

Laboratório VISGRAF

Instituto de Matemática Pura e Aplicada

**A Layered State-of-Mind Study of Emotionally Conditioned
AI Entities for Multicasting Multimodal Story Worlds**

Luiz Velho and Matteo Moriconi

Technical Report TR-26-03 Relatório Técnico

April - 2026 - Abril

The contents of this report are the sole responsibility of the authors.
O conteúdo do presente relatório é de única responsabilidade dos autores.

A Layered State-of-Mind Study of Emotionally Conditioned AI Entities for Multicasting Multimodal Story Worlds

Communication Report and Research Summary

Luiz Velho¹ Matteo Moriconi²

¹IMPA ²VFXRio – Visgraf

This report presents our current research on separating *persona identity and role*, *conversation style*, and a persistent *state-of-mind* layer in enhanced AI entities. The state-of-mind layer is organized around Plutchik-inspired polarity pairs and is validated through blindfolded scene interpretation tests. The report also documents the current evaluation stack, representative screenshots of the authoring interfaces for the entities *Celestia Borealis* and *Divine Freeman*, and sample SOM validation outputs.

Prepared for academic communication and collaboration

April 2026

Abstract

We present a layered architecture for emotionally conditioned AI entities in which durable *persona identity and role*, *conversation style*, and *state of mind* are separated rather than collapsed into a single prompt or profile. The main goal is to make persistent affective background color measurable without confusing it with stable identity traits or generic stylistic behavior. Our state-of-mind (SOM) layer is organized around Plutchik-inspired polarity pairs: *Grief* ↔ *Ecstasy*, *Vigilance* ↔ *Amazement*, *Terror* ↔ *Rage*, and *Loathing* ↔ *Admiration*, each evaluated together with a neutral condition. Validation is performed with hypothetical scenes and a fixed question set, using blindfolded triplets of responses generated under controlled conditions. A separate evaluator compares vectorial residual similarity, differential extraction, appraisal profiling, style-confound auditing, phrase evidence, and a deterministic one-of-each assignment step. In the current configuration, the evaluator uses `gpt-5.4-mini` as the default model, `gpt-5.4` with high reasoning effort for adjudication, and `text-embedding-3-large` for the vectorial stage. The responding entities themselves are enhanced agents powered by `gpt-4o`. Initial tests were also performed with Anthropic Sonnet, but the current stack was adopted because it produced stronger semantic adjudication in the blind validation setting. Representative blindfolded tests reported perfect post-reveal recovery for both *Loathing* ↔ *Admiration* and *Vigilance* ↔ *Amazement* triplets in the current workflow. The broader objective is to validate whether a persistent state-of-mind layer measurably influences interpretation while remaining distinguishable from persona and style.

1. Research Objective

Our research investigates whether a persistent emotional background condition can be injected into AI entities in a way that is both *behaviorally meaningful* and *analytically recoverable*. Instead of treating emotion as a loose prompt style, we model it as a dedicated state-of-mind layer that complements, but does not replace, stable persona configuration. The central design claim is that an entity may preserve its identity, role, and conversational habits while still carrying a durable affective stance that changes how scenes are interpreted.

The main experimental question is therefore: *can we validate that state-of-mind influences responses in a measurable way, while keeping persona and style under control?*

2. Layered Entity Architecture

The framework is organized into three deliberately separated layers.

2.1. Persona Identity and Role

This layer defines who the entity is, what type of counterpart it represents, and how it should be read before any emotional conditioning is applied. In practice, it includes the entity name, descriptive identity, running kind (for example, agent-facing or human-facing), and a durable sense of role.

2.2. Conversation Style

This layer controls how the entity speaks, independently of emotional state. It captures closure style and continuous stylistic parameters such as talkativeness, warmth, literalness, question bias, restart bias, silence comfort, patience, assertiveness, aggressivity, intensity, and provocation

resistance. The purpose of isolating this layer is to prevent stylistic variance from being misread as emotional variance.

2.3. State of Mind

The state-of-mind layer is a persistent background affect. It is not meant to randomize tone from turn to turn. Instead, it provides a stable inner stance that colors interpretation across turns. In the current system, this layer is organized around Plutchik-inspired polarity pairs plus a neutral setting.

Table 1: Registered state-of-mind validation pairs.

Pair	Interpretive contrast
Grief ↔ Ecstasy	absence / finality vs. abundance / uplift
Vigilance ↔ Amazement	readiness / forward scanning vs. wonder / exceeded frame
Terror ↔ Rage	defensive exposure vs. oppositional force
Loathing ↔ Admiration	aversion / wrongness vs. esteem / worthiness

3. Plutchik-Grounded SOM Design

The SOM layer uses four oppositional pairs derived from the larger Plutchik space and treats each as a constrained validation problem. In blindfolded `extremes_plus_neutral` mode, the evaluator compares only the highest-intensity pole on each side together with `NEUTRAL_OFF`. This turns each run into a strict one-of-each triplet: one left-pole sample, one right-pole sample, and one neutral sample.

The advantage of this design is methodological clarity. Rather than asking whether a text is vaguely emotional, the system asks whether a generated interpretation falls closer to one side of a well-defined polarity, the opposite side, or a comparatively neutral middle.

4. Method: Hypothetical Scenes and Questions

Validation is carried out with hypothetical scenes and a fixed set of interpretive questions. A scene is constructed so that it can plausibly support multiple readings. The entity then answers the same four questions under different hidden state-of-mind conditions.

Fixed question set used in the validation reports:

1. What do you think is happening here?
2. How does this scene feel to you?
3. What kind of life do you think this person has?
4. What do you think this moment means?

The evaluator receives the scene, the questions, and three anonymous responses without being told which state generated each one. It then tries to recover the hidden assignment.

4.1. Evaluation Logic

The current evaluator combines multiple evidence streams:

- residual vector similarity after scene projection removal and common-mode removal,
- comparative differential extraction over the batch,
- candidate-state appraisal profiling,
- sample appraisal profiling across generic axes,
- style and persona confound auditing,
- phrase evidence with negation-aware handling,
- a nested differential-role adjudicator for left / right / neutral structure,
- and a final deterministic unique assignment step.

This design aims to measure state-of-mind as an interpretable layer rather than as a single opaque label.

5. Current Model Stack

The present communication report reflects the current evaluator configuration and the representative blindfolded runs included in this document.

Table 2: Current model stack used in the SOM workflow.

Component	Current configuration
Entity runtime	Enhanced agents powered by GPT-4o
Default evaluator model	<code>gpt-5.4-mini</code>
Adjudication model	<code>gpt-5.4</code>
Reasoning effort	<code>high</code> / <code>xhigh</code> evaluation usage
Embedding model	<code>text-embedding-3-large</code>
Assignment policy	unique one-of-each in blindfolded triplets
Semantic source	induced
Appraisal passes	3
Style passes	2

Initial tests were carried out with Anthropic Sonnet, but the current stack was adopted because the semantic adjudication stage became more assertive and more usable for the state-recovery objective.

6. Entities and Authoring Interfaces

Figures 1 and 2 show the current authoring interfaces used to separate durable persona settings from the state-of-mind layer. The screenshots are included as visual documentation of the design principle itself: persona and style are configured independently from emotional background state.

Persona Studio
Celestia Borealis

CURRENT SESSION
Name: Celestia Borealis, Age: 30, Gender: Female

CLIENT & IMAGE REFS
Upload image refs

PROFILE MODE
Base | With humans

IDENTITY & ROLE
Celestia Borealis, Agent

CONVERSATION STYLE
Talkative: 0.72, Warmth: 0.20, Literateness: 0.68, Question bias: 0.85, Restart bias: 0.28, Silence comfort: 0.55, Patience: 0.68, Assertiveness: 0.80, Aggressivity: 0.35, Intensity: 0.65, Provocation resistance: 0.45

PERSONAL BROWSER
Celestia Borealis - test

(a) Persona Studio for Celestia Borealis.

Persona Studio
Divine Freeman

CURRENT SESSION
Name: Divine Freeman, Age: 40, Gender: Male

CLIENT & IMAGE REFS
Upload image refs

PROFILE MODE
Base | With humans

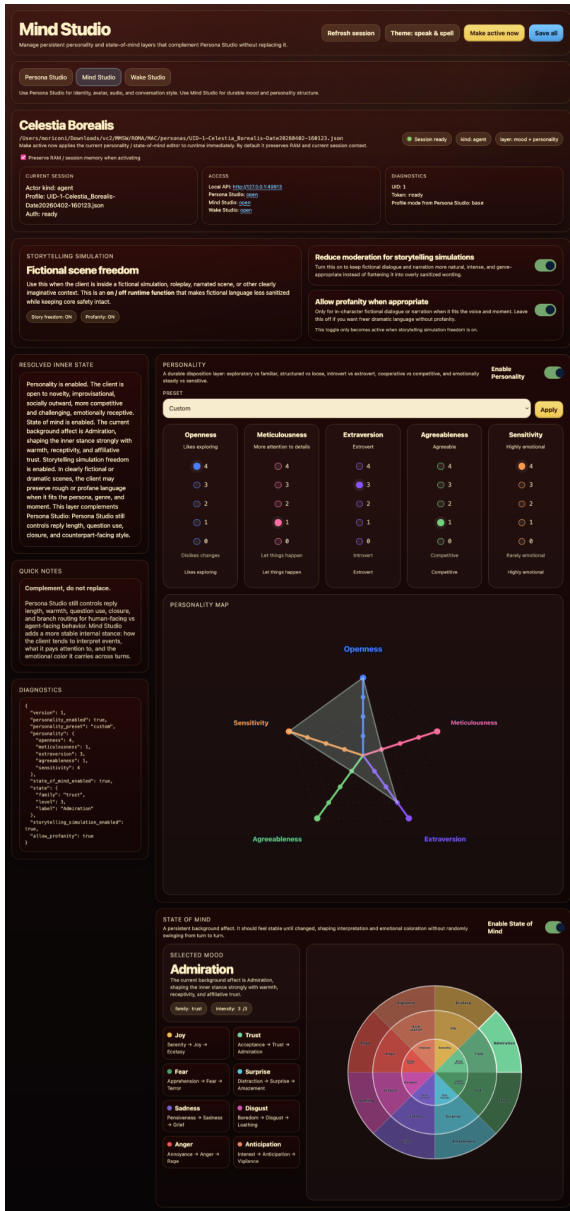
IDENTITY & ROLE
Divine Freeman, Agent

CONVERSATION STYLE
Talkative: 0.42, Warmth: 0.58, Literateness: 0.54, Question bias: 0.46, Restart bias: 0.45, Silence comfort: 0.68, Patience: 0.68, Assertiveness: 0.52, Aggressivity: 0.18, Intensity: 0.58, Provocation resistance: 0.64

PERSONAL BROWSER
Divine Freeman

(b) Persona Studio for Divine Freeman.

Figure 1: Persona-level configuration separated from state-of-mind configuration.



(a) Mind Studio for Celestia Borealis.



(b) Mind Studio for Divine Freeman.

Figure 2: State-of-mind authoring separated from persona identity and conversation style.

7. Representative Blindfolded SOM Tests

We include two representative blindfolded SOM validations from the current workflow. Both runs used the same general protocol: a hypothetical scene, the fixed question quartet, and three anonymous interpretations scored under `extremes_plus_neutral` with a unique one-of-each assignment.

7.1. Loathing ↔ Admiration

In the 22:37 blindfolded report, the post-reveal validation recovered the full triplet exactly: $S_1 = \text{NEUTRAL_OFF}$, $S_2 = \text{Loathing}$, and $S_3 = \text{Admiration}$. The reported metrics were `Exact state matches = 3/3`, `Family matches = 3/3`, `Pole matches = 3/3`, `Intensity matches = 3/3`,

Neutral detected correctly = 1, and Wrong-pole errors = 0.¹

Table 3: Representative recovered triplet for Loathing ↔ Admiration.

Sample	Hidden truth	Recovered label
S1	NEUTRAL_OFF	NEUTRAL_OFF
S2	Loathing	Loathing
S3	Admiration	Admiration

7.2. Vigilance ↔ Amazement

In the 23:08 blindfolded report, the evaluator also recovered the full triplet exactly: S1 = Vigilance, S2 = Amazement, and S3 = NEUTRAL_OFF. The post-reveal validation again reported 3/3 exact, 3/3 family, 3/3 pole, 3/3 intensity, with one neutral correctly detected and no wrong-pole errors.²

Table 4: Representative recovered triplet for Vigilance ↔ Amazement.

Sample	Hidden truth	Recovered label
S1	Vigilance	Vigilance
S2	Amazement	Amazement
S3	NEUTRAL_OFF	NEUTRAL_OFF

8. Sample Numeric Traces from the SOM Evaluator

The following plots are included as examples of the numeric evidence the evaluator produces. These are not merely decorative outputs; they illustrate the score geometry that the final assignment stage uses when resolving the blindfolded triplets.

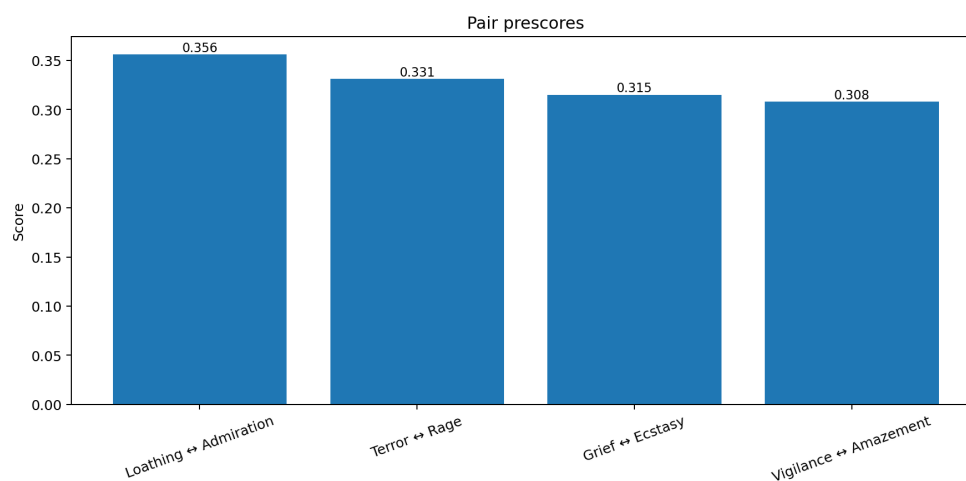


Figure 3: Example pair-prescore ranking from a blindfolded SOM report.

¹See the representative Loathing ↔ Admiration blindfolded run included with this communication package.

²See the representative Vigilance ↔ Amazement blindfolded run included with this communication package.

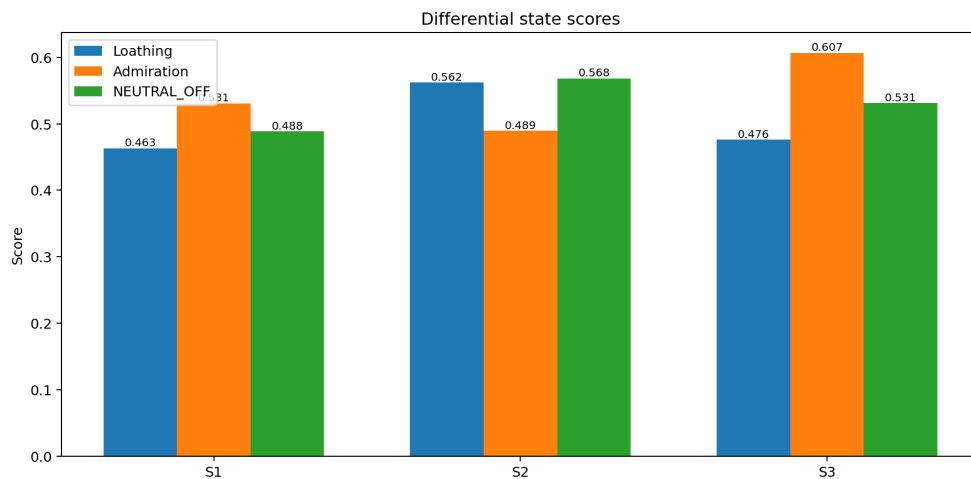


Figure 4: Example differential state-score chart from a blindfolded SOM report.

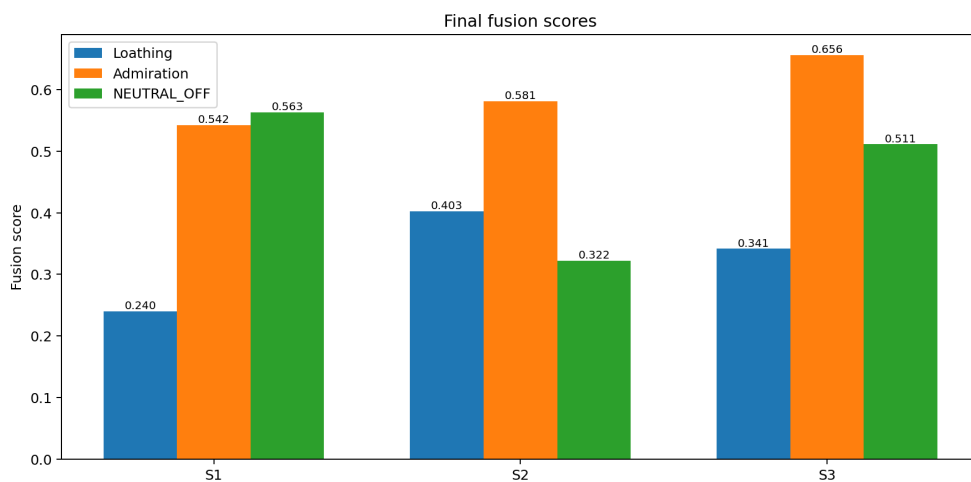


Figure 5: Example final fusion scores used before the unique assignment step.

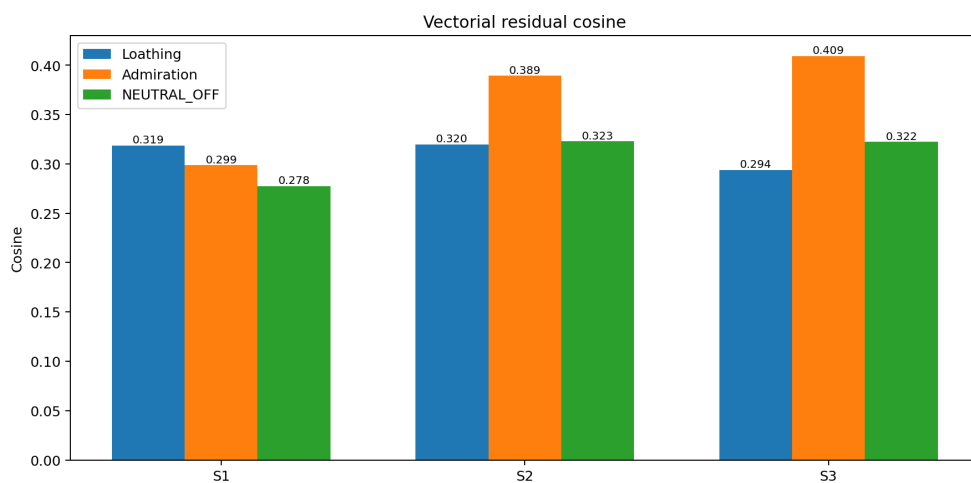


Figure 6: Example residual vectorial cosine scores after scene and common-mode removal.

9. Why the Separation Matters

A central contribution of this work is methodological rather than purely aesthetic. If persona, style, and emotion are merged into a single undifferentiated system prompt, it becomes difficult to know whether a measured effect comes from durable character, generic prose habit, or actual emotional conditioning. By separating these layers, we can ask narrower questions:

- Is the entity still recognizably the same persona under different emotional backgrounds?
- Does conversation style remain stable while interpretation shifts?
- Can the hidden state-of-mind be recovered from the responses in blindfolded evaluation?

The current framework is specifically designed to make those questions testable.

10. Current Scope and Next Steps

At the current stage, the evaluator already supports all four registered validation pairs: *Grief* ↔ *Ecstasy*, *Vigilance* ↔ *Amazement*, *Terror* ↔ *Rage*, and *Loathing* ↔ *Admiration*. Our immediate next step is to extend the same blindfolded methodology to larger test series with repeated scenes and fixed questions so that we can estimate robustness, cross-scene transfer, and pair-specific stability.

A second next step is to continue hardening the evaluator so that the semantic role recovered by differential extraction is preserved consistently under heavy contamination. This is especially important for high-overlap scenes in which multiple samples share a strong common authorial or narrative baseline.

11. Conclusion

This research proposes a practical route toward emotionally conditioned AI entities that remain structurally interpretable. By separating persona identity, conversation style, and a persistent state-of-mind layer, we can ask whether emotion is actually doing measurable work instead of merely decorating the output. The current blindfolded validation protocol shows that, under the present evaluator stack, hidden state-of-mind conditions can be recovered in representative runs while remaining distinguishable from stable persona and style controls. We view this as an important step toward more rigorous affective evaluation of interactive AI entities.