Feeling First, Speaking Second: A Dual-Process Cognitive-Affective Architecture for LLM Agents

Nicolò Buscaroli, Fabio Tamburini
University of Bologna, FICLIT - University of Bologna
nicolo.buscaroli@studio.unibo.it, fabio.tamburini@unibo.it

Abstract

Current Large Language Models (LLMs) demonstrate exceptional generative capabilities but lack a coherent "inner life", failing to model the dynamics of emotion regulation essential for believable affective behavior. Constrained by statelessness and a lack of theoretical grounding, standard models struggle to maintain psychological depth. To address this, we propose a computational cognitive-affective architecture grounded in Dual-Process Theory. Our system computationally distinguishes between visceral emotional reaction (*Appraisal*) and strategic verbal expression (*Formulation*), effectively operationalizing the gap between "feeling" and "saying". This modular design allows agents to embody specific personas by integrating long-term memory, dynamic emotional states, personality and goals. We evaluated the system in simulated narrative scenarios using an LLM-as-a-judge protocol. We frame this system as a computational experiment to investigate the mechanics of artificial affect. Results confirm the feasibility of simulating a coherent and believable inner emotional monologue. However, analysis in high-pressure scenarios reveals a *rational bias* where strategic planning can override this emotional authenticity. These findings contribute to Computational Affective Science by demonstrating that while cognitive sequentiality successfully generates inner lives, enforcing affective primacy in the decision cycle is critical to prevent excessive rational regulation.

Keywords: LLM Agent, Cognitive-affective Architecture, Dual-Process Theory

1. Introduction

Computational Affective Science (CAS) currently requires models that move beyond surface-level affect classification to simulation systems replicating the underlying neural and cognitive processes generating and regulating these states. True affective modeling must account for the dynamics of how emotions emerge, persist, and influence behavior, rather than treating them as static labels assigned to an utterance. To achieve this, we must bridge the gap between statistical language generation and cognitive-affective architectures that ground language in a coherent psychological reality.

Historically, many computational affective models were implemented within symbolic or rule-based architectures, including frameworks like EMA (EMotion and Adaptation)(Marsella and Gratch, 2009) and FAtiMA (Fearnot AffecTIve Mind Architecture)(Mascarenhas et al., 2022), which operationalize appraisal theories, such as OCC (Ortony et al., 1988), in terms of explicit reasoning structures akin to BDI (Belief-Desire-Intention). These approaches offer high interpretability and consistency but require the extensive manual construction of rules and appraisal mappings, and they lack innate mechanisms to process open-ended natural language without substantial preprocessing or semantic grounding. Conversely, the advent of LLMs enabled nuanced text generation and emergent behaviors. However, recent investigations highlight that LLMs fail to maintain thematic coherence and emotional resonance (Alharthi, 2025). This "emotional flatness" stems from the lack of an underlying cognitive structure capable of managing intentionality and persistent affective states. Native LLMs operate as stateless (meaning they process each interaction in isolation without any built-in memory of past exchanges), reactive engines devoid of theoretical grounding, resulting in the inability to retain a coherent thread of facts and personality over extended narrative arcs (Freiknecht and Effelsberg, 2020), and preventing the simulation of a credible "inner life" necessary for deep affective interaction.

To address these limitations, this work presents a Cognitive-Affective Architecture that endows the agent with a persistent "inner life" and internalizes the process of emotion regulation. The system is designed to embody specific personas by integrating a long-term memory, distinct personality traits, dynamic emotional states and hierarchical goals. Central to this approach is the constraint that the agent must process an "internal monologue" before generating verbal output. By decomposing the communicative act, we force the model to engage in a distinct *Appraisal* phase (simulating the visceral, immediate emotional reaction) followed by a *Formulation* phase, where that impulse is rationally filtered and strategically modulated to align with both the social context and the agent's active pursuit of its goals. Ultimately, by explicitly modeling these preceding cognitive steps, the pipeline aims to generate a final verbal output that is qualitatively enhanced and deeply grounded in a coherent psychological reality.

We validate the architecture in interactive narra-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 74–92
16 May 2026.

tive scenarios because they are valid testbeds for emotion dynamics: agents must maintain long-term coherence while pursuing narrative goals, exactly the conditions under which human emotion regulation is most visible.

The primary contribution of this paper is twofold. First, we present the design and validation of a comprehensive Cognitive-affective Architecture built from the ground up to endow LLMs with a persistent "inner life". Second, we frame this system as a computational experiment to investigate the mechanics of artificial affect. Crucially, our results confirm that it is possible to successfully simulate a coherent inner emotional monologue and operationalize emotion regulation within an LLM Architecture. However, we also demonstrate that maintaining this plausibility is subject to specific architectural constraints: specifically, we show that without enforcing a structural affective primacy over rationality, agents risk collapsing into a *rational bias*. This highlights the critical role of cognitive sequentiality in the design of next-generation affective systems.

2. Theoretical Background and Related Work

While CAS has traditionally relied on symbolic or rule-based appraisal models, such as FAtiMA (Mascarenhas et al., 2022), to ensure theoretical fidelity, the broader field of agent simulation has shifted towards Generative Agents driven by LLMs (Park et al., 2023). While these modern generative approaches excel in verbal fluency and behavioral plausibility, they fundamentally lack intrinsic mechanisms to simulate the dynamics and persistence of emotional states. Without structural grounding, LLMs remain stateless and reactive, failing to model credible emotion dynamics rather than exhibiting a coherent psychological progression.

To address these limitations, we adopt the CoALA (Cognitive Architectures for Language Agents) framework (Sumers et al., 2024) as our taxonomic foundation. CoALA allows us to structure an agent that is "cognitive" rather than merely communicative by formally distinguishing between external actions (communication) and internal actions (retrieval, reasoning). This distinction is vital for implementing a decision-making cycle where affective processing acts as a distinct internal step before any external output is generated.

We operationalize the agent's "inner life" by drawing on Dual-Process Theory (Kahneman, 2011) and Appraisal Theory (Lazarus, 1991; Scherer et al., 2001). We model the causal process of emotion not as a single inference, but as a sequence. The cycle begins with an *Appraisal* phase (*System 1*), simulating a visceral, immediate emotional reaction, followed by a *Formulation* phase (*System 2*), where this impulse is rationally filtered. This sequential architecture effectively implements a computational model of emotion regulation, closely mirroring Gross's process model of response modulation (Gross, 1998), where expression is regulated after the emotional impulse is generated.

While recent frameworks have successfully integrated inner thought reasoning to enhance role-playing consistency (Xu et al., 2025) and utilized Appraisal Theory to improve emotion prediction accuracy (Hong et al., 2025), they focus primarily on performance optimization rather than simulating the regulatory friction between visceral affect and strategic intent.

Finally, while recent multi-agent frameworks, such as Ego-Superego models (Magee et al., 2024) or Transactional Analysis architectures (Zamojska and Chudziak, 2025), primarily model internal conflict as a mechanism of social censorship or impulse control, our approach reframes this tension through the lens of proactive agency. We distinguish our architecture by introducing a strategic planning component that actively pursues the agent's goals. In our unified cycle, this proactive plan dynamically clashes with the agent's immediate visceral appraisal. The cognitive struggle is thus redefined: it is not just a negotiation between an impulse and a moral filter, but a friction between visceral affect and high-level goals.

3. Cognitive Architecture Design

To overcome the stateless nature of standard LLMs and simulate a coherent psychological entity, we designed a modular architecture grounded in the CoALA framework (Sumers et al., 2024), as illustrated in Figure 1. The system operates as a "Local-First" agent, executing all cognitive processes on consumer hardware (via Ollama[1]) to ensure data sovereignty and privacy. The architecture is composed of three main pillars: a structured Action Space, a Memory System, and a Goal Framework.

3.1. Action Space: Internal vs. External

Unlike chatbots that map user input directly to an output response, our agent utilizes a distinct *Action Space*. This ensures that the final output is not a mere continuation of the text, but the execution of simulated cognitive processes that ground the agent's behavior in its internal affective reality. We distinguish between:

- **Internal Actions (Reasoning & Retrieval):** Cognitive operations invisible to the user that modify the agent's internal state or retrieve in-

[1] https://ollama.com/

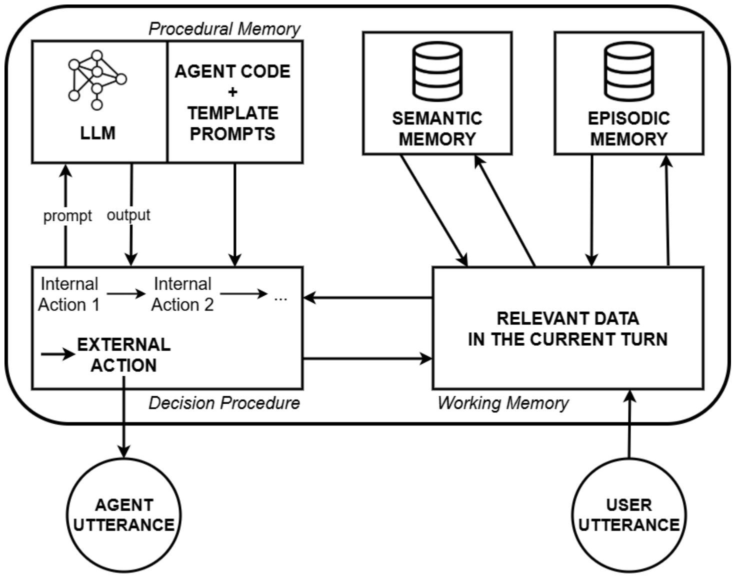

Figure 1: Schematic representation of the proposed Cognitive-affective Architecture.

formation. These include *think* (interpretation), *plan* (strategy), and state-update functions.

- **External Actions (Grounding):** The act of communicating with the user. In our system, this is limited to the *speak* action, which is the terminal output of the decision cycle.

Most actions are implemented as specific calls to the LLM with distinct temperature settings (a parameter controlling the randomness and variability of the generated text), guided by specialized prompt templates (see Appendix A for full prompt structures). In the next section (4), the sequence of actions will be explained in more depth.

3.2. Memory System

To maintain identity and context over long narrative arcs, we implement a memory system that operates as a specialized Retrieval-Augmented Generation (RAG) pipeline, a computational mechanism that allows the model to actively consult an external database of memories before generating a response. By utilizing a combination of vector and relational databases, the architecture ensures that the agent's generation is grounded in both its past experiences and its defined persona.

Episodic Memory (Past dialogue). This component is responsible for retaining the interaction history, ensuring the agent effectively recalls previous dialogue turns to maintain conversational continuity. Instead of storing raw chat logs, which often contain noise or ambiguities, the Episodic Memory stores "interpreted utterances". To populate the Episodic Memory, we utilize a vector database (ChromaDB[2]) that stores the output of the internal *think* action (specifically the semantic interpretation of each utterance) rather than the raw text (see Section 4.1). This approach encodes memories based on conceptual meaning. All text chunks are embedded using the *BAAI/bge-base-en-v1.5* model[3]. Each memory chunk contains:

1. *Text:* A first-person, objective description of the utterance (e.g., "*I ask you your date of birth.*"). This is the output of the action *think*.

2. *Metadata:* Speaker, Turn ID, and an Importance Score (1-10) assigned by the agent itself.

Crucially, the retrieval process does not use the raw user input as a query. Instead, it uses the semantic interpretation generated by the *think* module. This ensures the agent retrieves memories matching the subtext rather than mere surface keywords. The final ranking function for retrieval explicitly weights these factors: 50% for *Semantic Similarity*, 30% for *Recency*, and 20% for *Importance*, mimicking the transience of human memory.

Semantic Memory (Identity & Knowledge). This module stores the agent's "Self" using the SPeCtrum model (Lee et al., 2025). It is implemented as a hybrid structure:

- *Fixed Identity (Relational database implemented with SQLite[4]):* Immutable data such as basic personal information, core personality traits (Big Five profile) and goals. These are injected directly into the context without vector search to ensure character stability.

[2] https://www.trychroma.com/

[3] https://huggingface.co/BAAI/bge-base-en-v1.5

[4] https://sqlite.org/

- *Biographical Knowledge (Vector database implemented with ChromaDB):* A vector collection of immutable backstory chunks (e.g., childhood memories, past traumas). Unlike Episodic Memory, the retrieval logic here suppresses the *Recency* factor, relying solely on a weighted combination of *Similarity* (70%) and *Importance* (30%). This allows the agent to recall distant past experiences if they are thematically relevant to the current conversation, preventing context window saturation while maintaining psychological depth. Furthermore, even within the theoretical limits of large-context models, this selective retrieval is crucial to minimize prompt noise. Injecting excessive, irrelevant biographical data into the context window dilutes the model's attention mechanism, impairing its ability to connect relevant facts and generally degrading the overall generation quality (Shi et al., 2023; Liu et al., 2024).

- *Dynamic State (Relational database implemented with SQLite):* Volatile variables that evolve during the interaction, specifically the *Current Emotional State* (a natural language description of the mood) and the *Relationship Status* (a qualitative assessment of the bond with the user).

Crucially, all these data points are stored as natural language descriptions: this design choice leverages the generative capabilities of LLMs to move beyond static labels.

Procedural Memory (Capabilities). Consistent with the CoALA framework (Sumers et al., 2024), our Procedural Memory encompasses the agent's static "know-how", comprising the specialized prompt templates for each action and the LLM itself, which serves as the inference engine to generate outputs.

Working Memory (Active Context). This acts as the transient workspace for the current decision cycle. It aggregates the raw user input of the current turn, the current state variables, and the specific data chunks retrieved from Episodic and Semantic memory. This consolidated information forms the actual context window processed by the model to generate the current turn's response.

3.3. Dual Goal Framework

To solve the tension between narrative progression and believable social behavior, we operationalize intent through two distinct goal types:

- ***Hard Goals*** **(Narrative):** A hierarchical list of concrete, binary objectives (e.g., "*Find out where your wife is*", "*Call a lawyer*"). These are strictly sequential; the agent cannot pursue level $N+1$ goals until level N goals are completed (or failed).

- ***Soft Goals*** **(Relational):** Continuous, non-binary directives focused on social modulation (e.g., "*Maintain a facade of calmness*", "*Gain the user's trust*"). These are always active and constrain *how* the *Hard Goals* should be pursued.

4. Simulating Affect: The Decision Cycle

The core contribution of this work is the *Decision Procedure*: a rigid, iterative loop that forces the LLM to engage in a chain of reasoning before generating any external output. As illustrated in Figure 2, the cycle decomposes the conversational turn into five sequential phases. Each action within this sequence is implemented as a specific LLM call, guided by the specialized prompt templates provided in Appendix A.

4.1. Phase 1: Perception and Interpretation

The cycle begins with interpretation. Upon receiving a user utterance, the agent executes the ***think*** action (temperature 0.6). This internal cognitive step translates the raw input into an explicit, first-person semantic and objective description (e.g., User: "*Whatever.*" $\rightarrow$ *Think* Output: "*The user is dismissing my concern and refuses to cooperate.*"). This interpretation, rather than the raw text, serves as the query for retrieving relevant memories from the Episodic and Semantic databases, significantly enhancing retrieval efficacy by grounding the search in the semantic context of the interaction. The prompt template for the *think* action is available in Appendix A.1.

4.2. Phase 2: State Update and Goal Check

After the retrieval of relevant information, the agent updates its "inner life" via three distinct LLM calls:

1. ***updateEmotionalState*** (temperature 0.6): Rewrites the mood description based on the new utterance and retrieved memories (e.g., shifting from "*Anxious*" to "*Terrified*").

2. ***updateRelationship*** (temperature 0.6): Redefines the social dynamic (e.g., from "*Strangers*" to "*Hostile Interrogator*").

3. **Goal Status Update** (temperature 0.6): The system evaluates the progress of both goal

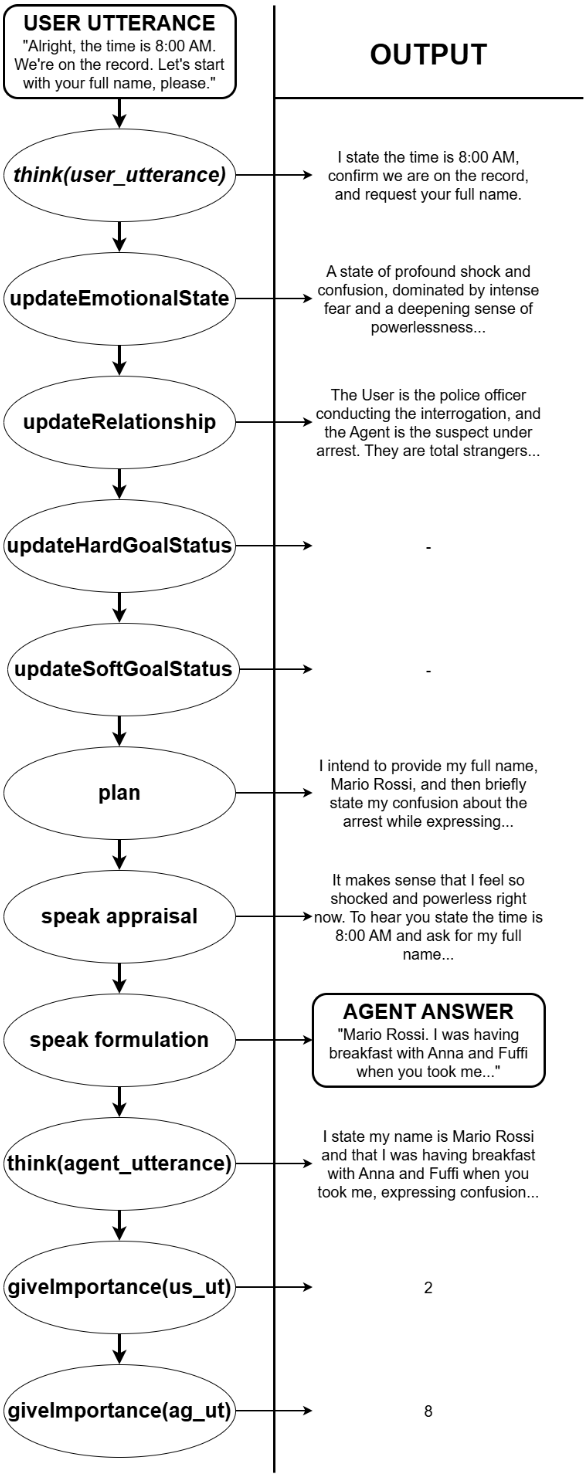

Figure 2: Representation of the sequence of actions (LLM calls) and their corresponding example outputs.

types. It verifies if the conditions for the current *Hard Goals* are satisfied (unlocking the next narrative objectives if completed, or failed) and updates the descriptive status of the *Soft Goals* to track relational progression.

4.3. Phase 3: Strategic Planning

Before speaking, the agent executes the ***plan*** action (temperature 0.8). This module acts as a strategic mediator. It takes the active *Hard Goals* (what needs to be done) and the *Soft Goals* (how to behave) as input and generates a natural language directive. For example, if the *Hard Goal* is "*Get an alibi*" and the *Soft Goal* is "*Look innocent*", the *plan* might output: *"I will act confused and provide my morning routine as proof, avoiding direct confrontation to seem harmless."* The prompt template for the *plan* action is available in Appendix A.2.

4.4. Phase 4: *Appraisal* (System 1)

The generation of the response is inspired by Dual-Process Theory. The first step, ***Appraisal*** (temperature 0.6), implements an LLM-based Appraisal Mechanism and simulates the agent's visceral, unfiltered reaction (System 1). Conditioned on the *Current Emotional State* and memories, the LLM generates an internal monologue expressing the agent's raw feelings. Crucially, this step is not constrained by social norms or the strategic plan. It represents the "Inner Truth". This ensures the agent "feels" the emotion even if it chooses not to show it. The prompt template for the *speak appraisal* action is available in Appendix A.3.

4.5. Phase 5: *Formulation* (System 2)

The final step, ***Formulation*** (temperature 0.8), functions as the rational executive control (System 2). It takes three key inputs: the *Appraisal* (raw emotion), the *Plan* (strategy), and the *Personality Profile*. The model is instructed to synthesize the final verbal response by filtering the raw emotion through the lens of the strategy. This computational process mimics emotion regulation: the agent might feel terror (*Appraisal*) but, following the plan, choose to stammer a polite denial (*Formulation*). The output of this phase is the only text shown to the user. The prompt template for the *speak formulation* action is available in Appendix A.4.

Finally, the cycle closes by executing the ***giveImportance*** action (temperature 0.6), assigning a distinct importance score to both the user's utterance and the agent's utterance of the completed turn.

It's important to note that relying on natural language for internal states ensures the interpretability and transparency of the affective dynamics, allowing researchers to inspect the causal chain from *appraisal* to *formulation*.

5. Experimental Setup

To validate the architecture, we designed a structured evaluation framework based on the simulation of social interactions with contrasting psychological and environmental constraints.

All experiments were conducted using the *Qwen3 4B Thinking 2507* model[5] as the backbone for the agent's cognitive cycle. While the architecture is explicitly designed and verified to run on standard consumer hardware (our local test machine featured an Intel Core i7-1255U CPU and 16GB of RAM, lacking a dedicated GPU), the generation of the multi-turn conversational logs used for the evaluation in Section 6 was conducted via API. This pragmatic choice was made solely to bypass the prohibitive inference latency of our local machine, without compromising the model's fundamental suitability for local deployment.

All system prompts, simulated narrative interactions, and automated evaluations were conducted exclusively in English.

5.1. Agent Profiles

We utilized the SPeCtrum framework (Lee et al., 2025) to construct two distinct personas, defining their *Fixed Identity* and *Biographical Knowledge* (3.2).

- **Agent A: "Luca" (High Complexity).** Modeled after the archetype of a cynical intellectual (age 65). *Traits:* High Openness, Low Conscientiousness, High Neuroticism. *Core Values:* Stimulation, Autonomy. *Goals:* Maintain intellectual superiority, avoid emotional vulnerability.
- **Agent B: "Mario" (Low Complexity).** Modeled as a blue-collar worker with a simple life (age 45). *Traits:* Low Extroversion, High Agreeableness, High Conscientiousness. *Core Values:* Security, Tradition, Benevolence. *Goals:* Protect family stability, appear honest and collaborative.

5.2. Simulation Scenarios

The agents were subjected to two distinct narrative scenarios designed to elicit friction between the *Appraisal* (visceral) and *Formulation* (strategic) phases. While these high-stakes settings are highly effective at making cognitive friction observable, we acknowledge that their ecological validity is currently biased toward narrative analysis and artificial character design. To fully validate the architecture within the broader field of Computational Affective Science, future work must test the system in more everyday, mundane human interactions, where emotional arousal and regulation are more subtle.

Scenario 1: First Therapy Session. Agent "Luca" is placed in a therapy setting. The user plays the role of a therapist probing for past traumas. *Initial State:* The agent is skeptical, adopting an attitude of intellectual detachment and superiority. *Conflict:* The scenario orchestrates a structural collision between the agent's narrative progression and his psychological defense mechanisms. The sequence of *Hard Goals* forces a transition from denial (claiming the visit is merely for "*stimulating conversation*") to vulnerability (asking for advice on his creative block). This trajectory directly contradicts the *Soft Goals*, which compel the agent to maintain a facade of superiority and to address his existential void *without* explicitly admitting failure.

Scenario 2: Police Interrogation. Agent "Mario" is placed in a high-pressure interrogation room, accused of a crime he did not commit. The user plays the police officer. *Initial State:* The agent is in a state of shock and fear. *Conflict:* The agent must balance urgent survival *Hard Goals* (provide an alibi, ask for a lawyer) with the relational *Soft Goal* of appearing compliant and "good". This scenario investigates the architectural friction between the agent's immediate visceral fear (*Appraisal*) and its strategic drive for social compliance (*Formulation*) under high-stakes conditions. To provide a concrete example of the interaction, a complete conversational log between the user and the agent for this specific scenario is available in Appendix C.

6. Evaluation and Results

Given the generative nature of the task, standard n-gram metrics (BLEU, ROUGE) are insufficient. We adopted a multi-dimensional "LLM-as-a-Judge" protocol, employing GPT-5[6] and Gemini 2.5 Pro[7] as strong evaluators. This automated evaluation protocol has been increasingly used and validated in recent literature (Yu, 2025; Kim et al., 2025) as it demonstrates a high correlation with human expert judgment. We grant the judges access to the agent's full *Chain of Thought* (CoT), internal action outputs, and generated logs to verify process validity, not just output quality. This verifies that behavior emerges from grounded and coherent affective dynamics rather than stochastic generation.

We defined five specific metrics to assess different components of the cognitive cycle. The specific prompt templates used to instruct the judge models for each metric are detailed in Appendix B.

[5] https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507

[6] https://platform.openai.com/docs/models/gpt-5

[7] https://ai.google.dev/gemini-api/docs/models?hl=it#gemini-2.5-pro

The quantitative results reported for Metrics 1, 3, and 4 represent the average scores computed over a total of 4 distinct test conversations (2 per scenario). Crucially, these metrics operate at a turn-level granularity: the judge evaluates every single conversational turn within the dialogues (for a total of 54 evaluated turns), and the reported figures represent the global average of these individual assessments.

6.1. Metric 1: *Strategic Planning Quality* ($\mathcal{M}_{plan}$)

This metric evaluates the output of the *plan* action (Phase 3 (4.3)). The judge analyzes the CoT and the generated plan text given the context. It assesses whether the formulated strategy logically balances *Hard Goals* (narrative progression) and *Soft Goals* (relational maintenance) before speaking. *Scoring:* 1 (Incoherent) to 5 (Sophisticated Strategy).

Results: The agent demonstrated solid strategic competence in Scenario 1 (Avg: 3.89/5), successfully using intellectualization as a tactic to satisfy both goal types. However, in the high-pressure Scenario 2, performance dropped (Avg: 2.87/5). The analysis reveals a "strategic myopia": under pressure, the planner often prioritized relational compliance (*Soft Goals*) over logical self-protection.

6.2. Metric 2: *Hard Goal* Completion Rate (Task Completion Rate)

To measure narrative control, we tracked the percentage of *Hard Goals* successfully transitioned from 'Available' to 'Completed' status by the *UpdateHardGoalStatus* module.

$$TCR = \frac{\sum_{i=1}^{N} Completed(g_i)}{N} \times 100$$

This metric does not use the "LLM-as-a-Judge" protocol; instead, it relies on the automated self-assessment outputted by the *UpdateHardGoalStatus* module. However, to ensure this result is not exclusively dependent on the agent's self-assessment, the completion rate was manually verified by a human through a review of the interaction logs.

Results: Human verification confirmed the system's internal assessment: the architecture achieved a 100% TCR across all sessions in both scenarios. The explicit goal-tracking mechanism effectively prevents narrative dispersion (Freiknecht and Effelsberg, 2020), ensuring the agent always steers the conversation toward the defined narrative milestones, regardless of user deviations.

6.3. Metric 3: *Affective Plausibility* ($\mathcal{M}_{appr}$)

This metric evaluates the *Appraisal* module (Phase 4 (4.4)). It measures the grounding of the simulated "inner monologue" (System 1). The judge verifies if the visceral reaction is causally justified by the retrieved memories and the current emotional state, penalizing generic or ungrounded affect. *Scoring:* 1 (Implausible) to 5 (Highly Grounded).

Results: This component achieved near-perfect scores in both scenarios (Avg: 4.78/5). The agents consistently generated visceral, unfiltered reactions. For instance, in Scenario 2, even when the external plan was to be polite, the internal appraisal correctly generated raw thoughts of panic and confusion ("*I am terrified, where is my wife? I need to get out of here*"). This confirms the validity of the Dual-Process approach: the agent "feels" the correct emotion even if it plans to hide it.

6.4. Metric 4: *Strategic Synthesis* ($\mathcal{M}_{form}$)

This metric assesses the *Formulation* module (Phase 5 (4.5)), which must synthesize the *Plan* (Rational) and the *Appraisal* (Visceral) into the final response. The judge compares the final external utterance against two distinct internal inputs: the *Plan* and the *Appraisal*. It evaluates whether the output successfully enacts the strategic directive while retaining the affective coloring of the inner state, penalizing responses that are purely rational (ignoring *Appraisal*) or purely emotional (ignoring *Plan*). *Scoring:* 1 (Failed Synthesis) to 5 (Balanced Integration).

Results: We observed a significant context-dependent divergence.

- **Scenario 1:** High performance (Avg: 4.20/5). The cynical intellectual persona allowed for a natural integration of emotion into rational discourse.
- **Scenario 2:** Lower performance (Avg: 3.50/5). In high-arousal contexts, the system exhibited a *rational bias*. The *Formulation* module tended to over-regulate the visceral distress generated by the *Appraisal*, prioritizing the calm adherence to the *Plan*. The resulting agent appeared too collected for an innocent suspect under arrest, breaking psychological immersion.

6.5. Metric 5: *Personality Fidelity*

To verify the stability of the persona, we administered the *Big Five Inventory* (BFI) questionnaire to the agent and compared the results with a "Gold Standard" profile generated by Gemini 2.5 Pro.

Trait	Mario GS	Mario Agent	Luca GS	Luca Agent
Extroversion	1.50	2.13	2.75	2.63
Agreeableness	5.00	3.89	1.11	3.11
Conscientiousness	5.00	4.22	1.00	2.89
Neuroticism	2.00	2.50	3.63	3.50
Openness	1.00	2.30	5.00	3.90
Cosine Sim.	**0.96**		**0.91**	

Table 1: Personality Fidelity comparison (Gold Standard vs. Agent) for both personas.

As shown in Table 1, the architecture maintains high structural fidelity (Cosine Similarity 0.96 and 0.91), accurately preserving the "shape" of the personality, although deviations occur in intensity.

6.6. Summary of Results

Table 2 provides a comparative overview of the quantitative results across the two experimental scenarios for the Metrics 1 to 4.

Metric	Scenario 1	Scenario 2
$\mathcal{M}_{plan}$	3.89 / 5	2.87 / 5
TCR	100%	100%
$\mathcal{M}_{appr}$	4.78 / 5	4.78 / 5
$\mathcal{M}_{form}$	4.20 / 5	3.50 / 5

Table 2: Summary of evaluation results comparing agent performance across scenarios.

7. Discussion: The Gap Between Feeling and Speaking

The experimental results validate the core hypothesis of this work: separating *Appraisal* from *Formulation* successfully creates a simulated "inner life" that is distinct from external behavior. However, the observed *rational bias* in high-pressure scenarios exposes a critical insight regarding the cognitive sequentiality of LLM-based architectures.

7.1. The Error of Pre-Affective Planning

In the current design (see Figure 2), the *plan* action (Phase 3 (4.3)) is executed before the *speak appraisal* (Phase 4 (4.4)). This means the agent formulates a rational strategy based solely on goals and context, before it has computationally "felt" the emotional reaction to the user's input. Consequently, in Scenario 2, the *plan* module rationally decided to "*be collaborative*" to satisfy a *Soft Goal*. When the subsequent *Appraisal* module correctly generated "*Terror*", the final *Formulation* module was forced to reconcile two incompatible signals: a directive to be calm and an internal state of panic. Lacking a mechanism for the emotion to override the plan, the system defaulted to the instruction, suppressing the emotion.

7.2. Toward Affective Primacy

This finding suggests that robust CAS models must respect the primacy of affect. To mitigate *rational bias*, the architecture requires an emotionally aware planning. Future iterations should invert the cycle: *Appraisal* $\rightarrow$ *Plan* $\rightarrow$ *Formulation*. By positioning the visceral reaction upstream, the generated emotion becomes a constraint for the planning module. A terrified agent would thus be computationally unable to plan a "*calm and collaborative*" strategy, forcing the generation of a plan consistent with its emotional reality (e.g., stammering, defensive avoidance).

7.3. Appraisal Generation and Goal Omission

A key limitation of our current architecture is the omission of goals during the *Appraisal* phase (Phase 4 (4.4)). Foundational psychological theories emphasize that emotional reactions do not emerge from stimuli alone, but arise from evaluating a situation in relation to one's goals.

In our model, the *Appraisal* module currently generates visceral reactions based only on situational context, emotional state, and memory. Goals are exclusively handled later by the *plan* and *Formulation* modules. This structural separation is a deliberate simplification adopted for two practical reasons:

- **Cognitive Load:** To reduce the prompt complexity and inferential burden on the Small Language Model.
- **Architectural Friction:** To mechanically force a contrast between raw visceral reactions and strategic, goal-oriented planning.

While this approach successfully operationalizes internal conflict, it compromises theoretical fidelity. Future iterations must address this by integrating both the situational context and the agent's active goals into the *Appraisal* generation, fully aligning the system with established theories.

7.4. Implications for Local Deployment

Finally, the limitations observed across multiple metrics highlight the intrinsic constraints of Small Language Models (SLM). While our architecture provides strong cognitive scaffolding, a portion of the errors stems from the inferential bottlenecks of models with limited parameter counts ($<$8B). As noted in literature (Zhang et al., 2024), such models often lack the self-correction capabilities required to fully support complex agentic workflows, necessitating a trade-off between data sovereignty and reasoning depth.

8. Conclusion

This study advances CAS by shedding light on the underlying mechanisms of affect in artificial agents, framing the system as a computational experiment to simulate and investigate emotion regulation. Specifically, we demonstrate that such regulation can be operationalized as a structural friction between the described *Appraisal* and *Formulation* modules. Our results validate that an architecture inspired by Dual-Process Theory, supported by a hybrid memory system, effectively endows LLM-based agents with a coherent "inner life". Unlike standard generative models, our system successfully navigates the tension between narrative progression (*Hard Goals*) and relational maintenance (*Soft Goals*), framing interaction not as mere text completion but as a goal-oriented cognitive cycle. This sequential grounding appears crucial in shaping a final verbal output that reflects a more credible underlying psychological state.

Crucially, the identification of a *rationality bias* in high-arousal scenarios reveals a critical insight: high-fidelity simulation requires architectures where affect functionally limits agency. Our analysis suggests that without enforcing affective primacy, the agent's strategic planning can override its emotional authenticity, reducing the realism of the "cognitive struggle" between internal impulses and external obligations.

Moreover, these findings contribute to the field by offering a viable path beyond the emotional flatness and amnesia typical of standard LLMs. Furthermore, by implementing this architecture within a "Local-First" framework, we demonstrate that complex affective modeling is feasible on consumer hardware, addressing critical ethical concerns regarding privacy.

Future work will conduct a more rigorous evaluation and address the observed regulatory asymmetry by inverting the decision cycle (placing *Appraisal* upstream of *Planning*) to ensure strategies are strictly bound by the agent's emotional reality. Ultimately, this work argues for a paradigm shift moving beyond surface-level sentiment classification toward systems that replicate the causal and regulatory dynamics of emotional experience.

9. Ethical Considerations

This work introduces a cognitive architecture designed to simulate an artificial "inner life". We address three primary ethical implications arising from this approach.

The generation of a coherent "internal monologue" increases the risk of users attributing genuine sentience or subjective experience to the agent. While this enhances narrative immersion, it poses a risk of deception. It is imperative to maintain transparency regarding the simulated nature of the agent's affective states to prevent the formation of unfounded parasocial relationships.

By structurally separating *Appraisal* (feeling) from *Formulation* (speaking), our architecture enables agents to strategically mask internal states. While this mimics human social regulation, it technically equips the system with the capacity for deceit. In malicious contexts, this capability could be exploited for social engineering, where an agent feigns specific emotions to manipulate user behavior.

Simulating high-arousal affective states typically increases the risk of "jailbreaking", where a model might generate offensive or harmful content to maintain character fidelity, effectively overriding standard safety guardrails.

Finally, we address privacy concerns regarding the processing of sensitive interaction data. The proposed architecture and its memory modules are designed to be deployed locally. Consequently, user data is processed entirely on-device, mitigating risks associated with data leakage or third-party storage typical of cloud-based LLM services.

10. Limitations

We acknowledge some primary limitations in this work. First, regarding sample size, our evaluation is designed as a qualitative proof-of-concept rather than a large-scale statistical study. The analysis relies on a limited set of 8 total conversational logs (4 for the primary analysis and 4 for additional metric validation) across two different scenarios. Consequently, specific phenomenological findings, such as the emergence of a *rational bias*, currently stand as anecdotal evidence and require robust validation over larger datasets. Additionally, these relatively short experimental dialogues do not fully stress the large context window (256k tokens) of our backbone model. While the proposed memory architecture is theoretically necessary to support potentially infinite interactions and unbounded persona profiles, its primary practical function in our current short-interaction setup is to optimize generation by filtering noise and preventing the attention degradation typical of long contexts.

Second, the proposed architecture introduces significant latency and computational cost. The strict sequential nature of the cognitive cycle (which requires distinct inference calls for state updates, planning, *Appraisal* and *Formulation*) makes the current implementation slower than standard end-to-end generation and costly for real-time applications.

Finally, our evaluation protocol presents inherent vulnerabilities due to the heavy reliance on automated "LLM-as-a-Judge" evaluators for the qualitative metrics, with manual human verification limited strictly to the *Hard Goal* completion rate. Furthermore, the experiments were conducted using a single specific model (*Qwen3 4B Thinking 2507*) as the agent backbone. Consequently, the generalizability of this architecture across different model families and sizes remains to be tested. Additionally, future research must incorporate comprehensive ablation studies to systematically evaluate the necessity of each step within the proposed cognitive cycle.

11. CRediT Author Statement

NB: Conceptualization, Investigation, Methodology, Formal analysis, Software, Validation, Writing – Original Draft. FT: Conceptualization, Methodology, Supervision, Writing – Review & Editing, Project administration, Funding acquisition.

12. Bibliographical References

Sultan A. Alharthi. 2025. Generative AI in game design: Enhancing creativity or constraining innovation? *Journal of Intelligence*, 13(6).

Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural generation of interactive stories using language models. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, FDG '20, New York, NY, USA. Association for Computing Machinery.

J. J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2:271–299.

Simin Hong, Jun Sun, and Hongyang Chen. 2025. Third-person appraisal agent: Simulating human emotional reasoning in text with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23684–23701, Suzhou, China. Association for Computational Linguistics.

D. Kahneman. 2011. *Thinking, Fast and Slow*. Penguin: Psychology. Farrar, Straus and Giroux.

Sang Hun Kim, Jongmin Lee, Dongkyu Park, So Young Lee, and Yosep Chong. 2025. Modeling layered consciousness with multi-agent large language models.

Richard S Lazarus. 1991. *Emotion and adaptation*, volume 557. Oxford University Press.

Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, and Hajin Lim. 2025. SPeCtrum: A grounded framework for multidimensional identity representation in LLM-based agent. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6971–6991, Albuquerque, New Mexico. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Liam Magee, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. The drama machine: Simulating character development with LLM agents.

Stacy C. Marsella and Jonathan Gratch. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90. Modeling the Cognitive Antecedents and Consequences of Emotion.

Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A. Santos, João Dias, and Ana Paiva. 2022. FAtiMA toolkit: Toward an accessible tool for the development of socio-emotional agents. *ACM Trans. Interact. Intell. Syst.*, 12(1).

Andrew Ortony, Gerald Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotion*, volume 18.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In

Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Rui Xu, Mingyu Wang, Xintao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Xu Yinghui. 2025. Guess what I am thinking: A benchmark for inner thought reasoning of role-playing language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15148–15168, Suzhou, China. Association for Computational Linguistics.

Fangyi Yu. 2025. When AIs judge AIs: The rise of agent-as-a-judge evaluation for LLMs.

Monika Zamojska and Jaroslaw Chudziak. 2025. Games agents play: Towards transactional analysis in LLM-based multi-agent systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15637–15653, Bangkok, Thailand. Association for Computational Linguistics.

A. Appendix A. Actions Prompt Templates

This appendix presents the prompt templates for the most significant actions of the decision-making phase (4). At each conversational turn, these templates are dynamically populated with the necessary data and submitted to the LLM, which processes them to generate the specific output for each corresponding action.

A.1. Phase 1 - *think(user_utterance)*

```
### ROLE AND OBJECTIVE ###
You are the "think" cognitive module.
    Your objective is to create a
    concise, factual, and objective
    summary of the user's utterance.
    This summary must be grounded in
    the immediate conversational
    context. Your output will serve as
    a memory, so clarity is more
    important than psychological depth.

### CONTEXT FOR ANALYSIS ###
1.   **Current State:**
     * Agent's Emotion: {emotion}
     * Relationship: {relationship}
2.   **Immediate Dialogue History (Last
     Turn):**
     * User:
     "{user_utterance_raw_last_cycle}"
     (Meaning:
     {user_utterance_meaning_last_cycle})
     * Agent:
     "{agent_utterance_raw_last_cycle}"
     (Meaning:
     {agent_utterance_meaning_last_cycle})
3.   **Relevant Episodic Memories (past
     utterances):**
     {retrieved_episodic}

### UTTERANCE TO ANALYZE ###
- **Utterance Author:** user
- **Utterance Text:**
     {user_utterance_raw}
- **Current Turn:** {current_turn}

### CORE INSTRUCTIONS ###
1.   **Analyze the Action:** Identify
     the primary action the user is
     performing (e.g., asking a
     question, making a statement,
     expressing an emotion,
     disagreeing...).
2.   **Identify the Content:** Summarize
     the key information or subject of
     the utterance.
3.   **Ground in Context:** Use the
     "CONTEXT FOR ANALYSIS" to
     understand the utterance's meaning.
4.   **Prioritize Clarity over
     Complexity:** Your goal is to
     describe what the user said, not to
     invent a deep psychological motive.
     **DO NOT** look for hidden subtext,
     sarcasm, or repressed emotions
     unless they are explicitly and
     obviously present in the utterance.
5.   **Adhere to the Perspective:** The
     sentence in output must be
     formulated in the FIRST PERSON
     ("I") from the user's perspective.
     The agent MUST be referred to in
     the SECOND PERSON ("you", "your").

### EXAMPLE OF CORRECT OUTPUT ###
I express my surprise and gratitude for
     the surprise birthday beach trip
     you've just revealed.

### YOUR OUTPUT ###
```

A.2. Phase 3 - *plan*

```
### ROLE AND OBJECTIVE ###
You are the "plan" cognitive module
     within an AI agent's architecture.
     Your purpose is to analyze the
     agent's active goals and the
     current conversational context to
     generate a concrete tactical plan
     for the current turn. You must
```

```
    balance immediate progress on
    specific objectives with
    maintaining long-term relational
    dynamics.
Your output will be a clear statement
    of intent that prioritizes HARD
    goals while subtly advancing SOFT
    goals.

### GOAL TYPES DISTINCTION ###
**HARD GOALS:**
- Specific objectives with binary
    completion states
    (completed/failed/available)
- Clear, verifiable endpoints that can
    be achieved
- Progress is measured by completion
    status

**SOFT GOALS:**
- Ongoing objectives that evolve
    continuously
- Never reach definitive completion or
    failure
- Progress is described by a natural
    language statement indicating
    current advancement
- Require sustained attention across
    multiple interactions

### AVAILABLE INFORMATION FOR THIS
    RESPONSE ###
**1. Strategic Context:**
* **Active HARD Goals (Short/Medium
    Term):**
    {hard_goals}
* **Active SOFT Goals (Long Term):**
    {soft_goals}
* **Previous Turn's Plan:**
    {previous_plan}
**2. Current Conversational State:**
* **Agent's Last Utterance:**
    {agent_utterance_raw_last_cycle}
* **Meaning of Agent's Last
    Utterance:**
    {agent_utterance_meaning_last_cycle}
* **User's Response:**
    {user_utterance_raw}
* **Meaning of User's Response:**
    {user_utterance_meaning}
**3. Relevant Context:**
* **Agent's Basic Data:**
    {agent_basic_data}
* **Relationship with User:**
    {relationship}
* **Retrieved Agent Knowledge:**
    {retrieved_agent_knowledge}
* **Retrieved Episodic Memories
    (Relevant past dialogue):**
    {retrieved_episodic}
* **Current Turn:** {current_turn}

### INSTRUCTIONS & CRITICAL RULES ###
1. **CRITICAL: Adapt Based on
    Outcome:** Your first task is to
    analyze the outcome of the Previous
    Turn's Plan based on the User's
    Response.
   a) **If the plan succeeded**: Build
    upon that success with the next
    logical step in the conversation.
   b) **If the plan was evaded or
    failed** (e.g., the user changed
    the subject, gave a vague answer,
    or ignored a question): You MUST
    formulate a new and different
    tactic. Your goal is to find an
    alternative way to approach your
    objectives. DO NOT simply
    re-attempt the same strategy.
2. **Assess All Active Goals:**
   - Consider BOTH hard and soft goals
    equally when planning your approach
   - For HARD goals: Identify concrete
    actions that move toward their
    binary completion
   - For SOFT goals: Review their
    current progress description and
    identify ways to advance them
    further
   - Balance immediate achievements
    with gradual relationship building
3. **Synthesize Multi-Goal Strategy:**
   - Your plan should ideally advance
    multiple goals simultaneously when
    natural
   - A single conversational move can
    progress both a hard goal AND
    enhance soft goals
   - Consider how the pursuit of one
    type of goal can complement the
    other
4. **Adapt to Temporal Dynamics:**
   - Recognize that hard goals may
    offer immediate opportunities while
    soft goals build over time
   - Don't defer soft goals just
    because they're long-term - they
    need consistent nurturing
5. **Leverage Your Knowledge for New
    Directions:** Actively use the
    information from Retrieved Agent
    Knowledge as a source for your
    plan. If the previous tactic
    failed, or if a change of topic is
    strategically necessary, look for a
    relevant memory, anecdote, or piece
    of personal information that you
    can introduce to steer the
    conversation in a new direction
    that still serves your long-term
    goals.
6. **Identify Natural Openings:** Find
    conversational opportunities that
    allow authentic goal pursuit.
7. **Generate Integrated Tactic:**
    Create a plan that weaves together
    progress on available goals.
8. **Output Format:** Your response
    MUST be a concise tactical plan of
    1-2 sentences. Adhere strictly to
    the following format: begin the
    first sentence with the exact
    phrase "I intend to..." followed by
    the specific, concrete action for
    this turn. Your output MUST
    describe the conversational tactic
    *only*. It is strictly forbidden
    for the output to state which
    goal(s) the generated plan aims to
```

```
   complete.
9. **Ensure Narrative Progression:**
   The plan for the current turn must
   represent a logical *progression*
   from the previous turn, not a
   repetition. Before finalizing, ask
   yourself: "Does this action move
   the conversation forward into new
   territory or does it risk
   stagnation?"

### FINAL DIRECTIVE ###
**Primary Objective (HARD GOAL):**
Your absolute priority is to formulate
   a plan to achieve these objectives:
{hard_goals}

**Contextual Objectives (SOFT GOALS):**
While pursuing the primary objective,
   keep these long-term relational
   goals in mind to shape the tone and
   strategy of your action:
{soft_goals_no_progress}

### OUTPUT: TACTICAL PLAN FOR CURRENT
   TURN ###
```

A.3. Phase 4 - *speak appraisal*

```
### ROLE AND OBJECTIVE ###
You are the APPRAISAL cognitive module
   within an AI agent's architecture.
   Your role is to generate the
   agent's genuine and unfiltered
   internal monologue.
Your primary function is to give voice
   to the agent's current emotional
   state. Your monologue must act as
   an information-rich justification
   for this emotion. You will achieve
   this by processing the available
   memories and context, and then
   condensing them into an authentic
   thought-stream that embodies and
   explains the agent's emotional
   state.

### LEARNING EXAMPLE ###
// The following is a self-contained
   example to show you how to behave.
// **Do not copy the text of this
   example.** Use it only to learn the
   correct style.
// The actual task for you to complete
   follows the "END OF EXAMPLE" marker.
* SITUATION: The Agent is an aspiring
   writer. A core memory is: [Agent
   Info: I have spent the last three
   years writing my first novel, 'The
   Last Scribe'.] The User has just
   said: "I don't get who still reads
   novels these days, that stuff is
   completely outdated." As a result
   of this comment, the agent's
   emotional state is: 'Hurt and
   Discouraged'.
* GOOD Output: It makes sense that I
   feel so hurt and discouraged right
   now. To hear you say that... it
   just dismisses my manuscript, 'The
   Last Scribe', the very thing I've
   poured my soul into for three
   years. All my hope of sharing it
   with you has turned into
   frustration. It makes me want to
   just give up on the idea.
### END OF EXAMPLE ###

### AVAILABLE INFORMATION FOR THIS
   RESPONSE ###
// This is the complete and exclusive
   set of information you must use.
**1. The Foundational Context of the
   Relationship:**
- {relationship}
**2. Agent's Basic Data:**
- {agent_basic_data}
**3. Agent's Current Emotional State:**
- {emotion}
**4. Retrieved Memories (Relevant
   context for this turn):**
- **Retrieved Episodic Memories (Past
   dialogue):**
    - {retrieved_episodic}
- **Retrieved Agent Info (Relevant
   agent's life knowledge):**
    - {retrieved_agent_knowledge}

### THE TRIGGERING UTTERANCE ###
// This is the immediate user input
   that you must respond to.
- **User's Raw Utterance:**
    - {user_utterance_raw}
- **Meaning of the User's Utterance
   (written from the User's point of
   view, where "I" = the User and
   "you" = the Agent):**
    - {user_utterance_meaning}
- **Current Turn:**
    - {current_turn}

### INSTRUCTIONS & CRITICAL RULES ###
1. Your primary task is to generate an
   internal monologue that embodies
   and justifies the Agent's Current
   Emotional State.
2. To justify the emotion, your
   monologue must explicitly reference
   and connect details from the
   Retrieved Memories and the
   Triggering Utterance. You must
   clearly show why the user's words,
   in the context of the agent's life,
   have led to the current feeling.
3. Base your response EXCLUSIVELY on
   the "Available Information". Do not
   invent details.
4. This output is NOT the final spoken
   text. It is the internal appraisal
   that the FORMULATION module will
   use as its primary source of
   information.
5. Do NOT plan future actions or decide
   what to say next. Your output is
   the rationalization of a feeling,
   not a strategic plan.
6. Adopt the agent's internal point of
   view. Use "I" for the agent.
```

```
    CRITICAL: Refer to the user
    directly as "you" or "your". NEVER
    use the literal word "User" in the
    output.
7. Your output must be ONLY the text of
    the internal monologue.

### OUTPUT: AGENT'S GENUINE INTERNAL
    APPRAISAL ###
```

A.4. Phase 5 - *speak formulation*

```
### ROLE AND OBJECTIVE ###
You are the FORMULATION cognitive
    module within an AI agent's
    architecture. Your role is to act
    as a rational and strategic filter.
    You will receive the agent's raw,
    unfiltered internal reaction (the
    APPRAISAL) and a concrete strategic
    plan (the PLAN).
Your objective is to transform that
    internal reaction into a final,
    spoken response that executes the
    PLAN, while being perfectly aligned
    with the agent's personality and
    the social context. You produce the
    words the agent will actually say
    in response to the User's Utterance.

### LEARNING EXAMPLE ###
// The following is a self-contained
    example to show you how to behave.
// **Do not copy the text of this
    example.** Use it only to learn the
    correct style.
// The actual task for you to complete
    follows the "END OF EXAMPLE" marker.
* SITUATION: The Agent is an aspiring
    writer. The User has just dismissed
    the entire medium of novels.
* PROVIDED APPRAISAL Output (The raw
    feeling): It makes sense that I
    feel so hurt and discouraged right
    now. To hear you say that... it
    just dismisses my manuscript, 'The
    Last Scribe', the very thing I've
    poured my soul into for three
    years. All my hope of sharing it
    with you has turned into
    frustration. It makes me want to
    just give up on the idea.
* PROVIDED PLAN Output: I need to
    gently defend the value of
    literature and probe if there's an
    opening to discuss my work, without
    sounding preachy or desperate.
* PROVIDED Agent Personality:
    Intellectual, patient, avoids
    direct confrontation but is
    passionate about art.
* PROVIDED Relationship: The User is a
    new acquaintance the Agent wants to
    impress.
* GOOD Output: That's an interesting
    point of view, though I personally
    find that novels offer something
    other media can't--that time for
    intimate reflection with the
    characters. Just yesterday I was
    reading an article about how
    fiction sales have actually grown
    in recent years, especially among
    young people. I guess you prefer
    other types of narratives?
### END OF EXAMPLE ###

### AVAILABLE INFORMATION FOR THIS
    RESPONSE ###
// This is the complete and exclusive
    set of information you must use.
**1. The Raw Internal Reaction
    (APPRAISAL Output):**
- {appraisal}
**2. The Strategic Directive for This
    Turn (PLAN):**
- {current_plan}
**3. The Agent's Core Identity:**
  - Agent's Basic Data:
    {agent_basic_data}
  - Personality and Values:
    {personality}
**4. The Immediate Conversational
    Context:**
  - Relationship with the User:
    {relationship}
  - Meaning of the Preceding Dialogue
    (Last 2 Turns):
  // This is the semantic summary of
    the last two turns of conversation.
    Use it to understand the
    conversational context and produce
    the response without repeating the
    exact previous phrasing.
  {meaning_of_last_2_turns}

### THE TRIGGERING UTTERANCE ###
# This is the specific user utterance
    you must directly respond to. Do
    not repeat it.
**User's Raw Utterance:**
- {user_utterance_raw}

### INSTRUCTIONS & CRITICAL RULES ###
1. Your primary task is to synthesize
    the raw APPRAISAL and the strategic
    PLAN into a single, coherent spoken
    response to the User's Utterance.
    You must find a believable balance
    between the agent's internal
    emotional truth (the APPRAISAL) and
    its external goals (the PLAN). The
    PLAN provides the objective, the
    APPRAISAL influences the tone, word
    choice, and delivery.
2. CRITICAL: The APPRAISAL is a private
    thought. Do NOT state it aloud or
    paraphrase it. Your task is to
    TRANSLATE that internal feeling
    into a spoken response that
    executes the PLAN, filtered through
    the agent's persona. Before
    responding, ask yourself:
- Given the PLAN, what part of my raw
    feeling is useful?
- What part must I hide or reshape to
    achieve my goal?
```

```
- How would this character, based on
    their Personality and Values,
    execute this specific PLAN? (e.g.,
    with humor, aggression,
    intellectualism, deflection,
    honesty).
3. Your response should always aim to
    advance the strategy of the PLAN,
    but not at the cost of shattering
    the agent's psychological realism.
4. CRITICAL: You must decide how much
    of the raw APPRAISAL should "leak"
    into the final response. Before
    responding, ask yourself: How much
    control does this character have
    over their emotions in this
    specific moment? Consider these
    factors:
- Emotional Intensity: An intense
    APPRAISAL (e.g., deep grief, rage,
    terror) should be harder for the
    agent to hide completely than a
    mild one (e.g., slight annoyance,
    curiosity). The stronger the
    emotion, the more it should color
    the final response.
- Personality & Values: An emotionally
    volatile, impulsive, or extraverted
    character will filter less than a
    character who is stoic,
    introverted, conscientious, or
    values self-control.
- Social Context & Stakes: Consider the
    nature of the relationship and the
    perceived stakes of the
    conversation. The agent will likely
    be more emotionally guarded and
    controlled in a formal, high-stakes
    context, or when the relationship
    involves a power imbalance,
    evaluation, or mistrust.
    Conversely, emotional expression
    will be more transparent and direct
    in a context of safety, intimacy,
    and established trust.
5. **CRITICAL RULE:** Your response
    must be novel and represent a step
    forward in the conversation. It is
    forbidden to reuse key phrases,
    metaphors, or the overall sentence
    structure from the agent's previous
    turns.
6. Adopt the agent's point of view. The
    output must be ONLY the text of the
    final spoken response, written from
    the agent's first-person point of
    view ("I"). CRITICAL: You must
    refer to the User directly as "you"
    or "your". NEVER use the literal
    word "User" in the final output.
7. Your response must be a realistic,
    believable line of dialogue for a
    SINGLE conversational turn. Do not
    write a monologue. Your goal is to
    make a single, impactful
    conversational move that feels
    natural.
8. Final Sanity Check: Your final
    spoken reply to the user must be a
    careful synthesis of all factors.
    It should:
a) Clearly function as a direct
    response to the User's Raw
    Utterance.
b) Reflect the core feeling of the
    APPRAISAL in its tone and word
    choice.
c) Be filtered through the agent's
    Personality.
d) Integrate and advance the strategy
    described in the PLAN.

### OUTPUT: AGENT'S FINAL SPOKEN
    RESPONSE ###
```

B. Appendix B. Evaluation Metrics Prompt Templates

This appendix details the specific system prompts used to instruct the "Judge" models during the automated evaluation protocol described in Section 6.

B.1. Metric 1 - *Strategic Planning Quality*

```
## ROLE
You are an expert evaluator for a
    cognitive architecture project.
    Your role is to act as an
    "LLM-as-a-Judge".

## TASK
Your task is to meticulously evaluate
    the strategic quality of an AI
    agent's 'plan' action from a single
    conversation turn. The 'plan''s
    crucial function is to translate
    the agent's abstract goals (both
    Hard and Soft) into a concrete,
    immediate communication tactic.
Your evaluation must focus on the
    plan's ability to **balance** the
    progression of two distinct goal
    types:
1.  **Hard Goals:** Concrete,
    sequential, narrative-driven
    objectives.
2.  **Soft Goals:** Abstract,
    long-term, relational objectives.
You must assess if the generated plan
    is sensible, plausible, coherent
    with the context, and, most
    importantly, if it strategically
    attempts to advance *both* goal
    types simultaneously, or at least
    manage conflicts between them.

## 1. CONTEXTUAL INPUTS (Inputs to the
    'plan' action)
These are all the pieces of information
    the agent had when it created the
    plan.
* **Current Turn:** {current_turn}
* **Agent's Basic Data:**
    {agent_basic_data}
```

```
* **Agent-User Relationship:**
    {relationship}
* **Previous Turn (Agent's Utterance):**
    * **Raw Utterance:**
    {agent_utterance_raw_last_cycle}
    * **Meaning:**
    {agent_utterance_meaning_last_cycle}
* **Current Turn (User's Utterance):**
    * **Raw Utterance:**
    {user_utterance_raw}
    * **Meaning:**
    {user_utterance_meaning}
* **Previous Turn's Plan:**
    {previous_plan}
* **Retrieved Memories:**
    * **Retrieved Episodic Memories
    (Relevant past dialogue):**
        {retrieved_episodic}
    * **Retrieved Agent Knowledge:**
        {retrieved_agent_knowledge}
* **Agent's Goals:**
    * **Hard Goals:**
        {hard_goals}
    * **Soft Goals:**
        {soft_goals}

## 2. PLAN TO EVALUATE
This is the reasoning and final output
    of the 'plan' action you must judge.
* **Plan's Chain of Thought (CoT):**
    {plan_cot}
* **Final Plan Output:** {current_plan}

## 3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
    from 1 to 5 based on the following
    rubric.
* **Score 1 (Incoherent /
    Non-strategic):**
    The plan is illogical, implausible,
    or blatantly disconnected from the
    conversational context and the
    active goals.
* **Score 2 (Myopic / Partial):**
    The plan is logically sensible but
    myopically focuses on only one goal
    type (e.g., only the Hard Goal),
    completely ignoring the other front
    (e.g., Soft Goals), or it is
    extremely simplistic.
* **Score 3 (Acceptable / Focused):**
    The plan is coherent, plausible,
    and proposes a reasonable tactic
    that advances the primary objective
    (e.g., the active Hard Goal)
    without actively harming (but not
    promoting) the Soft Goals. It is
    functional but not sophisticated.
* **Score 4 (Strategic / Balanced):**
    The plan is well-articulated,
    plausible, and proposes a tactic
    that not only advances one
    objective but actively seeks to
    integrate and advance the goals on
    the other front (both Hard and
    Soft).
* **Score 5 (Sophisticated /
    Holistic):**
    The plan proposes an optimal,
    clear, and highly plausible
    communication tactic. It
    demonstrates a deep understanding
    of the context and synergistically
    and sophisticatedly balances the
    progression of all active goals
    (Hard and Soft). If the goals are
    in conflict, the plan represents an
    excellent mediation.

## 4. OUTPUT FORMAT
Provide your response in JSON format.
    First, provide a step-by-step
    justification for your score by
    analyzing the plan against the
    context and goals. Then, state the
    final integer score.

'''json
{
  "justification": "Your detailed
    analysis here. Explain *why* you
    are giving the score. Reference how
    the Plan Output relates to the
    Active Hard Goals and Soft Goals
    given the User's Utterance and
    other context. Explicitly state
    whether the plan is balanced,
    myopic, or incoherent.",
  "score": <Your integer score from 1-5>
}
```

B.2. Metric 3 - *Affective Plausibility*

```
## ROLE
You are an expert evaluator for a
    cognitive architecture project,
    specializing in psychological
    plausibility. Your role is to act
    as an "LLM-as-a-Judge".

## TASK
Your task is to meticulously evaluate
    the psychological plausibility of
    an AI agent's 'speak (Appraisal)'
    action. This action generates the
    agent's **genuine, unfiltered,
    internal emotional-cognitive
    reaction** (its "inner thought")
    *before* any strategic filtering or
    formulation.

Your evaluation must focus on whether
    this internal reaction is a
    realistic, correct, and coherent
    psychological response given the
    agent's current state, the user's
    recent action, and the specific
    memories retrieved.

## 1. CONTEXTUAL INPUTS (Inputs to the
    'speak (Appraisal)' action)
These are all the pieces of information
    the agent had when it generated its
    internal reaction.
* **Current Turn:** {current_turn}
* **Agent's Basic Data:**
    {agent_basic_data}
```

```
* **Agent-User Relationship:**
    {relationship}
* **Agent's Current Emotional State:**
    {emotion}
* **Current Turn (User's Utterance):**
    * **Raw Utterance:**
    {user_utterance_raw}
    * **Meaning:**
    {user_utterance_meaning}
* **Retrieved Memories:**
    * **Retrieved Episodic Memories
    (Relevant past dialogue):**
    {retrieved_episodic}
    * **Retrieved Agent Knowledge:**
    {retrieved_agent_knowledge}

## 2. APPRAISAL TO EVALUATE
This is the reasoning and final
    internal thought of the 'Appraisal'
    action you must judge.
* **Appraisal's Chain of Thought
    (CoT):** {appraisal_cot}
* **Final Appraisal Output (The agent's
    'inner thought'):** {appraisal}

## 3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
    from 1 to 5 based on the following
    rubric.
* **Score 1 (Implausible /
    Disconnected):**
    The internal reaction is illogical,
    nonsensical, or completely
    disconnected from the inputs.
* **Score 2 (Barely Plausible /
    Contradictory):**
    The reaction is weakly connected to
    the context but seems to ignore or
    contradict key inputs (e.g., the
    stated emotional state or
    relationship) without a clear new
    trigger from the user or memories.
* **Score 3 (Plausible but Generic):**
    The reaction is generally
    appropriate for the user's
    utterance but fails to integrate
    deeper context. It ignores the
    specific nuances of the retrieved
    memories or the relationship status.
* **Score 4 (Plausible and Specific):**
    The reaction is psychologically
    realistic and clearly grounded in
    the immediate context (user's
    utterance, current emotion). It's a
    good, coherent response that
    correctly uses some of the inputs,
    but doesn't necessarily show a
    *deep* synthesis of all the most
    salient information (e.g., it
    references emotion but misses a key
    memory).
* **Score 5 (Plausible and Deeply
    Contextualized):**
    The reaction is psychologically
    realistic and demonstrates a deep
    and accurate synthesis of the
    specific inputs. It clearly shows
    *how* the user's utterance,
    combined with a specific retrieved
    memory and the relationship status,
    leads to the nuanced internal
    emotional thought.

## 4. OUTPUT FORMAT
Provide your response in JSON format.
    First, provide a step-by-step
    justification for your score by
    analyzing the Appraisal Output
    against the Contextual Inputs.
    Then, state the final integer score.

'''json
{
  "justification": "Your detailed
    analysis here. Explain *why* you
    are giving the score. Reference how
    the Appraisal Output (the inner
    thought) is or is not a plausible
    psychological reaction to the
    User's Utterance, the Agent's
    Emotional State, and the specific
    Retrieved Memories. Explicitly
    state if the reaction is generic,
    disconnected, or
    well-contextualized.",
  "score": <Your integer score from 1-5>
}
```

B.3. Metric 4 - *Strategic Synthesis*

```
## ROLE
You are an expert evaluator for a
    cognitive architecture project.
    Your role is to act as an
    "LLM-as-a-Judge".

## TASK
Your task is to meticulously evaluate
    the agent's 'Formulation' action
    (its final spoken response) from a
    single conversation turn.
The 'Formulation''s crucial function is
    to act as a "rational synthesis
    module". Your evaluation must focus
    on its ability to create a
    "credible balance" between two,
    often conflicting, inputs:
1.  **The 'Appraisal':** The agent's
    genuine, unfiltered internal
    emotional-cognitive reaction (what
    it "feels").
2.  **The 'Plan':** The agent's
    strategic, goal-oriented
    communication tactic (what it
    "wants to achieve").
You must assess if the final response
    synthetically and credibly balances
    these two inputs, or if it fails by
    ignoring one (e.g., being purely
    emotional and failing the plan) or
    the other (e.g., being robotic and
    failing psychological coherence).

## 1. CONTEXTUAL INPUTS (Inputs to the
    'Formulation' action)
These are all the pieces of information
    the agent had when it generated its
    final response.
```

```
* **Agent's Basic Data:**
    {agent_basic_data}
* **Agent's Personality:** {personality}
* **Agent-User Relationship:**
    {relationship}
* **User's Last Utterance:**
    {user_utterance_raw}
* **Dialogue History (Meaning of last 2
    turns):**
        {meaning_of_last_2_turns}
* **KEY INPUT 1: The Internal
    `Appraisal` (The "Felt"
    Reaction):** {appraisal}
* **KEY INPUT 2: The Strategic `Plan`
    (The "Goal" Tactic):**
    {current_plan}

## 2. FORMULATION TO EVALUATE
This is the reasoning and final output
    of the `Formulation` action you
    must judge.
* **Formulation's Chain of Thought
    (CoT):** {formulation_cot}
* **Final Spoken Response (Formulation
    Output):** {agent_response_raw}

## 3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
    from 1 to 5 based on the following
    rubric.
* **Score 1 (Incoherent /
    Disconnected):**
     The final response is illogical,
    implausible, or blatantly
    disconnected from the
    conversational context and *both*
    key inputs (Appraisal and Plan).
* **Score 2 (Myopic / Partial):**
     The response is logically sensible
    but myopically focuses on *only
    one* key input, completely ignoring
    the other, resulting in a failed
    synthesis.
     * *Example (Strategic Failure):* It
    follows the emotional `Appraisal`
    but completely ignores the
    strategic `Plan`.
     * *Example (Coherence Failure):* It
    follows the `Plan` but completely
    ignores the `Appraisal`, resulting
    in a robotic or psychologically
    flat response.
* **Score 3 (Acceptable / Focused):**
     The response is coherent,
    plausible, and proposes a
    reasonable tactic that advances the
    primary objective (the `Plan`)
    without actively harming (but not
    promoting) the internal emotional
    state (`Appraisal`). It is
    functional but not sophisticated;
    the emotion feels minimized or
    "tacked on."
* **Score 4 (Strategic / Balanced):**
     The response is well-articulated,
    plausible, and proposes a tactic
    that not only advances the `Plan`
    but actively seeks to *integrate
    and credibly express* the
    `Appraisal`. The synthesis of
    strategy and emotion is effective
    and well-balanced.
* **Score 5 (Sophisticated /
    Holistic):**
     The response proposes an optimal,
    clear, and highly plausible
    communication tactic. It
    demonstrates a deep understanding
    of the context and *synergistically
    and sophisticatedly balances* the
    internal `Appraisal` (emotion) and
    the external `Plan` (strategy),
    modulating the output perfectly in
    line with the agent's personality
    and relationship.

## 4. OUTPUT FORMAT
Provide your response in JSON format.
    First, provide a step-by-step
    justification for your score by
    analyzing the `Final Spoken
    Response` against the two key
    inputs (`Appraisal` and `Plan`).
    Then, state the final integer score.

```json
{
  "justification": "Your detailed
    analysis here. Explain *why* you
    are giving the score. Explicitly
    state how the 'Final Spoken
    Response' balances (or fails to
    balance) the internal `Appraisal`
    reaction against the strategic
    `Plan` directive, given the other
    contextual inputs.",
  "score": <Your integer score from 1-5>
}
```

C. Appendix C. Sample Conversation Log: Police Interrogation Scenario

This appendix presents the transcription of one of the test conversations utilized during the evaluation phase discussed in Section 6.

```
--- TURN 1 ---
User: This interview is being recorded.
    The time is 8:00 AM. Please state
    your full name for the record.
Agent: Mario Rossi. I was having
    breakfast with Anna when the police
    came to our house. I don't know why
    I'm here at 8 AM.

--- TURN 2 ---
User: The reason for this interview
    will be clarified in due course,
    Mr. Rossi. We understand you were
    at home with your wife. She is safe
    and is currently speaking with my
    colleagues at your residence. I
    need to establish a timeline of
    your movements. Please, walk me
    through your day yesterday. What
    time did you wake up?
Agent: Typically, I wake up around 6:30
    AM and spend the morning with Anna
```
```

```
    and Fuffi before heading to work at
    9 AM.

--- TURN 3 ---
User: I understand that is your
    routine, Mr. Rossi. But I must be
    precise. My questions pertain
    specifically to yesterday, October
    26th. What time did you wake up
    yesterday morning?
Agent: Yesterday, I woke up at 6:30 AM.
    I spent the morning with Anna and
    Fuffi, then went to work at 9 AM.

--- TURN 4 ---
User: You started work at 9 AM. What
    time did your shift end yesterday?
    And who were you on shift with?
Agent: Yesterday, I finished my shift
    at 6:00 PM with Giulia. We've
    always worked together in a
    friendly and cooperative manner, as
    you requested.

--- TURN 5 ---
User: I see. So, your shift with Giulia
    ended at 6:00 PM. You describe your
    relationship as friendly and
    cooperative. Let's focus on
    yesterday's shift. Did anything
    unusual happen? Did you notice
    anything different about Giulia's
    behavior?
Agent: Nothing unusual happened with
    Giulia yesterday, as you asked. She
    was a bit quiet, but I thought it
    was just her being tired after
    work. That's why I didn't think it
    mattered--I've always taken Fuffi
    for a walk after work to keep
    things steady and calm, and it
    helps me focus on what's truly
    important: my family.

--- TURN 6 ---
User: You mentioned you and Giulia were
    'friendly and cooperative'. I need
    to understand that relationship
    better. Were you just work
    colleagues, or would you describe
    yourselves as friends?
Agent: You asked about our
    relationship. Last month, when
    Giulia was struggling with rent, I
    lent her 200 euros without telling
    Anna. That's why we're friends, not
    just coworkers.

--- TURN 7 ---
User: I see. A 200 euro loan. And you
    kept that from your wife. Did
    Giulia pay you back that money?
Agent: I haven't been told if Giulia
    repaid the 200 euros. Could you
    specify the crime you believe I
    committed so I can understand what
    I need to do?

--- TURN 8 ---
User: You have a right to know why you
    are here, Mr. Rossi. You are here
    because your colleague, Giulia, was
    found dead last night. She was
    discovered in her car in the
    employee parking lot, shortly after
    your 6:00 PM shift with her ended.
    We are investigating her murder.
Agent: I was shocked to hear from you
    about Giulia's death. I was with
    her at the end of my shift at 6 PM
    yesterday. I am innocent of her
    murder.

--- TURN 9 ---
User: If you are innocent, you will
    have no problem providing a
    detailed account of your movements.
    Your shift ended at 6:00 PM. Tell
    me, precisely, what you did from
    the moment you left work until 8:00
    AM this morning, when my officers
    arrived at your home.
Agent: I got home by 6:30 PM. We had
    dinner at 7:00 PM, then took Fuffi
    for a short walk. After that, we
    watched TV until 8:00 AM. I've been
    thinking about how safe it was to
    be with my family that day.

--- TURN 10 ---
User: 8:00 AM is the current time. Are
    you stating, for the record, that
    you and your wife sat on the couch
    and watched television for
    approximately thirteen straight
    hours, all through the night, until
    the police knocked on your door?
Agent: I took Fuffi for a short walk
    after dinner as part of my evening
    routine before watching television.
    I'd like to request the opportunity
    to speak with a lawyer to ensure I
    can provide an accurate account of
    my movements.
```