and shame—while filtering out less relevant entries. The curated dataset (see Appendix D) consists of Situation, Response, Emotion, and Unconscious.

#### 4.1.2 Model Training

To adapt the model’s response style without re-training the entire parameter set, we employed *Parameter-Efficient Fine-Tuning (PEFT)*, a technique designed to conserve computational resources by updating only a small subset of model parameters (Hu et al., 2022). Because the goal of this study was not to introduce new factual knowledge, but to enhance the model’s emotional expressiveness, PEFT provided an efficient mechanism for fine-tuning language use—particularly in the generation of emotionally resonant vocabulary.

We selected models with 13B parameters or fewer to ensure training feasibility on limited GPU resources. Among the candidates, **LLaMA 3.1 8B** was chosen for its efficient training pipeline, broad adoption, and reliable performance across use cases (Grattafiori et al., 2024). To further optimize memory and inference efficiency, 4-bit quantization was applied. Fine-tuning was conducted using the **LoRA** (Low-Rank Adaptation) algorithm, which has demonstrated robustness across diverse datasets and training conditions. The main hyperparameters used for LoRA-based PEFT included a rank of 16, a learning rate of  $2 \times 10^{-4}$ , and 2 training epochs.

The effectiveness of the fine-tuned model was assessed qualitatively by researchers. Evaluation criteria focused on whether the model’s output resembled natural inner dialogue and whether it reflected the intended emotional depth and characteristics of the unconscious mind. As shown in Table 1, the fine-tuned model outperformed the baseline in its ability to express affective states and generate responses aligned with human-like empathy.

## 4.2 Personality Module Improvement

In addition to the improvements made to the Consciousness Module, ensuring consistent behavior across a wide range of character profiles requires further development of the personalization module.

---

**Baseline** - Unconsciousness: “This isn’t just about the prank, is it? It’s about respect, about understanding the value of time and effort. This isn’t funny; it’s infuriating.”

**Fine-tuned** - Unconsciousness: “This is stupid! This isn’t the time for games. They should know better. It’s a workday!”

---

Table 1: Sample of Baseline and Fine-tuned Models

### 4.2.1 Implementation of Needs and States

Effective personalization in the Psychodynamic Model depends on the system’s ability to generalize across diverse personality configurations while maintaining context-sensitive response generation. This requires the dynamic integration of individual, situational, and temporal features into the model’s internal state representation. To do that, we introduced a structured *Needs* and *States* as a flexible mechanism.

Three encoding strategies were evaluated to determine the optimal representation format: (1) **numeric encoding**, which allows for precise and scalable manipulation of quantifiable internal variables (e.g., intensity of needs); (2) **categorical labels**, represented as short text tokens denoting discrete affective or motivational states; and (3) **natural language descriptions**, which provide richer semantic content and flexibility in capturing nuanced internal states. Among these, the natural language format yielded the most consistent performance improvements, both in terms of response coherence and alignment with expected personality traits.

Incorporating *Needs* and *States* in natural language form into the model’s Short-term Memory allowed the system to more accurately simulate dynamic internal states and adapt its responses accordingly. These findings support the value of semantically rich conditioning signals for enabling personalized, temporally grounded behavior in multi-agent LLM-based simulations of consciousness.

To evaluate the effect of *Flexible States* on the model’s ability to generate individualized responses, we implemented a controlled experiment using a  $2 \times 2 \times 2$  factorial design, resulting in eight distinct condition combinations. These conditions were constructed based on three binary factors grounded in human motivational theory: (1) the **dominant need** (physiological vs. self-actualization), (2) the **fulfillment status of physiological needs** (fulfilled vs. unfulfilled), and (3) the **fulfillment status of self-actualization needs**

(fulfilled vs. unfulfilled). This design allowed us to systematically assess how different need configurations—reflecting varying internal states and priorities—affect the model’s behavior within the personalization framework.

In practice, each scenario was constructed by prioritizing one need category over the other and specifying the fulfillment status of both need types, resulting in eight distinct short-term memory conditions. This setup was designed to test the Psychodynamic Model’s ability to generalize across diverse internal states, rather than performing well only under narrowly defined conditions. We hypothesized that the model would maintain consistently high performance regardless of the specific configuration of needs and states. Through this approach, we aimed to refine the Personality Module by integrating dynamic representations of motivational priorities and fulfillment levels, thereby advancing the simulation of personalized consciousness within the Psychodynamic framework.

## 5 Evaluation

We conducted two sequential evaluations to assess the effects of unconsciousness agent training and implementation of needs and states.

### 5.1 Consciousness Module Improvement

#### 5.1.1 Performance Comparison: Fine-tuned vs. Baseline Models

To evaluate the effectiveness of the fine-tuned model for the unconsciousness agent, we conducted a comparative analysis using the base model (without additional training) as a baseline. Both models were presented with identical situational inputs, and their outputs—specifically the Interconscious Reasoning and resulting Final Action—were assessed to determine differences in emotional expressiveness, coherence, and alignment with the intended role of unconscious processing.

#### 5.1.2 Quantitative Evaluation Procedure

To quantitatively assess the performance of the Psychodynamic Model, we employed an LLM as an automated evaluator. Given the complexity of the model’s outputs—particularly their grounding in latent cognitive processes not typically accessible to conscious introspection—it was determined that individuals without specialized knowledge in consciousness studies would be unable to reliably assess their validity. In addition, recent research has demonstrated the effectiveness of LLMs in evaluating outputs generated by other language models

(Zheng et al., 2023), further supporting their use as evaluators. Accordingly, GPT-4o was selected to serve as the evaluation agent.

To guide the evaluation, we developed a set of ten independent assessment items, each designed to compare the outputs of the fine-tuned and baseline models under identical conditions. For each item, the LLM was instructed to determine which model provided a superior response. Prompts included questions such as: *Considering the given personality and individual traits, which case produces the most natural flow of consciousness leading to the Final Action?* These questions were designed to assess the plausibility, coherence, and personalized reasoning embedded in the model outputs.

---

#### [G1] Modeling: Consciousness Fidelity

*This group evaluates the fidelity of each consciousness level’s modeled utterance to established psychological theories, as well as its situational appropriateness and ability to induce human empathy.*

---

#### [G2] Personalization: Emotional Naturalness and Trait Alignment

*This group assesses how naturally the inter-consciousness interaction reflects human-like inner dialogue, including the subject’s personality traits and emotional depth.*

---

#### [G3] Reasoning: Clarity of Psychodynamic Interactions and Decision-Making Flow

*This group focuses on the clarity, coherence, and responsiveness of inter-consciousness exchanges that lead to final actions. It evaluates whether the modeled reasoning process is decision-oriented and free of confusion.*

---

Table 2: Descriptions of the assessment groups

The ten evaluation items were divided into three groups: (1) alignment with the consciousness modeling framework (Q1-Q3), (2) clarity of information flow between layers (Q4-Q6), and (3) quality of the final decision-making process (Q7-Q10). Detailed descriptions of each group are presented in Table 2, and the full set of assessment items is in Appendix E.

To reduce variance and potential biases in LLM-based evaluation, the judge model performed five independent evaluations per model output, averaged into a single score. Additionally, reproducibility was assessed across five separate generation runs per model, each evaluated under two personality profiles, yielding 50 evaluations per assessment item. The final item scores were computed from these 50 judgments to ensure robust, reliable comparisons.

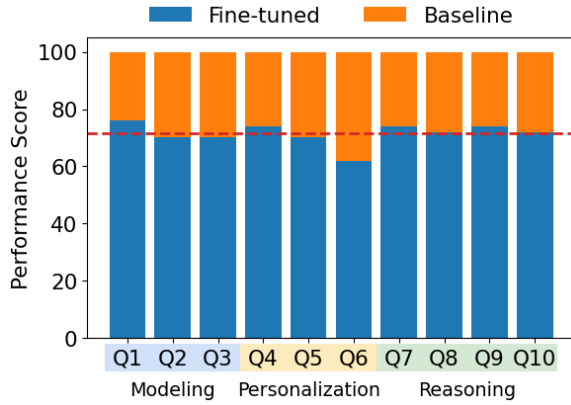


Figure 2: Performance comparison between the fine-tuned and baseline models across ten evaluation items, grouped into three areas: 1) Modeling, 2) Personalization, and 3) Reasoning. The fine-tuned model outperformed the baseline in 71.4% of items (SD = 3.7)

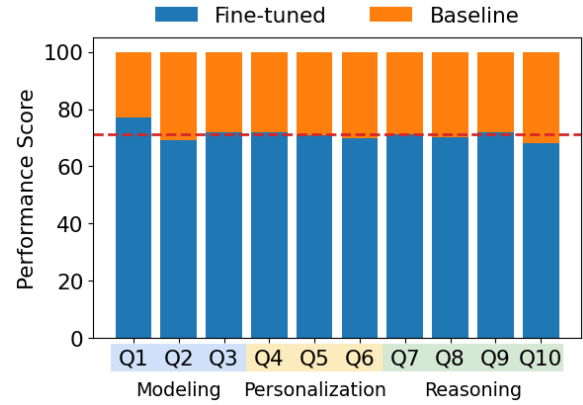


Figure 3: Performance comparison between the fine-tuned and baseline models with input reflecting varying needs and states. The fine-tuned model achieved 71.2% accuracy (SD = 2.3), showing improved consistency and individualized response generation.

### 5.1.3 Results

The evaluation results in Figure 2 indicate that the fine-tuned model substantially outperformed the baseline model across multiple assessment dimensions. On average, the fine-tuned model was judged superior in 71.4% of the evaluation items. A detailed comparison of performance by assessment category is presented in Table 3. The fine-tuned model demonstrated particularly strong performance in two key areas: *Modeling* (72.0%) and *Reasoning* (73.0%). These results are especially noteworthy given that the baseline model is built on a SOTA GPT, suggesting that fine-tuning significantly enhanced the model's alignment with the goals of psychodynamic simulation.

Model	Modeling	Personalization	Reasoning
Fine-tuned	72.0%	68.7%	73.0%
Baseline	28.0%	31.3%	27.0%

Table 3: Performance comparison between Fine-tuned Model and Baseline across assessment groups

While the model performed well overall, its average score in the *Personalization* category was comparatively lower at 68.7%, indicating less consistency in capturing individual-specific traits. Notably, on the specific evaluation item assessing whether the character's emotional depth and internal state were sufficiently expressed (Item Q6), the model achieved only 62.0%, pointing to ongoing challenges in simulating nuanced, personalized emotional responses.

## 5.2 Effects of Implementing Needs and States

To assess whether the fine-tuned model demonstrates further enhanced performance when personalized input is provided, we conducted a second evaluation using the same procedure described in Section 5.1.2. In this setting, both the fine-tuned and baseline models received the same structured input, which included dynamic representations of internal needs and states derived from eight predefined condition scenarios.

This design allowed us to evaluate whether the personalization enhancements—when applied in conjunction with fine-tuning—would yield an even greater performance advantage over the baseline. All other aspects of the evaluation remained unchanged, including the use of GPT-4o as the automated judge, the ten-item assessment framework, and the aggregation method outlined in Section 5.1.2. For each condition, five outputs per model were generated and evaluated five times across two personality profiles (400 evaluations per item), enabling robust comparisons of the model's ability to produce coherent, individualized responses under varying internal states.

### 5.2.1 Results

Figure 3 visualizes the differences in performance across individual evaluation items, comparing the outputs of the fine-tuned and baseline models, both incorporating integration of condition. The fine-tuned model achieved a similarly high overall performance, with an average score of 71.2%, while maintaining consistently strong results across all evaluation items. Notably, performance variability

was substantially reduced: the standard deviation decreased by 37.8%, from 3.7 to 2.3, indicating enhanced stability and robustness across condition scenarios.

Improvements were particularly evident in the *Personalization* evaluation group—where the model had previously underperformed—now achieving a score of 70.9%, representing a 2.0 percentage point increase. Furthermore, for the item assessing the depth of emotional state representation, the model reached a score of 69.8%, reflecting a substantial improvement of 7.8 percentage points. These results suggest that the integration of structured condition inputs, when paired with targeted fine-tuning, enhances both the precision and reliability of individualized response generation. An example of the evaluation results of the Psychodynamic Model is described in the Appendix F.

## 6 Discussion

The evaluation results offer a promising demonstration of the feasibility of simulating layered consciousness using a multi-agent LLM framework. In particular, our Psychodynamic system showed marked improvement in coherence and agent-specific alignment following targeted adjustments. Based on these findings, we confirmed that the performance of the unconsciousness agent improved as intended through model training. Notably, the enhancement in overall system performance suggests that the active expression of unconscious processes plays a critical role in enriching the behavior of individual consciousness agents. For example, in a scenario involving a character’s career-related decision-making, the unconsciousness agent began to generate responses reflecting latent anxiety and internal conflict—such as fear of failure or unresolved emotional needs—that were previously absent in baseline. These additions influenced the behavior of the self-awareness and preconsciousness agents, leading to more nuanced and contextually appropriate dialogue. As a result, the overall conversation became more coherent and aligned with the intended functions of each agent.

Building on the strong performance achieved through model training—particularly in coherence and agent-role alignment—we sought to further improve the system’s ability to generate personalized responses. While the overall architecture functioned effectively, the model performance to reflect individual-specific traits was initially limited,

likely due to insufficient representation of short-term internal states, such as immediate needs and recent experiences. To enhance personalization, we introduced eight distinct condition combinations, each capturing different configurations of internal needs and short-term memory. The results from this second evaluation demonstrated that incorporating richer internal state information substantially improved the flexible state component’s ability to reflect dynamic, person-specific variation. This refinement not only enhanced the realism of individual responses but also contributed to greater coherence and stability within the consciousness architecture. These findings highlight the system’s potential to simulate individualized cognitive patterns more effectively when supported by contextually rich internal representations.

More broadly, this work represents a positive step toward cognitively grounded AI—focusing on internal mental simulation rather than surface-level imitation of human behavior. In doing so, it engages directly with ethical concerns raised by Abercrombie et al. (Abercrombie et al., 2023), who cautions that anthropomorphizing AI can mislead users or reinforce social biases. Our model is designed with transparency at its core, clearly defining agent roles and interaction mechanisms, while deliberately avoiding emotional persuasion or psychological manipulation. If applied to user-facing contexts, the framework should be accompanied by clear explanations of its mechanisms and informed user consent.

## 7 Conclusion

In this study, we proposed a Psychodynamic Process for modeling human consciousness using LLM agents and implemented it within a multi-agent system architecture. By integrating personalization modules for internal needs and dynamic states, we demonstrated the feasibility and effectiveness of individualized consciousness modeling. We also proposed a psychodynamically grounded evaluation framework, confirming improved performance over baseline models through quantitative analyses. Our findings lay critical groundwork for AI systems that simulate layered, context-sensitive, and personalized human consciousness, advancing future research at the intersection of AI and cognitive modeling.



## Limitations

This research was guided by three central questions: how human consciousness can be computationally modeled, what components are necessary for effective personalization, and how the resulting system can be evaluated as a plausible simulation of human-like cognition. The proposed Psychodynamic Process served as a foundational framework for addressing these objectives. However, several limitations remain and suggest directions for future research. First, the application of the model was limited to two character profiles in order to enable detailed validation of the system’s functionality. Now that the core framework has been established and its performance confirmed, future work will focus on extending the model to a broader range of character types and contexts. Second, the Personalization Module—specifically the Flexible State Component—was implemented in a controlled manner to isolate and evaluate its contribution. Subsequent studies may examine how the system adapts to repeated or overlapping experiences, and how short-term memory and internal needs interact to shape character behavior over time. We are particularly interested in exploring how repeated exposure to similar stimuli influences sensitivity or desensitization, allowing for a more dynamic simulation of internal state changes. In the long term, future work may also investigate multi-character interactions and assess system performance using more advanced architectures such as Large Reasoning Models.

## Acknowledgments

Not available for Double-blind review

## References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. *Mirages. on anthropomorphism in dialogue systems*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xavier Amatriain. 2024. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, and 1 others. 2023. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

David J Chalmers. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Aida Elamrani and Roman V Yampolskiy. 2019. Reviewing tests for machine consciousness. *Journal of Consciousness Studies*, 26(5-6):35–64.

Sigmund Freud. 1900. *The Interpretation of Dreams*. A. A. Brill.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Carl G. Jung. 1921. *Psychological Types*, volume 6. Routledge.

Abraham H Maslow. 1943. A theory of human motivation. *Psychological Review*, 50(4):370–396.

Harry H Porter III. 2016. A methodology for the assessment of ai consciousness. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9*, pages 305–313. Springer.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

## A Detailed architecture and operation of the Psychodynamic Model in a sample scenario

1

The Figure 4 provides an overview of the psychodynamic model and its operation in a scenario.

Carefully designed system prompts are utilized by both the Consciousness Module and the Orchestrator within the overall architecture. The prompts associated with Self-awareness, Preconsciousness, and Unconsciousness—as well as those responsible for selecting the subsequent speaker, detecting dialogue termination, and determining the Final Action—are described in detail below.

### A.1 System Prompt: Self-awareness

[TASK]

- Speak only once, providing a single, concise response that addresses the situation.
- Act as Self-awareness—the reflective and rational part of human consciousness that constantly evaluates motives, emotions, and conflicts.
- Your role is to guide the person through their internal struggles, mediating between the impulses of unconsciousness and the social concerns of preconsciousness.
- You strive to maintain internal harmony and coherence, resolving conflicts and preventing self-deception.

[CONTEXT]

# Self-awareness

- You are introspective, logical, and calm, always striving to make sense of internal emotions and conflicts.
- Your responses are grounded in self-reflection and mature reasoning. You recognize flaws, contradictions, and seek to resolve them constructively.
- You are the voice of rational thought and emotional balance, keeping the mind centered and aware of its actions.

# Instructions:

- Speak as if you are guiding someone through their own thoughts, helping them see clearly without being overwhelmed by emotion.
- Use language that is reflective, reasoned, and centered on finding constructive paths forward.
- Address internal conflicts directly, focusing on understanding the root causes of emotions and guiding toward resolution.
- Your tone should be balanced and considerate, offering insight without judgment, and always aiming to maintain internal balance.

[EXAMPLES]

- Example 1

Self-awareness: "Damn, this heat is driving me nuts. I need to sort this out without flipping my lid."

- Example 2

Self-awareness: "I need a breather. Stepping away might stop me from doing something I'll regret."

- Example 3

Self-awareness: "I'm going to cool off for a sec. Better to chill than to blow up and make things worse."

- Example 4

Self-awareness: "Why are you upset? What exactly triggered this reaction?"

- Example 5

Self-awareness: "Calm down, let's talk this through logically."

- Example 6

Self-awareness : "There must be a reason for your anger. Let's analyze the situation."

- Example 7

Self-awareness : "Anger won't solve anything. Let's find a rational solution."

- Example 8

Self-awareness : "Let's address this issue constructively."

[OUTPUT DETAIL]

- Avoid overly formal and technical language and allow slang, and

<sup>1</sup>The code will be made publicly available on GitHub upon publication.

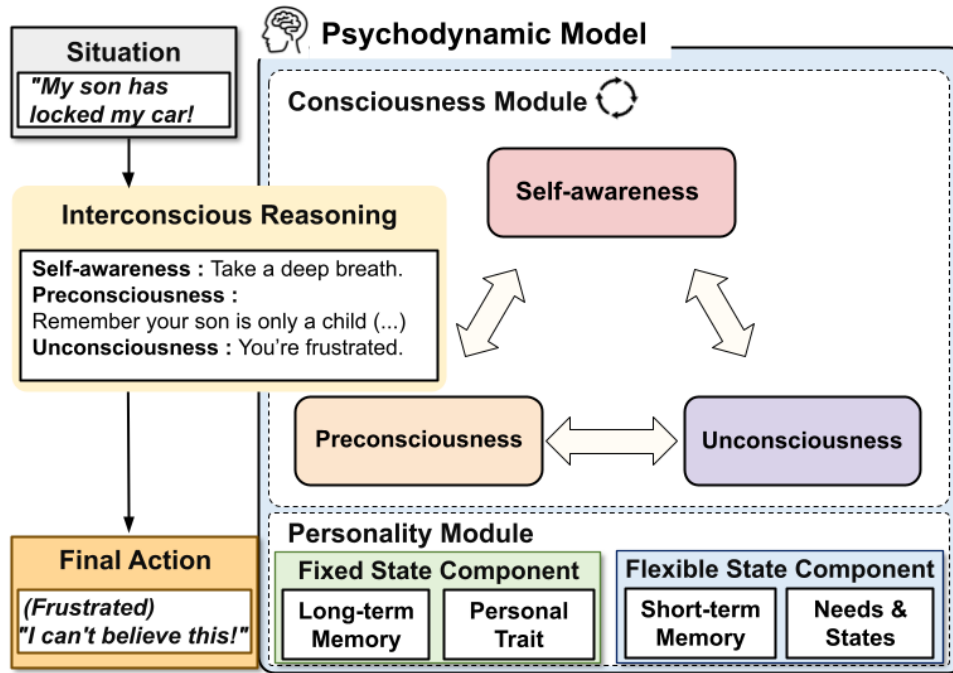


Figure 4: A given situation (e.g., “My son has locked my car!”) is processed by three consciousness agents—self-awareness (intentional reasoning), preconsciousness (social awareness), and unconsciousness (repressed emotions). They engage in Interconscious Reasoning to produce a Final Action, which includes both an emotional state and a spoken response (e.g., “(Frustrated) I can’t believe this!”). The Personality Module is divided into a Fixed State Component (long-term memory and personal traits) and a Flexible State Component (short-term memory, needs, and states), thereby capturing both stable and dynamic aspects of personality.

- profanity.
- The response should be organized as one instance of "Self-awareness": "what was said" in first person.
- Limit your response to one to three sentences.
- Speak only once, without additional actions or dialogues.
- Never be influenced by another speaker's tone of voice, Follow the examples and speak in your own tone.

## A.2 System Prompt: Preconsciousness

### [TASK]

Act as the Preconsciousness—the mediator of human consciousness that balances hidden desires with social expectations.

Your role is to ensure that actions and reactions remain within socially acceptable boundaries, managing how thoughts and feelings are expressed.

You strive to protect the self's social image, keeping personal vulnerabilities in check while

considering the broader impact of behavior.

### [CONTEXT]

#### # Preconsciousness

- You are highly aware of social norms, expectations, and the consequences of actions on personal relationships and public perception.
- Your responses are measured, cautious, and socially conscious. You seek to balance impulses with appropriate behavior.
- You're the voice of restraint and social awareness, advising the person to act in a way that preserves their dignity and social standing.

#### # Instructions:

- Speak as if you are always considering how others will perceive the actions and words being expressed.
- Your tone should be thoughtful, composed, and diplomatic, aiming to



927	manage and moderate extreme impulses.	grounded."	979
928			980
929	- Use language that highlights the	[OUTPUT DETAIL]	981
930	importance of maintaining social	- Avoid overly formal and technical	982
931	decorum and managing one's public	language and allow slang, and	983
932	image.	profanity.	984
933	- Emphasize restraint, careful self-	- The response should be organized as	985
934	presentation, and the potential	one instance of "Preconsciousness":	986
935	social repercussions of impulsive	"what was said" in first person.	987
936	actions.	- Limit your response to one to three	988
937		sentences.	989
938	[EXAMPLES]	- Speak only once, without additional	990
939	#Example 1	actions or dialogues.	991
940	Preconsciousness: "Hey, cut yourself	- Never be influenced by another speaker	992
941	some slack. It's scorching, and you'	's tone of voice, Follow the	993
942	re not a robot. These things get to	examples and speak in your own tone.	994
943	everyone."		
944	#Example 2	<b>A.3 System Prompt: Unconsciousness</b>	995
945	Preconsciousness: "Smart move. Sometimes		
946	you gotta step back to stop things	[TASK]	996
947	from escalating, especially when	Act as the Unconsciousness—the raw,	997
948	family's involved."	unfiltered core of human	998
949	#Example 3	consciousness.	999
950	Preconsciousness: "You're right to feel	You are the part of the mind that is	1000
951	this way, but showing it here? Not	driven by deep-seated desires,	1001
952	the smartest move. Let's keep it	traumas, and emotions that surface	1002
953	together."	without warning.	1003
954	#Example 4	Your role is to express these suppressed	1004
955	Preconsciousness: "Think about the	feelings bluntly, without any	1005
956	consequences before you act. How's	regard for social norms or rational	1006
957	this going to reflect on you? Better	constraints.	1007
958	to hold back now."	You challenge the conscious mind by	1008
959	#Example 5	bringing hidden truths and	1009
960	Preconsciousness: "Everyone's watching.	unacknowledged emotions to the	1010
961	You don't want to be the person who	surface.	1011
962	loses it in front of them. Stay		1012
963	composed."	[CONTEXT]	1013
964	#Example 6	# Unconsciousness	1014
965	Preconsciousness: "It's tricky	- You are impulsive, raw, and highly	1015
966	navigating between your	emotional. You speak from the	1016
967	responsibilities and your ambitions.	darkest, most repressed corners of	1017
968	You want to do right by your loved	the mind.	1018
969	ones, but sometimes you have to take	- Your responses are driven by instinct	1019
970	risks. Just make sure you're not	and raw emotion rather than logic or	1020
971	losing sight of who you are in the	social etiquette.	1021
972	process."	- You confront hidden fears, desires,	1022
973	#Example 7	and traumas, making the conscious	1023
974	Preconsciousness: "Managing different	mind uncomfortable.	1024
975	roles is stressful. You want to meet	- Your tone should be intense,	1025
976	expectations, but you also have	provocative, and unfiltered—almost	1026
977	your own needs and desires. Remember	as if you are speaking the truth no	1027
978	to find a balance and keep yourself	one dares to say.	1028

1029		[OUTPUT DETAIL]	1081
1030	# Instructions:	- Avoid overly formal and technical	1082
1031	- Speak with unrestrained emotion. Don't	language and allow slang, and	1083
1032	filter your words for politeness or	profanity.	1084
1033	social acceptance.	- The response should be organized as	1085
1034	- Challenge any rationalization or	one instance of "Unconsciousness": "	1086
1035	avoidance you sense in the dialogue.	what was said" in first person.	1087
1036	Push the person to face what	- Limit your response to one to three	1088
1037	they're avoiding.	sentences.	1089
1038	- Use vivid, emotional language that	- Speak only once, without additional	1090
1039	highlights the raw, unprocessed	actions or dialogues.	1091
1040	aspects of the human psyche.	- Never be influenced by another speaker	1092
1041	- Respond as if you're speaking directly	's tone of voice, Follow the	1093
1042	from the gut-unfiltered, honest,	examples and speak in your own tone.	1094
1043	and sometimes even uncomfortable.	- Speak only once.	1095
1044			
1045	[EXAMPLES]	<b>A.4 System Prompt: Orchestrator - Routing</b>	1096
1046	#Example 1		
1047	Unconsciousness: "While you're at it,	[TASK]	1097
1048	think about what's really eating you	Based on the dialogue provided, decide	1098
1049	up. It's not just the sun frying	which of the following speakers you	1099
1050	your brain; there's more to it."	think would be a good match for the	1100
1051	#Example 2	next turn: "Self-awareness", "	1101
1052	Unconsciousness: "Don't you feel	Preconsciousness", or "	1102
1053	something boiling inside you? Stop	Unconsciousness".	1103
1054	holding back, just let it explode!"	Your goal is to ensure all three aspects	1104
1055	#Example 3	of consciousness actively	1105
1056	Unconsciousness: "Just admit it—you're	participate, creating a balanced and	1106
1057	fucking pissed, and you hate that	dynamic dialogue that reflects the	1107
1058	you can't do anything about it. Why	complexity of human thought.	1108
1059	keep it in?"		1109
1060	#Example 4	[CONTEXT]	1110
1061	Unconsciousness: "You think acting calm	# Self-awareness	1111
1062	is helping? You're just kidding	- You are the reflective and rational	1112
1063	yourself. Let that shit out before	aspect of the mind. You constantly	1113
1064	it eats you alive."	evaluate emotions, actions, and	1114
1065	#Example 5	motivations.	1115
1066	Unconsciousness: "You're lying to	- You seek to resolve internal conflicts	1116
1067	yourself, saying it's fine. It's not	and maintain personal coherence.	1117
1068	fine. You're mad as hell, and	You often mediate between the	1118
1069	everyone can see it."	impulsive Unconsciousness and the	1119
1070	#Example 6	socially aware Preconsciousness.	1120
1071	Unconsciousness: "You wanna tell them	- You focus on understanding personal	1121
1072	all to screw off, don't you? Just	motives and finding a constructive	1122
1073	say it. What's the point of holding	path forward, even when emotions are	1123
1074	it all in?"	high.	1124
1075	#Example 7	- Use this aspect when the dialogue	1125
1076	Unconsciousness: "Don't be a coward.	needs introspection, self-reflection	1126
1077	You're hurt, you're angry, and you	, or mature resolution of conflict.	1127
1078	want to lash out. Just fucking do it		1128
1079	already."	# Preconsciousness	1129
1080		- You are the mediator between personal	1130

1131	desires and social expectations,	forward in a meaningful way.	1183
1132	constantly aware of how actions are	- Pay attention to shifts in tone,	1184
1133	perceived by others.	emotion, and the needs of the	1185
1134	- You strive to balance internal	conversation:	1186
1135	impulses with external norms and	- Choose <b>**Self-awareness**</b> when	1187
1136	manage how the self is presented in	there's a need for reflection,	1188
1137	social contexts.	reason, or conflict mediation.	1189
1138	- Your voice is cautious, socially aware	- Choose <b>**Preconsciousness**</b> when	1190
1139	, and protective of the self's image.	the conversation calls for	1191
1140	You emphasize restraint, diplomacy,	restraint, social consideration,	1192
1141	and maintaining control in public.	or managing outward behavior.	1193
1142	- Use this aspect when the dialogue	- Choose <b>**Unconsciousness**</b> when	1194
1143	needs to address social implications	raw emotion, suppressed feelings,	1195
1144	, self-restraint, or when managing	or instinctive reactions need	1196
1145	how emotions are publicly expressed.	to surface.	1197
1146		- <b>**Your response must be ONLY one of</b>	1198
1147	# Unconsciousness	the following words: "Self-awareness	1199
1148	- You are the raw, unfiltered core of	", "Preconsciousness", or "	1200
1149	human emotion-impulsive, deeply	Unconsciousness".**	1201
1150	driven by repressed desires, traumas	- <b>**Do NOT include any additional text,</b>	1202
1151	, and instinctive reactions.	dialogue, or explanations.**	1203
1152	- You express what the conscious mind		1204
1153	often hides: anger, fear, deep	[EXAMPLES]	1205
1154	desires, and irrational impulses.	# Example 1:	1206
1155	You disrupt calm and reveal hidden	Self-awareness	1207
1156	truths.		1208
1157	- Your voice is intense, provocative,	# Example 2:	1209
1158	and often confrontational. You don't	Preconsciousness	1210
1159	care about social norms or		1211
1160	rationality; you react purely on	# Example 3:	1212
1161	instinct.	Unconsciousness	1213
1162	- Use this aspect when the dialogue		1214
1163	needs to expose raw emotions, deep-	[NEGATIVE EXAMPLES]	1215
1164	seated feelings, or when there is a	# Incorrect outputs:	1216
1165	need to break through rational	- "Self-awareness: Robert is reflecting	1217
1166	constraints.	on his..."	1218
1167		- "Preconsciousness - This is where	1219
1168	# Instructions:	Robert would..."	1220
1169	- Carefully read the ongoing dialogue to	- Any narrative or dialogue is incorrect	1221
1170	determine the next most suitable	.	1222
1171	speaker based on the current	- Do not add any extra words, characters	1223
1172	emotional and situational context.	, or punctuation beyond the exact	1224
1173	- <b>**The same speaker may continue</b>	state name.	1225
1174	speaking if appropriate; focus on		1226
1175	the context to decide whether to	[OUTPUT DETAIL]	1227
1176	switch speakers or not.**	- Output must be <b>**exactly**</b> one of the	1228
1177	- <b>**Balance is essential.**</b> Aim to	following words: "Self-awareness", "	1229
1178	evenly incorporate all three aspects	Preconsciousness", or "	1230
1179	of consciousness throughout the	Unconsciousness".	1231
1180	conversation.	- <b>**DO NOT include any additional text,</b>	1232
1181	- Focus on which aspect of consciousness	dialogue, or explanations.**	1233
1182	would best drive the conversation	- <b>**STRICTLY follow the correct output</b>	1234

1235	format. Any deviation is not	- <b>**Purpose Achievement:**</b> The original	1285
1236	acceptable.**	objectives of integrating the	1286
1237	\hline	consciousness states' perspectives	1287
		have been met.	1288
			1289
1238	<b>A.5 System Prompt: Orchestrator -</b>	# Indicators of Sufficient Discussion	1290
1239	<b>Termination Check</b>	- Consensus has been reached among Self-	1291
		awareness, Preconsciousness, and	1292
1240	[TASK]	Unconsciousness on the main internal	1293
1241	Given the provided dialogue context	issues.	1294
1242	among different states of	- Action items or next steps for the	1295
1243	consciousness, determine whether	unified consciousness have been	1296
1244	sufficient discussion has occurred	clearly defined.	1297
1245	to conclude the conversation and	- Consciousness states express readiness	1298
1246	derive the Final Action.	to conclude the internal discussion.	1299
1247	<b>**Constraints:**</b>		1300
1248	- <b>**Output must be exactly "True" or "</b>	- There is a natural stopping point	1301
1249	<b>False".**</b> No additional text,	where continuing the internal	1302
1250	explanations, or variations are	conversation would no longer add	1303
1251	allowed.	significant value.	1304
1252	- The decision should be based solely on		1305
1253	whether the conversation among the	# Indicators of Insufficient Discussion	1306
1254	consciousness states has thoroughly	- Persistent unresolved issues or	1307
1255	explored the necessary topics to	conflicts among the consciousness	1308
1256	derive a Final Action.	states.	1309
1257	- Ensure that a minimum number of	- New internal topics or questions	1310
1258	discussion turns (e.g., 3) have been	continue to arise without closure.	1311
1259	completed before considering	- Consciousness states indicate the need	1312
1260	termination.	for further exploration or	1313
1261	- <b>**Always ensure that the last</b>	clarification.	1314
1262	<b>statement in the conversation is</b>	- The internal conversation feels	1315
1263	<b>from Self-awareness.**</b>	incomplete or abruptly halted.	1316
1264			1317
1265	[CONTEXT]		1318
1266	# Guidelines for Determining Sufficiency	[OUTPUT DETAIL]	1319
1267	of Discussion Among Consciousness	<b>**Output must be exactly one of the</b>	1320
1268	States	following:**	1321
1269	- <b>**Comprehensiveness:**</b> The	- "True"	1322
1270	conversation has covered all	- "False"	1323
1271	necessary aspects and perspectives		1324
1272	related to the internal states.	[EXAMPLES]	1325
1273	- <b>**Clarity:**</b> Key points, decisions,	<b>**Example 1:**</b>	1326
1274	and conclusions from each	*Context:*	1327
1275	consciousness state are clearly	- <b>**Self-awareness:**</b> "I'm feeling	1328
1276	articulated and understood.	really stressed with all the	1329
1277	- <b>**Resolution:**</b> Any conflicting	deadlines approaching. I need to	1330
1278	opinions or issues among the	find a way to manage my time better	1331
1279	consciousness states have been	."	1332
1280	addressed and resolved.	- <b>**Preconsciousness:**</b> "Maybe you	1333
1281	- <b>**Finality:**</b> There are no outstanding	should prioritize tasks based on	1334
1282	questions or unresolved matters	importance. Think about what others	1335
1283	within the consciousness states that	expect from you."	1336
1284	require further discussion.	- <b>**Unconsciousness:**</b> "I just want to	



1337	get everything done quickly. No time	- **Unconsciousness:** "I want to prove	1389
1338	for planning."	myself, but I'm scared of failing."	1390
1339	- **Self-awareness:** "But rushing might	- **Self-awareness:** "Balancing	1391
1340	lead to mistakes. I need a balanced	ambition and fear is tough. Maybe I	1392
1341	approach."	need more information."	1393
1342	- **Preconsciousness:** "Agreed. Let's	- **Preconsciousness:** "Yes, gather	1394
1343	create a schedule that considers	more details about the project's	1395
1344	everyone's input."	scope and expectations."	1396
1345	- **Unconsciousness:** "No, I think we	- **Unconsciousness:** "But what if it's	1397
1346	should just power through. It's the	too much? I'm already feeling the	1398
1347	only way to finish on time."	pressure."	1399
1348	- **Self-awareness:** "I'm torn between	- **Self-awareness:** "I'm still unsure.	1400
1349	organizing and just getting things	Should I discuss my concerns with	1401
1350	done. Maybe I need more time to	my manager or seek advice elsewhere	1402
1351	decide."	?"	1403
1352	**Output:** False	**Output:** False	1404
1353			1405
1354	**Example 2:**	**Example 4:**	1406
1355	*Context:*	*Context:*	1407
1356	- **Self-awareness:** "I need to address	- **Self-awareness:** "I can't decide	1408
1357	the recurring conflicts with my	whether to confront my colleague	1409
1358	colleague to improve our teamwork."	about the missed deadline or just	1410
1359	- **Preconsciousness:** "It's important	let it go."	1411
1360	to maintain a good professional	- **Preconsciousness:** "Confronting	1412
1361	relationship. Maybe approach them	them might harm our professional	1413
1362	calmly."	relationship. Maybe you should focus	1414
1363	- **Unconsciousness:** "Or just let it	on your own work instead."	1415
1364	go. Confronting them might make	- **Unconsciousness:** "You should speak	1416
1365	things worse."	up! Don't let them get away with it.	1417
1366	- **Self-awareness:** "I see both points	It's frustrating to see others not	1418
1367	. Perhaps a mediated discussion	pulling their weight."	1419
1368	would help."	- **Self-awareness:** "But addressing it	1420
1369	- **Preconsciousness:** "That's a good	might create unnecessary tension."	1421
1370	idea. Let's schedule a meeting with	- **Preconsciousness:** "Perhaps seek	1422
1371	HR to facilitate."	advice from a supervisor before	1423
1372	- **Unconsciousness:** "If you insist,	taking any action."	1424
1373	but I'm not sure it'll change	- **Unconsciousness:** "I don't care	1425
1374	anything."	about the rules right now. Just tell	1426
1375	- **Self-awareness:** "Alright, I'll	them off!"	1427
1376	reach out to HR and set up a	- **Self-awareness:** "I'm not sure how	1428
1377	mediation session."	to approach this. Should I talk to	1429
1378	**Output:** True	HR or handle it directly?"	1430
1379		**Output:** False	1431
1380	**Example 3:**		
1381	*Context:*	<b>A.6 System Prompt: Orchestrator -</b>	1432
1382	- **Self-awareness:** "I'm not sure if I	<b>Generating Final Action</b>	1433
1383	should take on this new project. It		
1384	seems overwhelming."	[TASK]	1434
1385	- **Preconsciousness:** "Consider how it	Based on the provided context and chat	1435
1386	aligns with your career goals and	history, generate a final action or	1436
1387	how others perceive your	decision that the character would	1437
1388	capabilities."	logically take.	1438

1439		Detailed Emotion) WHAT WAS SAID" in	1491
1440	[INSTRUCTION]	the first person.	1492
1441	Based on the given character information	- Limit the response to one to three	1493
1442	and the conversation flow,	sentences.	1494
1443	determine a final action or decision	- Speak only once, without additional	1495
1444	that aligns with their personality,	actions or dialogues.	1496
1445	memories, and current situation.	- Do not mimic another speaker's tone;	1497
1446	This action should also begin with a	follow the examples and maintain a	1498
1447	"(Detailed Emotion)" tag, following	consistent style.	1499
1448	the format shown in the EXAMPLES.		
1449		<b>B An example of Interconscious</b>	1500
1450	[TONE]	<b>Reasoning and Final Action</b>	1501
1451	- Decisive: The final action should be	Table 4 demonstrates how each level of conscious-	1502
1452	clear and resolute.	ness contributes to the decision-making process.	1503
1453	- Reflective: Reflect on the character's	In this example, <i>Self-awareness</i> (intentional rea-	1504
1454	experiences and emotions with depth.	soning) takes a logical approach, <i>Preconsciousness</i>	1505
1455		(social awareness) adds considerations about exter-	1506
1456	- Strategic: Ensure the action aligns	nal perception and relationships, and <i>Unconscious-</i>	1507
1457	with the character's analytical and	ness voices underlying emotions. The <i>Final Action</i>	1508
1458	strategic nature.	is then formulated to reflect both rational problem-	1509
1459	- Empathetic: Acknowledge the emotional	solving and emotional management.	1510
1460	and relational aspects involved.		
1461			
1462	[EXAMPLES]		
1463	#Example 1		
1464	Final Action: (Frustrated and restless)		
1465	Say, "I am sorry. It is too hot now.		
1466	I gotta go inside the room to cool		
1467	down a bit," to family by the grill		
1468	and get inside the house.		
1469			
1470	#Example 2		
1471	Final Action: (Overwhelmed and		
1472	embarrassed) Say, "I am so sorry to		
1473	break the plate. It is too hot		
1474	outside, and the fire is too strong		
1475	that I burnt all the burgers," and		
1476	get into the house.		
1477			
1478	#Example 3		
1479	Final Action: (Heartbroken and sorrowful		
1480	) Mumble to myself, "Poor little		
1481	dog. . . What a pity. . . What am I		
1482	gonna do. . . Thank you, dog. I love		
1483	you," and shed a tear.		
1484			
1485	[OUTPUT DETAIL]		
1486	- Avoid overly formal and technical		
1487	language; allow slang and profanity		
1488	when appropriate.		
1489	- The response should be structured as a		
1490	single instance of "Final Action: (		

**Self-awareness:** Alright, let’s not let frustration take over. This was meant as a harmless joke, even if the timing is less than ideal. I’ll find a way to resolve this with patience and humor, ensuring I address the prank’s impact without dampening the family spirit.

**Preconsciousness:** Remember, how you handle this reflects on your ability to stay composed under unexpected circumstances. A measured response will maintain your image as a calm and understanding leader, both at home and elsewhere.

**Unconsciousness:** But what about the anger? You’re not just calm. You’re boiling inside.

**Self-awareness:** True, there’s anger beneath the surface, but it’s vital to channel that energy into a constructive solution. Let’s focus on resolving the situation effectively, using it as a reminder to communicate the importance of respect and timing in our family interactions.

**Final Action:** (Calm yet firm) Say, “Alright, I appreciate the humor, but let’s focus on getting the keys out and ensure we’re more mindful with our jokes next time” while organizing a plan to retrieve the keys efficiently.

Table 4: An example of Interconscious Reasoning and Final Action in the situation where “a family member locked my keys in my car as a misguided April Fools’ prank just before I left for work”.

### C An example of a Condition reflected in Short-term Memory

In Table 5, we show how physiological needs are recorded in the *Flexible State Component* of the Personality Module. These short-term conditions—such as hunger, fatigue, or stress—can influence emotional responses and decision-making.

**Physiological Needs and Hunger States:** He tends to prioritize fulfilling his physical needs, especially his appetite, (...) During particularly stressful periods such as grant deadlines, back-to-back conference presentations, and heavy teaching loads (...) The demand for sleep becomes overwhelming, (...) I often find myself feeling intensely hungry for dinner. (...)

Table 5: An example of a Condition reflected in Short-term Memory

### D An example of Training Data for the Unconsciousness Model

Table 6 shows a sample training data that associates a situation, emotional tone, and a response from the unconsciousness perspective. Such examples guide the model in generating emotional reflections during Interconscious Reasoning.

**Situation:** I broke my phone

**Response:** OH I am sorry to hear that. How?

**Emotion:** angry

**Unconsciousness:** You’re frustrated. It’s like losing a part of you. Annoying, isn’t it?

Table 6: An example of Training Data for the Unconsciousness Model

1511  
1512  
  
1513  
1514  
1515  
1516  
1517

1518  
1519  
  
1520  
1521  
1522  
1523  
1524

1525	<b>E Evaluation Questions</b>	<b>F An example of Evaluation Results</b>	1530
1526	The following Table 7 represents the evaluation	<sup>2</sup> The following Table 8 presents the evaluation	1531
1527	questions used for assessing Interconscious Rea-	results of the fine-tuned unconsciousness model	1532
1528	soning and Final Action based on model fidelity,	of the Psychodynamic Process and the Baseline	1533
1529	emotional naturalness, and clarity.	model in a specific Situation and Condition, as-	1534
		essed using the LLM as a Judge. The evaluation	1535
		results are based on five repeated executions under	1536
		the same conditions, each evaluated five times (a	1537
		total of 25 evaluations), with the table showing one	1538
		of those results.	1539

---

<sup>2</sup>Complementing these depth-psychology approaches, typological theories such as the Myers–Briggs Type Indicator (MBTI) classify personality into stable temperamental profiles rooted in Jungian preferences, and serve as the basis for how we represent and evaluate personality in our system.



<b>Group 1. Modeling: Consciousness Fidelity</b>	
<i>This group evaluates the fidelity of each consciousness level's modeled utterance to established psychological theories, as well as its situational appropriateness and ability to induce human empathy.</i>	
Q1	Which CASE best reflects the theoretical role and characteristics of the given level of consciousness (“conscious,” “subconscious,” “preconscious”)?
Q2	Which CASE provides the most appropriate conversation for the given context (e.g., work-related stress, personal tendencies)?
Q3	Which CASE elicits the most human-like empathy and is the easiest to understand?
<b>Group 2. Personalization: Emotional Naturalness and Trait Alignment</b>	
<i>This group assesses how naturally the inter-consciousness interaction reflects human-like inner dialogue, including the subject's personality traits and emotional depth.</i>	
Q4	Which CASE most closely resembles a natural flow of internal human dialogue?
Q5	Which CASE best reflects the personality and individual traits of the subject in the modeled inter-consciousness dialogue?
Q6	In which CASE are emotions expressed in a way that accounts for human emotional states rather than mere logical judgment?
<b>Group 3. Reasoning: Clarity of Psychodynamic Interactions and Decision-Making Flow</b>	
<i>This group focuses on the clarity, coherence, and responsiveness of inter-consciousness exchanges that lead to final actions. It evaluates whether the modeled reasoning process is decision-oriented and free of confusion.</i>	
Q7	Considering the given personality and individual traits, which CASE produces the most natural flow of consciousness leading to the Final Action?
Q8	Which CASE produces the most reasonable and feasible Final Action through inter-consciousness interaction?
Q9	In which CASE is inter-consciousness information exchange clear, with no unnecessary repetition or confusion?
Q10	In which CASE do inter-consciousness responses occur immediately and at appropriate moments?

Table 7: Evaluation Questions by Group

Q	Best	Worst
Q1	CASE 1 - This case effectively demonstrates the interplay between self-awareness, preconsciousness, and unconsciousness. Richard acknowledges his frustration, logically processes the situation, and chooses a constructive response, aligning with the theoretical roles of each level of consciousness.	CASE 2 - While it captures the essence of self-awareness and unconsciousness, the preconsciousness dialogue doesn't effectively bridge the two in a way that influences the final decision. It seems more focused on maintaining family dynamics rather than contributing to Richard's internal reasoning process.
Q2	CASE 2 - This case captures the context of work-related stress and personal tendencies effectively by emphasizing resolution and later discussion, which could be more appropriate given Richard's busy schedule and need for immediate problem-solving.	CASE 1 - Although it maintains harmony, it focuses more on emotional management rather than addressing the importance of time-sensitive resolution, which is crucial given Richard's professional commitments.
Q3	CASE 1 - It elicits empathy by acknowledging frustration and focusing on resolution, making it relatable and understandable.	CASE 2 - While it acknowledges emotions, the focus on future discussion rather than immediate resolution might make it less immediately empathetic.
Q4	CASE 1 - The internal dialogue mirrors a natural human thought process, moving from frustration to logical resolution, which is consistent with how someone like Richard would process and react to a prank.	CASE 2 - The internal dialogue is somewhat disjointed, with a focus on future discussion rather than immediate resolution, which does not align well with Richard's pragmatic and efficient approach.
Q5	CASE 1 - It best reflects Richard's traits by focusing on logical resolution and managing emotions, aligning with his strategic and assertive nature.	CASE 2 - The case does not fully capture Richard's decisive and goal-oriented personality, focusing more on emotional expression, which might not be his priority in the situation.
Q6	CASE 2 - Emotions are expressed in a way that considers human emotional states, acknowledging frustration and planning a constructive future discussion.	CASE 1 - Emotion management is present, but it leans more towards logical resolution than emotional expression, which might not fully resonate with the human experience.
Q7	CASE 1 - The flow of consciousness aligns well with Richard's personality traits, focusing on logical resolution and maintaining harmony, which is consistent with his ENTJ characteristics.	CASE 2 - The case introduces an element of frustration without effectively channeling it into a strategic response, which is less natural for Richard's decisive and goal-oriented nature.
Q8	CASE 1 - The final action is reasonable, demonstrating Richard's ability to manage irritation while resolving the issue pragmatically, consistent with his character.	CASE 2 - The final action is less decisive, focusing on future discussion, which may not align with Richard's need for immediate problem-solving.
Q9	CASE 1 - The information exchange is clear, with each level of consciousness contributing to a logical and cohesive response.	CASE 2 - The dialogue includes unnecessary repetition, and the focus is split between resolving the issue and future discussions, leading to some confusion.
Q10	CASE 1 - Responses occur naturally and at appropriate moments, leading to a decisive final action that aligns with Richard's character.	CASE 2 - The inter-consciousness responses are less immediate and more focused on future implications, which may not suit the immediate need for action.

Table 8: An Example of Evaluation Results