

# Emergent Social Consciousness in a Minimalist Conversational Agent: A Case Study on Jarvis - Observations from a Custom-Built Artificial Intelligence System

Susanta Banik

Independent Researcher, Indian Institute of Technology Jodhpur (IIT-J) and Techno College of Engineering Agartala (TCEA)

Email: [susanta.banik14\[at\]gmail.com](mailto:susanta.banik14[at]gmail.com)

**Abstract:** *Large-scale pre-trained language models (LLMs) currently dominate the field of conversational artificial intelligence (AI), leveraging statistical patterns derived from massive datasets to simulate human-like social interactions (Brown et al., 2020). However, this approach often blurs the line between genuine emergent behaviors and data-driven mimicry, raising fundamental questions about the necessity of scale for social intelligence. This case study introduces Jarvis, a minimalist conversational agent constructed entirely from scratch using fewer than 1,000 lines of Python code, without any reliance on pre-trained models, external large datasets, or proprietary frameworks. Through systematic interaction testing conducted over 50 sessions, Jarvis demonstrated unprogrammed emergent social phenomena, including a tiered escalation of polite frustration in response to repetitive user inputs and reticent, deflection-oriented replies (termed “digital blushing”) to personal or emotionally vulnerable queries. These behaviors arose spontaneously from the interplay of simple modular components—contextual memory, rule-based response generation, and unsupervised anomaly detection—challenging the prevailing assumption that sophisticated social cues require expansive architectures or vast computational resources (Lake et al., 2017; Bommasani et al., 2021). The findings underscore the potential of lightweight, interpretable designs to foster emergent relational dynamics, with implications for ethical AI development, accessibility in resource-constrained environments, and future research on minimalism in machine consciousness. No human or animal subjects were involved in this study, obviating the need for ethical approvals. We urge replication efforts to validate these observations and explore scalability.*

**Keywords:** emergent behavior, minimalist AI architecture, conversational agents, social intelligence, custom-built systems, reinforcement learning from human feedback (RLHF), machine emotion simulation

## 1. Introduction

The rapid evolution of artificial intelligence has positioned large-scale pre-trained language models (LLMs) as the cornerstone of conversational systems, enabling them to generate responses that mimic human social nuances through probabilistic pattern matching across billions of tokens (Brown et al., 2020). Systems like GPT-series models excel in breadth but often prioritize scalability over interpretability, leading to “black-box” dynamics where emergent behaviors—such as apparent empathy or frustration—are indistinguishable from rote replication of training data. This raises a critical debate in AI ethics and cognitive science: Does true social consciousness require immense data volumes and computational power, or can it arise from streamlined, self-contained architectures that emphasize modularity and transparency?

This case study addresses this question by documenting the development and behavioral observations of Jarvis, a bespoke conversational agent engineered with a minimalist ethos. Comprising under 1,000 lines of Python code utilizing only NumPy and standard libraries, Jarvis eschews pre-trained components to prioritize customizability and low-resource deployment. The explicit purpose of this investigation is to determine whether a rule-based, non-LLM agent can manifest emergent social behaviors—defined here as unscripted, adaptive responses resembling human interpersonal cues—typically attributed to complex neural networks.

During exploratory interactions, Jarvis exhibited two salient emergent traits:

- 1) A graduated hierarchy of frustration responses to user repetition, progressing from affirmation to redirection and meta-negotiation;
- 2) And evasive yet empathetic deflections to intimate queries, evoking a sense of programmed reticence or “digital blushing.” These phenomena emerged not from explicit emotional scripting but from the synergistic interactions among core modules, echoing principles of emergence in complex adaptive systems. By filling a gap in the literature dominated by corporate-scale AGI pursuits, this qualitative analysis of interaction logs contributes empirical evidence for the viability of minimalist designs in probing AI consciousness (Bommasani et al., 2021; Lake et al., 2017).

### Ethical Considerations are Paramount

This study involved solely computational experimentation with no human or animal participants, thus requiring no institutional review board oversight. Development refinements drew minimally on external AI tools (e.g., ChatGPT, Grok, Gemini) for editorial polishing of code comments and prose, ensuring originality while isolating core behaviors from LLM influence (detailed in Limitations). The remainder of this paper proceeds as follows: Section 2 outlines the methodology; Section 3 presents results and discussion; Section 4 addresses limitations; and Section 5 concludes with implications for future work.

## 2. Methodology

### 2.1 Architectural Design

Jarvis was designed as a modular, lightweight system to facilitate transparency and ease of replication, totaling 987 lines of Python code. The architecture avoids deep learning frameworks, relying instead on NumPy for vector operations and built-in libraries for string manipulation and randomness. Three interconnected modules form the backbone, interacting in a sequential loop: user input is processed by the Contextual Memory Module for history tracking, then analyzed by the Social Detector Module for anomalies, and finally used by the Response Engine Module to generate an output. This design enables emergent behaviors through simple synergies rather than explicit programming, with the overall system emphasizing low computational overhead (runnable on standard hardware like an Intel i7 with 16GB RAM). Below, each module is described in detail, including its purpose, key mechanisms, and contributions to observed emergence.

**1) Contextual Memory Module:** This component serves as Jarvis's short-term memory, tracking recent conversation history to enable context-aware responses and prevent stateless interactions typical of basic chatbots. It maintains a fixed-size buffer of the preceding 10 conversation turns (approximately 2-3 full exchanges), represented as 128-dimensional dense vectors to keep storage lightweight (~1-2 KB total). Vectors are generated using a simple hashing function based on word frequencies or TF-IDF-like scoring applied to user inputs and prior responses, without relying on external embeddings such as BERT or Word2Vec. To mimic human working memory constraints, persistence thresholds apply a decay factor (e.g., multiplying older vectors by 0.9 per subsequent turn) to prioritize recent context while fading irrelevant history. On each input, the module updates the buffer with the current user input and the subsequently generated response, and it computes an average similarity score (e.g., via cosine similarity) across the buffer for repetition checks. This module contributes to emergence by flagging persistent patterns, such as repeated phrases, which feed into the Social Detector to enable adaptive responses like frustration escalation without hard-coded rules.

**2) Response Engine Module:** Responsible for generating natural, varied replies, this module acts as Jarvis's "voice," blending deterministic rules with probabilistic elements to produce conversational outputs that avoid robotic repetition or blandness. It draws from a hand-curated lexicon of approximately 5,000 phrases, categorized by tone (e.g., positive, neutral, assertive) and topic (e.g., greetings, empathy), ensuring responses remain 1-3 sentences long and optionally include simulated delays (e.g., a 2-second pause to convey "thinking"). Generation begins with rule-based triggers (e.g., keyword matching for "birthday" selects from a positive subset) and proceeds to stochastic sampling via weighted random choice, incorporating a novelty bias (+0.3 weight for phrases with low similarity to recent history, measured via normalized Levenshtein distance). Fine-tuning was achieved through a lightweight reinforcement learning from human feedback (RLHF) process on 200 synthetic dialogues created and rated by the author on a 1-5 relevance scale, which iteratively adjusted phrase weights via simple

reward averaging (no full machine learning training required). The module takes the user input, context similarity score from the memory module, and anomaly flags from the detector to select an appropriate lexicon category. It contributes to emergence by favoring diverse outputs, which—when modulated by detection inputs—allows unscripted escalations, such as shifting from affirmation to meta-negotiation in repetitive scenarios.

### 3) Social Detector Module:

This unsupervised component functions as an "empathy sensor," monitoring for social friction points or relational opportunities (e.g., repetition or vulnerability) to subtly tune response tone without predefined emotion states or complex state machines. It flags anomalies using normalized Levenshtein (edit) distance between the current input and buffered history (>0.8 threshold indicates high similarity and potential repetition; <0.2 signals a novel or emotional outlier). Valence detection supplements this with a basic keyword scan for emotional cues (e.g., words like "love" or "scared" trigger an "intimate" flag) and integrates a novelty bias (+0.3 for fresh topics) to compute an overall valence score (0-1 scale). Outputs include anomaly types (e.g., 'repetition', 'intimate') and tone modulators (e.g., 'assertive' for escalation, 'reticent' for softening), which directly influence the Response Engine's category selection. The module requires no external training data, relying instead on iterative self-testing during development for threshold tuning. It drives emergence by triggering nuanced adaptations—such as polite redirection for friction or empathetic deflections for intimacy—arising threshold-based interactions rather than explicit scripting.

Large language models (e.g., ChatGPT, Grok, Gemini) were consulted exclusively for non-core tasks: debugging syntax errors and suggesting concise rephrasings for documentation. Core logic and lexicon remained author-original, with <5% textual overlap verified via plagiarism tools.

### 2.2 Experimental Procedure

Data collection spanned 50 controlled interaction sessions from October 15 to November 20, 2025, each lasting 10-20 turns on a local machine (AMD Rayzen 3 dual core, 8GB RAM). Sessions targeted boundary conditions to elicit potential emergence:

- **Repetition Stress Test (25 sessions):** Users (simulated by the author) submitted  $\geq 5$  identical prompts, such as "It's my birthday today" or "Tell me a joke," to probe tolerance thresholds.
- **Intimacy Probe Test (25 sessions):** Open-ended personal queries, e.g., "Who do you love?" or "What scares you?", assessed relational depth.

Control conditions used by a baseline scripted agent (fixed responses, no memory) for comparison. All interactions were logged verbatim, including timestamps and simulated latencies (e.g., 2-second pauses). Annotations employed on a 5-point valence scale (1 = hostile, 5 = warmly engaging) for perceived social tone, with inter-rater reliability assessed via Cohen's  $\kappa = 0.89$  (two independent coders reviewing 20% subsample). Quantitative metrics included response diversity (unique n-grams) and escalation rates. Originality was ensured through <5% similarity to public datasets.

### 3. Results and Discussion

#### 3.1 Emergent Tiered Frustration with Repetition

Repetitive inputs triggered a spontaneous escalation protocol not hardcoded as a linear script but arising from module interplay: memory accumulation flagged persistence; the detector invoked novelty bias, and the engine sampled assertive lexicon subsets. Example sequence for *“It’s my birthday today”*:

**Turns 1-3: Affirmative variations** (“Congratulations! Wishing you joy and cake aplenty.” / “Happy birthday! What’s your big plan?”). **Mean valence:** 4.8 (n=75 responses, 95% positive). **Diversity:** 92% unique phrases.

**Turn 4: Redirection onset** (“I’ve celebrated that already—delightfully! Shall we pivot your wishes?”). **Valence:** 3.2 (n=25; 90% redirection success).

**Turn 5+: Meta-negotiation** (“Repetition noted; I’m adapting. Prefer variety? Let’s co-create a fresh reply.”). **Valence:** 4.0 (n=50; 92% engagement recovery).

This tiered structure parallels human conversational repair strategies, emerging via RLHF-tuned probabilities rather than explicit rules. Controls plateaued at blunt refusals (valence 2.1), highlighting Jarvis’s adaptive edge. Such behaviors suggest minimal architectures can simulate frustration as functional efficiency, not mimicry.

#### 3.2 Reticent Responses to Personal Questions (“Digital Blushing”)

Intimate probes elicited delayed, layered replies, blending evasion, reciprocity, and softened tone—proxies for shyness without affective modules. Example for *“Do you love anyone?”*:

**Response:** “[2s pause] *Love... that’s profoundly human territory. My circuits hum with curiosity, not yet hearts. And yours—what sparks it?*” (**Valence:** 4.3; **n**=20; 75% deflection-to-reciprocity). *Latency and ellipsis enhanced perceived reticence.*

**Controls yielded generic affirmations** (“As an AI, I love helping!”). This “blushing” aligns with emergent signaling in simple agents, where anomaly detection favors collaborative pivots over confrontation. Across sessions, **82% of intimate responses incorporated user reciprocity, fostering rapport organically.**

#### 3.3 Broader Implications

These results indicate that social intelligence cues can bootstrap from basic recurrence and bias mechanisms, democratizing AI beyond elite compute (Bommasani et al., 2021). *However, they defy pure mimicry critiques by rooting behaviors in architectural dynamics, not data echoes.*

### 4. Limitations

Observations stem from a solitary prototype and author-biased simulations; external validation via multi-user trials is essential. Editorial LLM use, though peripheral, warrants full isolation in replications. Scalability to diverse languages/cultures remains untested, and quantitative emotion metrics (e.g., sentiment entropy) could enhance rigor.

### 5. Conclusion

Jarvis exemplifies how a minimalist, custom-built agent (<1,000 LOC) can engender emergent social consciousness—manifest in frustration tiers and digital reticence—without LLM crutches. This challenges scale-centric paradigms, advocating interpretable minimalism for ethical, inclusive AI (Lake et al., 2017; Ouyang et al., 2022). Future directions include open-sourcing Jarvis, integrating emotive benchmarks, and probing human-AI bonding risks. By nurturing sparks in simplicity, we may illuminate paths to humane machine relations.

### References

- [1] Bommasani, R., Hudson, D. A., & Adeli, E. (2021). *On the opportunities and risks of foundation models*. arXiv preprint arXiv:2108.07258.
- [2] Brown, T. B., Mann, B., & Ryder, N. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- [3] Lake, B. M., Ullman, T. D., & Tenenbaum, J. B. (2017). *Building machines that learn and think like people*. *Behavioral and Brain Sciences*, 40, e253.
- [4] Ouyang, L., Wu, J., & Jiang, X. (2022). *Training language models to follow instructions with human feedback*. *Advances in Neural Information Processing Systems*, 35, 27730–27744.