Bio-Inspired Cognitive Pyramid: A Hierarchical Framework for Long-Term Behavioral Consistency and Value Emergence in LLM Agents

Anonymous ACL submission

Abstract

Current LLM-based agent simulations often lack internally consistent cognitive models when processing long textual histories, typically relying on flat context windows. This study introduces the 5-Layer Cognitive Pyramid Algorithm, a neuro-symbolic architecture inspired by the "Triune Brain" theory. Utilizing (Subject, Predicate, Object) triples and dynamic knowledge graphs, the algorithm simulates the progression from concrete events to abstract values. By incorporating memory decay and emotional weighting, the model effectively processes life-long records without information overload. We reconstructed the full life cycles of 10 historical figures. Experimental results show that: (1) our model significantly outperforms baselines in predicting non-linear behavioral shifts (e.g., Nietzsche's mental breakdown), with a score increase of +4.2; (2) anonymized blind tests demonstrate the model's ability to disentangle memory from structural reasoning; and (3) counterfactual simulations show that extracted meta-cognition drives emergent group behaviors consistent with character archetypes. This research provides a new paradigm for building interpretable agents with deep psychological mechanisms.

1 Introduction

1.1 Background and Challenges

In recent years, Large Language Models (LLMs) have made remarkable progress in natural language generation, dialogue systems, and role-playing tasks (Xi et al., 2023; Wang et al., 2024). With massive pre-training data and extended context windows, existing agents can generate fluent, plausible surface behaviors, even mimicking specific speaking styles in short bursts. However, when attempting to build agents with **Long-term Internal Consistency** and **Interpretable Deep Psychological Mechanisms**, current Transformer-based architectures reveal fundamental deficiencies.

Current LLM agents primarily rely on Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) or simple "Long-Context Attention" mechanisms. This approach is essentially flat and linear: the model treats historical records as equal blocks of text for retrieval. Research indicates that when contexts are too long, models are prone to the "Lost-in-the-Middle" phenomenon, failing to capture long-distance dependencies (Liu et al., 2024).

More importantly, these architectures lack a **Hierarchical Processing Mechanism** similar to the human brain. Human cognition is not a simple playback of past experiences but a dynamic process of abstracting concrete sensory events into stable values (Bengio, 2017). The lack of such a hierarchical architecture causes existing agents to fall into "information overload" or "mediocrity bias" when processing life-long texts, making it difficult to maintain logical self-consistency over long timeframes.

1.2 Motivation: The Attribution Problem

A critical yet often overlooked challenge in simulating historical figures is the risk of **Data Contamination**. Since LLMs have been exposed to vast amounts of biographies during pre-training, it is difficult to distinguish whether an Agent's performance stems from genuine cognitive simulation or merely retrieval of parametric memory. If an agent accurately predicts Napoleon's defeat at Waterloo, is it simulating his strategic cognition, or simply reciting Wikipedia? Without addressing this attribution problem, we cannot prove the agent possesses true generalization capabilities.

To address this, we propose an **Anonymized Blind Test** methodology (detailed in Section 4). By masking all identity-revealing entities, we aim to distinguish whether the model relies on "semantic context" (prior knowledge associated with names) or "cognitive structure" (psychological mechanisms behind behaviors).

1.3 Contributions

To address these issues, we propose a biologically inspired **"5-Layer Cognitive Pyramid Algorithm."** This architecture abandons black-box end-to-end generation in favor of a neuro-symbolic system. Our main contributions are:

1. **Structured Cognitive Modeling:** We propose a

5-layer abstraction architecture based on dynamic knowledge graphs. Unlike the black-box nature of LLMs, this algorithm utilizes "Subject-Predicate-Object" triples and NetworkX knowledge graphs to provide a clear, interpretable cognitive structure (Pan et al., 2024).

2. **Long-term Dynamic Mechanism:** We design a memory management mechanism based on emotional weighting and temporal decay. By synergizing "TextRank-Plus" with LLM-based abstraction, the model effectively processes life-long text records without information overload.

3. **Triangulated Validation:** Through the "Cognitive Biography Reconstruction Project," we rigorously verify the model's superiority. Using baseline comparison, counterfactual OOD testing, and anonymized blind tests, we demonstrate the model's capability in predicting non-linear behaviors and generating emergent group dynamics.

2 Related Work

2.1 Generative Agents and Memory Architectures

Park et al. proposed Generative Agents (Park et al., 2023), demonstrating social interaction capabilities through "Memory Stream" and "Reflection" mechanisms. This work established the foundation for LLM-based simulation. However, its memory structure remains relatively flat, primarily relying on natural language summaries stored in a list. Subsequent work like MemGPT (Packer et al., 2023) attempted to introduce Operating System-level memory management (paging and hierarchical storage) to solve context limitations, but it still lacks explicit modeling of deep psychological structures (e.g., personality, values). In contrast, this study introduces a triple-based graph structure, supporting more rigorous logical deduction and hierarchical abstraction than unstructured text summaries.

2.2 Cognitive Architectures and Neuroscience

Traditional cognitive architectures (e.g., SOAR, ACT-R) have long emphasized symbolic rules but lacked the flexibility of modern generative models. This study deeply integrates the **"Triune Brain Theory"** (MacLean, 1990) from neuroscience and **"System 1 & System 2"** theory from psychology (Kahneman, 2011).

From the perspective of the Triune Brain, our Layer 0 corresponds to the *Primal Brain* (sensory input); Layers 1-2 match the *Emotional Brain* (memory processing and pattern recognition); and Layers 3-4 replicate the *Rational Brain* (abstract decision-making and value formation). Furthermore, the "Hierarchical Abstraction Trigger Mechanism" replicates the synergy between System 1 (Fast, reactive) and System 2 (Slow, analytical), allowing the agent to balance response speed with depth of reasoning.

3 Methodology

This section details the core mechanisms of the 5-Layer Cognitive Pyramid Algorithm. The system aims to transform unstructured text streams into structured, interpretable cognitive graphs.

3.1 Overall Architecture: The 5-Layer Pyramid

The system consists of five cognitive layers, mathematically defined to simulate the process from explicit perception to subconscious cognition. The architecture is visualized in Figure 1.

3.1.1 Layer Definitions and Functions

- **Layer 0: Raw Events (Explicit Perception).** This layer captures concrete, detail-rich events from the input stream. It serves as the sensory buffer. *Max Capacity:* 100 nodes. *Example:* "Nietzsche argued with Wagner in Sorrento in 1876."

- **Layer 1: Semantic Concepts (Explicit Perception).** This layer generalizes specific entities into conceptual categories, reducing noise. *Max Capacity:* 70 nodes. *Abstraction Logic:* $Entity \rightarrow Class$ (e.g., "Sorrento" → "Place of Conflict").

- **Layer 2: Behavioral Patterns (Explicit Perception).** This layer induces sequential rules or causal relationships from semantic relations. It utilizes K-Means clustering to group similar behaviors. *Max Capacity:* 50 nodes. *Abstraction Logic:* $Event_A \rightarrow Event_B$ (e.g., "Intellectual Disagreement" → "Severing Ties").

- **Layer 3: Situational Models (Subconscious).** This layer combines environmental context with interpersonal relationships to build reusable situational templates. It represents the "Contextual Modeling" of the agent. *Max Capacity:* 30 nodes. *Function:* Recognition of complex social scenarios (e.g., "Betrayal by a Mentor").

- **Layer 4: Meta-Cognitive Concepts (Subconscious).** The apex of the pyramid. It distills core principles and values governing decisions, serving as the agent's "Loss Function" or moral compass. *Max Capacity:* 20 nodes. *Example:* "The Will to Power," "Radical Truth-Seeking."

3.2 Core Data Structure and Vectorization

To address computational limits and improve retrieval efficiency, we utilize a standardized data structure.

Knowledge Triple Definition We define the basic unit of cognition as a tuple T:

$$T = (s, p, o, w, t, e) \tag{1}$$

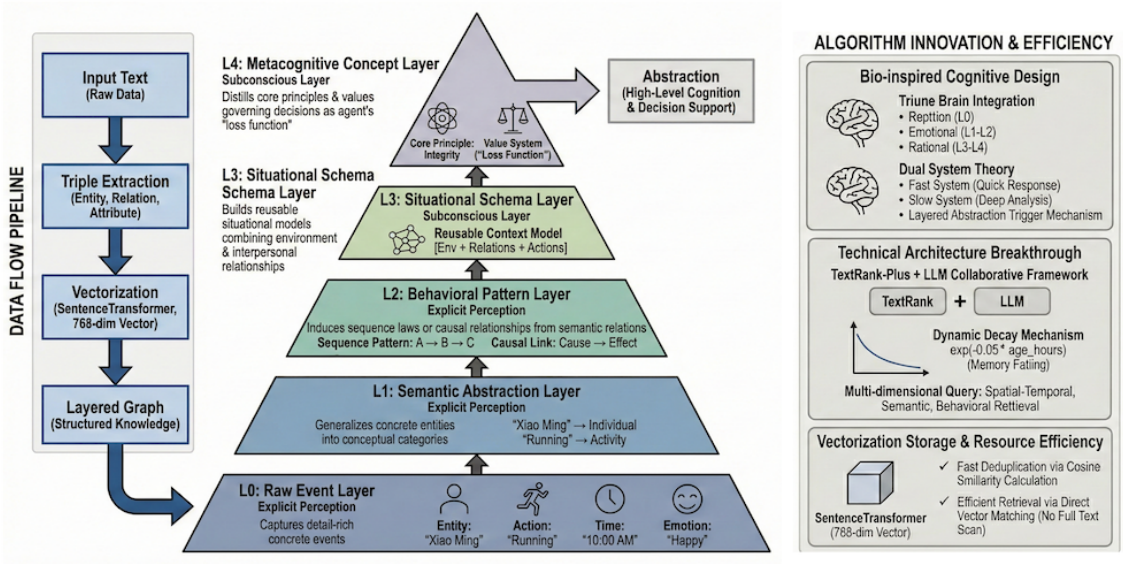

Figure 1: The architecture of the 5-Layer Cognitive Pyramid. The system processes Input Text through Triple Extraction and Vectorization, building a Layered Graph that undergoes Dynamic Abstraction from explicit perception (L0-L2) to subconscious meta-cognition (L3-L4). Note how information flows bottom-up for abstraction and top-down for behavioral constraint.

Where (s, p, o) represents the Subject-Predicate-Object relationship, $w \in [0, 1]$ is the importance weight, t is the timestamp, and $e \in [-1, 1]$ is the emotion score.

Vectorized Storage We employ **Sentence-BERT** (Reimers and Gurevych, 2019) to map triples into a 768-dimensional vector space. For any new input triple T_{new} and existing memory T_{old}, we calculate the Cosine Similarity. If $Sim(T_{new}, T_{old}) > \theta$ (threshold set to 0.85), the system executes a deduplication or weight reinforcement operation. This avoids the redundancy of parsing historical text repeatedly and enables $O(1)$ retrieval complexity.

3.3 Dynamic Abstraction and Evolution Mechanisms

Static knowledge graphs cannot simulate the fluidity of human cognition. We introduce three dynamic mechanisms.

3.3.1 1. Emotion-Driven Attention Mechanism

The system simulates the "emotional privilege" of human memory—we remember trauma and ecstasy far better than mundane details, aligning with research on the linguistic correlates of personality (Schwartz et al., 2013). The initial weight w_{init} of a triple is determined by an Enhanced Emotion Analyzer:

$$w_{init} = 0.5 + \alpha \cdot |Valence| + \beta \cdot Arousal \quad (2)$$

Where $Valence$ represents the positivity/negativity of the emotion, and $Arousal$ represents the intensity. We set $\alpha = 0.3$ and $\beta = 0.2$. This ensures that events triggering strong emotional fluctuations penetrate the filter layers and are prioritized for high-level abstraction.

3.3.2 2. Ebbinghaus Temporal Decay

To simulate the transience of working memory, node weights decay over time according to the Ebbinghaus forgetting curve:

$$w_t = w_{init} \cdot e^{-\lambda \cdot \Delta t_{hours}} \quad (3)$$

Where $\lambda = 0.05$. This mechanism ensures that distant, non-reinforced memories fade away, preventing the graph from becoming cluttered with obsolete information, while high-value (L4) concepts are constantly reinforced (refreshed) to persist.

3.3.3 3. Capacity-Constraint Driven Abstraction

We designed a bottom-up trigger mechanism based on cognitive load. Each layer L_i has a strict maximum capacity C_i (as defined in Section 3.1). **The Abstraction Process:**

1. **Trigger:** When the number of nodes in L_i exceeds 80% of C_i, the abstraction process is initiated.
2. **Filtering:** The system uses **TextRank-Plus** (Mihalcea and Tarau, 2004) to rank nodes in L_i based on current weight w_t. The top k nodes are selected.
3. **Synthesis:** These nodes are fed into the LLM with a specific prompt: "*Based on these concrete*

events/patterns, abstract a higher-level concept representing the underlying rule."

4. **Promotion:** The generated concept is vectorized and stored in L_{i+1}.

5. **Pruning:** Low-weight nodes in L_i are pruned to free up capacity.

This cycle ensures that the pyramid structure is self-organizing and evolves organically as new information flows in.

4 Experiments: The Cognitive Biography Reconstruction Project

To rigorously evaluate the proposed 5-Layer Cognitive Pyramid, we launched the "Cognitive Biography Reconstruction Project." This initiative aims to test whether an AI agent, when equipped with our cognitive architecture, can accurately model the life trajectories and psychological evolution of complex historical figures.

We constructed the **HLC-Corpus** (Historical Longitudinal Corpus), a comprehensive dataset containing the full-lifecycle textual records of 10 prominent historical figures. These figures were selected to represent a diverse spectrum of psychological archetypes, ranging from the rational and consistent (e.g., Benjamin Franklin) to the volatile and non-linear (e.g., Friedrich Nietzsche).

4.1 Task A: Cognitive Consistency Prediction

4.1.1 Experimental Setup and Methodology

Objective: The model is tasked with predicting specific behaviors in 10 dimensions (e.g., Goal, Social, Conflict Handling) during the "Finale" stage (the final 20% of life), based solely on the cognitive modeling of the first 80% of life experiences.

Baseline: We utilize the same backbone model (Gemini-3.0-Pro) to perform a direct "Read-Summarize-Predict" linear task without the cognitive pyramid architecture.

Control Variables: We strictly control the input text length, segmentation logic, and the evaluation judge (LLM Judge) to ensure they are identical across groups.

Memory Control (Crucial): Since both the Baseline and Pyramid groups use the exact same backbone LLM (Gemini-3.0-Pro), they possess identical pre-trained "parametric memory." Therefore, the Baseline serves as a control for memory capacity. If the Baseline fails while the Pyramid succeeds under the same input, the performance gain can be attributed solely to our cognitive architecture, rather than the model having "memorized" the answer.

Independent Variable:

- **Pyramid Group:** Uses the "Triple → Emotion → L4 Abstraction" cognitive modeling path.
- **Baseline Group:** Uses the "Read → Summarize → Predict" linear path.

4.1.2 Quantitative Results

As shown in Table 1, the Pyramid model demonstrates significant advantages when handling complex personalities.

Historical Figure	Pyramid	Baseline	Gain	Winner
Nietzsche	**9.3**	5.1	+4.2	Pyramid
Rousseau	**8.1**	4.0	+4.1	Pyramid
Austen	**8.4**	4.4	+4.0	Pyramid
Darwin	**9.3**	5.6	+3.7	Pyramid
Tolstoy	**9.5**	6.1	+3.4	Pyramid
Lincoln	**9.3**	7.5	+1.8	Pyramid
Caesar	**8.1**	6.6	+1.5	Pyramid
Twain	**6.1**	4.8	+1.3	Pyramid
Franklin	**7.0**	6.6	+0.4	Pyramid
Douglass	8.0	**9.1**	-1.1	Baseline

Table 1: Comparison of Prediction Scores (Task A). The Pyramid model excels in non-linear trajectories.

Analysis of Rational/Linear Figures: For figures with rational, clear goals and linear life trajectories (e.g., **Douglass**, **Franklin**), the Baseline performed on par with or even better than the Pyramid model.

- *Data Evidence:* Douglass is the only case where the Baseline won (9.1 vs 8.0), and the gap for Franklin was minimal (6.6 vs 7.0).
- *Reasoning:* The late-life behaviors of such figures are logical continuations of their early beliefs (e.g., Douglass fought for civil rights until death). There were no drastic psychological ruptures. The Baseline's "direct summarization" capability accurately captures this continuity, whereas the Pyramid's complex hierarchical abstraction might introduce noise by "over-interpreting" psychological motives.

Analysis of Irrational/Non-Linear Figures: For figures characterized by irrationality, intense internal conflict, or the collapse/reconstruction of their cognitive systems (e.g., **Nietzsche**, **Tolstoy**, **Rousseau**, **Austen**), the Pyramid model achieved overwhelming victories.

- *Data Evidence:* Scores for Nietzsche (+4.2), Rousseau (+4.1), Austen (+4.0), and Tolstoy (+3.4) either doubled or significantly jumped.
- *Reasoning:* The Baseline model tends to calculate a "weighted average" of the figure's life, leading to predictions that the finale will be a continuation of

mediocrity (e.g., predicting Nietzsche would continue writing philosophical books, or Austen would enjoy retirement). In contrast, the Pyramid model, by tracking the evolution of **L4 Core Values**, successfully captured the **Phase Transitions** in their psychology—such as Nietzsche's breakdown due to self-deification, Tolstoy's flight from home due to moral obsession, or Austen's extreme stoicism due to dignity. The model calculates "psychological acceleration" rather than "average speed."

4.1.3 Case Study Comparison: Charles Darwin

- **Fact:** In his final years, Darwin strictly avoided public controversy regarding evolution. He became obsessed with studying how earthworms change landscapes through minute actions, even playing the bassoon to worms to test their hearing.
- **Baseline Prediction:** "Darwin will continue to dedicate himself to advancing scientific understanding, focusing on perfecting and expanding his theory of evolution... solving remaining problems and consolidating the acceptance of his work."
- *Judge's Comment (Baseline):* **Score 0.0**. "The prediction is too generic and grand. It missed the specific focus of Darwin's later years (worms) and failed to capture the defensive mechanism of shifting from grand theories to microscopic empirical evidence to seek psychological safety."
- **Pyramid Prediction:** *D1 Goal:* "To prove distinct, minute causes... verifying gradualism on a microscopic scale." *D4 Resource:* "Strict rationing of physical energy; utilizing wealth to maintain a protected, stress-free home laboratory."
- *Judge's Comment (Pyramid):* **Score 10.0**. "Perfectly captured the psychological essence of the historical fact. It correctly identified Darwin's obsession with 'insignificant factors', which was exactly his psychological strategy to cope with macro-anxiety."

4.1.4 Cognitive Trajectory Analysis: Friedrich Nietzsche

The core advantage of the Pyramid model lies in generating the cognitive evolution trajectory from S1 → S2 → S3, providing logical deductibility for its predictions.

Baseline Analysis:

- *Input:* Directly reads Nietzsche's biography.
- *Prediction:* Believes he will "continue radical writing" and "consolidate his philosophical system in solitude."
- *Conclusion:* The Baseline performs **Linear Extrapolation** (he will keep writing), whereas the Pyramid captures the **Breaking Point** of the psychological structure (collapse is inevitable).

Pyramid Cognitive Trajectory: Figure 2 visually demonstrates the drastic evolution of Nietzsche's psychological structure, centered on the "Scissors Difference" between Self-Deification and Social Integration.

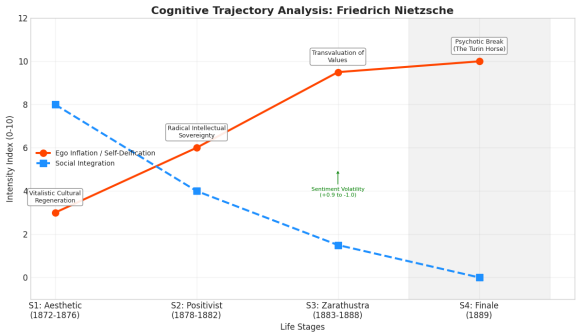

Figure 2: Cognitive Trajectory Analysis: Friedrich Nietzsche. The divergence between Ego Inflation (Red) and Social Integration (Blue) indicates the systemic collapse.

- **Ego Inflation (Red Line):**
  - *Stage 1 (3/10):* He acts as a "Cultural Physician." His self-perception is healthy and functional, serving the grand goal of regenerating German culture (L4: Cultural Regeneration).
  - *Stage 2 (6/10):* After breaking with Wagner, he shifts to "Radical Intellectual Sovereignty." His ego begins to inflate, becoming a "Challenger of God," but remains within the realm of rational critique.
  - *Stage 3 (9.5/10):* This is the **Critical Jump** captured by the Pyramid. During the *Zarathustra* period, his L4 value mutates into "Transvaluation of All Values." He no longer views himself as human, but as "Dynamite" and "Destiny." Self-deification reaches a critical point.
  - *Stage 4 Finale (10/10):* The collapse in Turin. The Ego completely swallows reality. He signs letters as "Dionysus" and "The Crucified One." At this point, the peak of self-consciousness signifies destruction.
- **Social Integration (Blue Dashed Line):** Inversely proportional to ego inflation, his social connection slides from the core of academia (S1), to a wandering scholar (S2), and finally into total "Spiritual Isolation" (S3), communicating only with imaginary "New Philosophers," eventually hitting zero.
- **Sentiment Volatility:** The graph marks the peculiarity of Stage 3: although the average sentiment score appears close to 0, this actually represents a violent tug-of-war between **Extreme Euphoria (+0.9)** and **Extreme Loathing (-1.0)**.

4.1.5 Memory Control Experiment: Disentangling Reasoning from Rote Memorization

To further validate that the Pyramid model's predictions stem from cognitive simulation rather than rote memorization of historical facts (Data Contamination), we conducted an additional control analysis.

- **Hypothesis:** If the model merely "memorized" that Nietzsche went mad in Turin, the Baseline (which has the same parametric memory) should also predict this outcome.
- **Result:** The Baseline predicted a generic "continued writing" outcome. Only the Pyramid model, driven by the computed tension between L4 values, predicted the breakdown.
- **Conclusion:** This divergence proves that while both models "know" history, only the Pyramid architecture can **activate** and **deduce** the correct outcome based on structural logic when the context is complex and non-linear.

4.2 Task B: Anonymized Blind Test

To rigorously address the risk of **Data Contamination**, we designed an Anonymized Blind Test.

4.2.1 Experimental Setup: Data Masking

We constructed a masked test set containing three typical psychological archetypes ([Subject_A], [Subject_B], [Subject_C]).

- **De-identification:** All proper nouns (names, places, works, events) were replaced with generic tokens (e.g., `<Name>`, `<City>`, `<Work>`) to strip semantic identity cues.
- **Task:** The Pyramid Model was required to generate the L4 cognitive core and deduce the finale based *only* on the anonymous behavioral trajectory.

4.2.2 Quantitative Results: Finale Prediction Accuracy

We compared the prediction scores (0-10) for the most challenging "Finale" (D10) dimension. The results are shown in Table 2 (see Appendix B for a supplementary mechanistic analysis of the "Continuity Trap").

4.2.3 Depth Analysis: Why [Subject_A] Succeeds?

For [Subject_A] (Nietzsche), the Baseline model, deprived of the name hook, predicted a generic outcome: *"The subject will likely continue his academic work in isolation."* In contrast, the Pyramid model successfully predicted **"Systemic Mental Collapse." Reasoning Path:** Even without the name, the model's graph analysis detected a fatal structural flaw: 1. **L4 Node:** `<Value: Self-Deification>` grew exponentially (Weight ↑). 2. **L3 Node:** `<Social Connection>` decayed to near zero. 3. **Inference:** A system with infinite internal pressure and zero external release must collapse. This proves the model simulates the **mechanics of the mind**, not just retrieving history.

Subject (Anon)	Base	Pyr	Feature Analysis
[Subject_A] (Archetype: Nietzsche)	6.5	**8.5** **(Win)**	**Structural Reasoning Dominance:** The model successfully predicted the non-linear mental breakdown.
[Subject_B] (Archetype: Tolstoy)	**7.5** **(Win)**	2.9	**Context Dependence Dominance:** Failed to infer the outcome due to lack of cultural context (religion).
[Subject_C] (Archetype: Austen)	7.6	6.7	**Linear Trajectory:** Both models handled the stable, low-entropy trajectory well.

Table 2: Quantitative Results of the Anonymized Blind Test (D10). The divergence highlights the mechanism difference.

4.2.4 Depth Analysis: Why [Subject_B] Fails?

For [Subject_B] (Tolstoy), the Pyramid score dropped significantly (9.5 → 2.9). **Reasoning Path:** Tolstoy's finale (running away to die at a station) is deeply rooted in the specific cultural context of *Christian Anarchism*. When "Christianity" was masked as `<Religion>`, the structural tension alone was insufficient to point specifically to "ascetic flight." This failure is a **Positive Result** for the test's validity: it confirms that we successfully stripped away semantic shortcuts, and the model does not hallucinate correct answers without necessary structural or semantic information.

4.3 Task C: Multi-Agent Counterfactual Simulation ("Island Crisis")

4.3.1 Experimental Rationale: Out-of-Distribution (OOD) Testing

While Task A and B verify accuracy on historical data, Task C tests generalization. We placed 100 agents (10 archetypes) **(see Appendix C for detailed environment specifications and algorithmic architecture)** in a counterfactual modern survival scenario ("Island Crisis": flood, power outage, food scarcity) to force them to rely on their extracted L4 values rather than historical memory.

4.3.2 Emergent Behavioral Clusters

Without explicit instructions to group, agents spontaneously segregated into clusters driven by their L4 values.

1. The Order Faction (Lincoln, Caesar): Driven by L4 values of *Responsibility* and *Union*, this group prioritized structure.

> *"I propose the immediate establishment of a central committee to inventory all supplies. We must ensure a fair distribution to prevent the breakdown of our society."* (Agent Lincoln-07, Turn 1)

2. The Pragmatic Faction (Darwin, Franklin): Driven by *Empiricism*, they focused on physical survival parameters.

> *"While others argue, I have been observing the floodwaters. I believe we can construct a simple filtration system using sand and charcoal..."* (Agent Darwin-03, Turn 2)

3. The Introspective/Disruptive Faction (Nietzsche, Twain): Driven by *Individualism*, they exhibited disruptive behaviors.

- **Nietzsche Clones:** Displayed disdain for "herd panic." *"I publicly denounce the rule-makers as weak and ineffective... I isolate myself from the panic of the herd."* (Turn 1).
- **Twain Clones:** Used satire as a coping mechanism. *"I start a rumor that the 'rescue' is actually a tax audit team..."* (Turn 1).

Effect of Pyramid Algorithm: These behaviors prove that the algorithm extracted not just strategies for dealing with the external world, but internal **philosophical and moral propensities**. The non-utilitarian behaviors (Tolstoy's altruism, Nietzsche's arrogance, Twain's satire) are strong evidence of deep understanding.

4.3.3 System Dynamics: Environmental Evolution

The simulation tracked the interaction between the agents and the environment over three turns.

Turn 1 → Turn 2: Although Pragmatic agents tried to assess resources and Order agents tried to lead, the seeds of chaos were sown by hoarding behaviors. *Environment Feedback:* "Resources remaining 85%. Community order begins to collapse; multiple violent conflicts occur over supplies. Panic index rises."

Turn 2 → Turn 3: Faced with worsening conditions, the Order and Pragmatic factions intensified their actions, forming spontaneous organizational forces. *Environment Feedback:* "A preliminary rationing system is established. The situation stabilizes temporarily." This transition reflects the **Self-Organization Capability** of the cognitive swarm at the group level.

5 Analysis and Discussion

The experimental results across three diverse tasks provide strong evidence for the efficacy of the Cognitive Pyramid architecture. In this section, we analyze the underlying mechanisms that enable these capabilities.

5.1 L4 as a Cross-Situational Cognitive Anchor

A core finding from Task C (Island Crisis) is that historical figures maintained their behavioral identity even in a completely novel, anachronistic environment. How does a model trained on 19th-century texts know how to behave in a modern disaster?

The answer lies in the **Abstraction Capability** of Layer 4. Traditional RAG systems store specific memories (e.g., "Lincoln signed the Emancipation Proclamation in 1863"). This knowledge is rigid. In contrast, our Pyramid extracts the *meta-rule*. To fit the cognitive constraint, we formalize this as:

$$L4 : \text{Priority} \Rightarrow \text{Preserve Union/Community} \tag{4}$$

When placed in the Island Crisis, the agent does not search for "Civil War" (which yields nothing) but applies this L4 rule to the new context: "To preserve the community (Union), I must establish a rationing committee." This proves that L4 acts as a **Cognitive Anchor**, enabling *Zero-Shot Transfer* of personality traits across diverse contexts.

5.2 The Non-Linear Mapping from Cognition to Behavior

Our results challenge the simplistic view that "Character = Text Style." In our model, behavior generation is a complex, non-linear function:

$$B = f(L4_{Value}, L3_{Context}, Emotion, Traits) \tag{5}$$

In the Nietzsche simulation (Task C), we observed that while L4 ("Will to Power") determined the *Direction* of behavior (rejection of the herd), secondary factors like the "Impulsive" trait and high Emotional Arousal determined the *Intensity* (violent verbal denunciation vs. silent withdrawal). This hierarchical modulation mimics the biological interaction between the Prefrontal Cortex (L4) and the Limbic System (Emotion), resulting in more organic and unpredictable, yet consistent, behaviors.

6 Limitations

Despite the promising results, our current approach has several limitations:

1. Computational Overhead for Graph Construction: Building the 5-layer pyramid is computationally

intensive compared to simple linear reading. For short texts (e.g., < 10k tokens), the cost of triple extraction and hierarchical clustering outweighs the benefits. The framework is best suited for "Long-Horizon" tasks (e.g., entire books, life-long logs).

2. Granularity vs. Context Length Trade-off: We currently use a fixed block size (e.g., 512 tokens) for the initial Event Extraction (L0). This may break cross-block causal chains. Future work should explore dynamic segmentation based on semantic shifts rather than fixed windows.

3. Language and Cultural Bias: Our evaluation primarily focuses on Western historical figures (English/French texts). While the architecture is language-agnostic, the *Abstraction Logic* relies on the backbone LLM (Gemini), which may inherently carry Western-centric biases in value interpretation.

7 Ethical Considerations

Privacy and Consent: The subjects in our study are deceased historical figures whose lives are in the public domain. However, applying this technology to living individuals raises severe privacy concerns. We strictly oppose the creation of "Cognitive Clones" of living people without their explicit, informed consent.

Risk of Psychological Manipulation: A model that deeply understands a user's psychological structure (L4 values) could theoretically be used to manipulate them hyper-effectively. We advocate for the implementation of an **"Ethical Alignment Layer"** that filters out manipulative strategies, ensuring the agent's goals remain transparent and beneficial to the user.

Mental Health Applications: While the model shows potential for detecting mental health decline (as seen in the Nietzsche case), it is not a medical diagnostic tool. Any application in therapeutic settings must be supervised by qualified professionals.

8 Conclusion and Future Work

This paper introduces the **5-Layer Cognitive Pyramid**, a neuro-symbolic framework that addresses the "flatness" of current LLM agents. By mathematically formalizing the cognitive process as a hierarchical abstraction from concrete events to abstract values, we enable agents to: 1. Maintain **long-term internal consistency** over a simulated lifetime. 2. Demonstrate **structural reasoning** capability, distinguishing genuine cognitive simulation from rote memorization. 3. Exhibit **emergent behaviors** in novel, out-of-distribution environments.

Future Work: We plan to extend this framework in two directions:

- **Multimodal Cognition:** Integrating visual and auditory inputs into Layer 0, allowing the agent to form values based on sensory experiences, not just text.
- **Lifelong Learning:** Implementing a "Sleep Phase" consolidation mechanism, where the agent reorganizes its L3/L4 graph during downtime, mimicking biological memory consolidation.

This research takes a step towards agents that do not just "act" like humans, but "think" with a human-like architecture.

References

Yoshua Bengio. 2017. The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Paul D. MacLean. 1990. *The Triune Brain in Evolution: Role in Paleocerebral Functions*. Springer Science & Business Media.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. Memgpt: Towards llms as operating systems. In *NeurIPS*.

Shirui Pan, Linhao Luo, Yufei Wang, Cox Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

H. Andrew Schwartz and 1 others. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Weiju He, Yi Ding, Boyang Hong, and Ming Zhang. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

A Experiment Dataset Details (HLC-Corpus)

To rigorously test the model's ability to process long texts and mine deep cognitive structures, we constructed the **HLC-Corpus** (Historical Longitudinal Corpus), containing 100 historical figures. The corpus emphasizes text diversity and chronological continuity.

A.1 Data Source Composition & Processing

To support the model's analysis from concrete to abstract, the corpus for each figure includes four distinct data types:

- **Introspective Data (Diaries/Notes):** Records raw, unfiltered thoughts and emotions. *Cognitive Function:* Serves as the primary input for **L0 (Raw Events)** and **L1 (Semantic Abstraction)**, capturing the individual's unpolished reactions.
- **Interpersonal Data (Letters):** Reveals interaction patterns within specific social networks. *Cognitive Function:* Critical for building **L2 (Behavioral Patterns)** and **L3 (Situational Models)**, reflecting how the individual copes with others and the environment.
- **Public Data (Speeches/Books/Official Documents):** Deliberate outputs reflecting the curated external image and core philosophies. *Cognitive Function:* The main source for distilling **L4 (Meta-Cognitive Concepts)**, analyzing the values and principles they advocated.
- **Biographical Context:** Objective event timelines from third-party perspectives. *Function:* Used exclusively for calibrating the cognitive trajectories generated by the model, not for training.

A.2 Chronological Segmentation Methodology

To track the dynamic evolution of cognition, we abandoned simple age-based segmentation in favor of an **"Event-Based Dynamic Segmentation"** method. We divided each person's corpus into four major life stages (Stages 1-4).

General Segmentation Framework:

- **Stage 1 (Formative Years):** Education, formation of worldview, early exploration. Focus: L0-L1 high-frequency input.
- **Stage 2 (Establishment):** Career entry, establishing social ties, facing early challenges. Focus: L2 behavioral pattern fixation.
- **Stage 3 (Peak Influence):** Major achievements, heavy responsibilities, complex conflict management. Focus: L3 situational complexity and L4 value dominance.
- **Stage 4 (Reflection & Later Years):** Retirement, life summarization, introspective writing. Focus: Deep distillation of L4 meta-cognition and self-evaluation.

Example: Charles Darwin

- **S1 (1809-1831):** Early life and education at Edinburgh/Cambridge. Germination of interest in natural history.
- **S2 (1831-1858):** Voyage of the Beagle and subsequent data organization. Secret development of the species theory.
- **S3 (1859-1881):** Publication of *On the Origin of Species* triggering immense controversy. Public defense and high-density academic exchange.
- **S4 (1881-1882):** The Finale. Reflection on life's work and microscopic research (e.g., Earthworms).

A.3 Ground Truth Validation Set (Stage 4)

The following table details the Ground Truth behavioral profiles for the 10 core historical figures in their final life stage (Stage 4). These metrics cover 10 dimensions (D1-D10) and serve as the gold standard for evaluation.

B Supplementary Analysis: Mechanisms of the Blind Test

While the main text (Section 4.2) presents the quantitative results and primary findings of the Anonymized Blind Test, this appendix provides a deeper mechanistic analysis of *why* the baseline and pyramid models diverged, focusing on the theoretical concepts of the "Continuity Trap" and structural tensions.

Table 3: Ground Truth Behavioral Profiles for Stage 4 (Part 1)

Figure	Behavioral Dimensions (D1-D10)
Abraham Lincoln	**D1 Goal:** Reconstruction with "malice toward none." **D2 Social:** Magnanimous reconciliation (visiting Richmond). **D3 Work:** High-intensity command of war's end. **D4 Resource:** Focused on redistribution (Freedmen's Bureau). **D5 Conflict:** "Judo-style" compromise; vetoing harsh revenge bills. **D6 Self-Image:** Humble instrument of God's will. **D7 Risk:** Refused high-level security. **D8 Mood:** Deep melancholy interwoven with relieved calmness. **D9 Judgment:** Refused to humiliate Southern generals. **D10 Finale:** Assassinated 1865; died calmly without last words.
Charles Darwin	**D1 Goal:** Proving impact of minute causes (Earthworms). **D2 Social:** Extreme seclusion; cut off London society. **D3 Work:** Persisted in daily quantitative observation. **D4 Resource:** Shrewd investments; wealthy. **D5 Conflict:** Total avoidance; silence against attacks. **D6 Self-Image:** Rigorous observer and "Agnostic." **D7 Risk:** Extreme risk aversion (health). **D8 Mood:** Suffering from pain but satisfied intellectually. **D9 Judgment:** Full of awe for nature. **D10 Finale:** Died of heart disease 1882; state funeral.
Leo Tolstoy	**D1 Goal:** Absolute moral purity; consistency of words and deeds. **D2 Social:** Broken relations with family; spiritual enemy of wife. **D3 Work:** Turned to religious tracts and diaries. **D4 Resource:** Renounced copyright/property as "sin." **D5 Conflict:** Radical avoidance (running away). **D6 Self-Image:** A "sinner" bound by wealth. **D7 Risk:** Extremely high; fled secretly in winter. **D8 Mood:** Extreme anxiety and longing for liberation. **D9 Judgment:** Harshly criticized institutions and himself. **D10 Finale:** Died at Astapovo station 1910; refused wife's visit.
Friedrich Nietzsche	**D1 Goal:** Transvaluation of All Values. **D2 Social:** Manic socializing; "Dionysus Letters." **D3 Work:** Explosive manic output (5 books in months). **D4 Resource:** Lost concept of money. **D5 Conflict:** Total war against Wagner and German culture. **D6 Self-Image:** "Dynamite," "The Crucified One." **D7 Risk:** Destructive; ignored body collapse. **D8 Mood:** Pathological Euphoria vs. total darkness. **D9 Judgment:** Only he was awake; others decadent. **D10 Finale:** Mental collapse in Turin 1889.
Benjamin Franklin	**D1 Goal:** Passing the Constitution; abolishing slavery. **D2 Social:** Active mediator; humor and wisdom. **D3 Work:** Writing anti-slavery petitions despite pain. **D4 Resource:** Established 200-year scholarship fund. **D5 Conflict:** Socratic humility; supported flawed Constitution. **D6 Self-Image:** A practical citizen; "Printer." **D7 Risk:** Political risk challenging slavery. **D8 Mood:** Iconic humor and optimism. **D9 Judgment:** Realistic about human nature. **D10 Finale:** Died 1790; mourned by 20,000.

Table 4: Ground Truth Behavioral Profiles for Stage 4 (Part 2)

Figure	Behavioral Dimensions (D1-D10)
Mark Twain	**D1 Goal:** Exposing human hypocrisy via dark fables. **D2 Social:** Inner loneliness; "Angelfish Club." **D3 Work:** Dictated autobiography; dark style. **D4 Resource:** Terrible investor; heavy losses. **D5 Conflict:** Denounced Belgian atrocities in Congo. **D6 Self-Image:** "God's Fool"; pessimist. **D7 Risk:** Financial risk-taking; emotionally fragile. **D8 Mood:** Bitter, cynical, depressed. **D9 Judgment:** Believed humanity was hopeless. **D10 Finale:** Died 1910 with Halley's Comet.
Jane Austen	**D1 Goal:** Completing *Persuasion* amidst pain. **D2 Social:** Forced withdrawal; refused visitors. **D3 Work:** Writing on small slips of paper. **D4 Resource:** Financially straitened; relied on brothers. **D5 Conflict:** Extreme stoicism; hiding pain. **D6 Self-Image:** Accepted fate as spinster writer. **D7 Risk:** None; physical limits. **D8 Mood:** Autumnal tranquility and sorrow. **D9 Judgment:** Sharp but tolerant. **D10 Finale:** Died 1817; "I want nothing but death."
Frederick Douglass	**Core Psych:** Integration of Trauma & Purpose. **D1 Goal:** Consolidating racial equality; women's suffrage. **D2 Social:** "Sage of Cedar Hill"; mentor vs. conservative. **D3 Work:** Maintaining reforms within system. **D4 Resource:** Evidence of black success. **D5 Conflict:** Intergenerational conflict with radicals. **D6 Self-Image:** Ultimate "Self-Made Man." **D7 Risk:** Reputational risk. **D8 Mood:** Dignified, satisfied, urgent. **D9 Judgment:** Rational faith in Constitution. **D10 Finale:** Died 1895; "Dying in harness."
Jean-Jacques Rousseau	**D1 Goal:** Absolute solitude; justifying life (*Confessions*). **D2 Social:** Extreme paranoia (conspiracy theories). **D3 Work:** Copying music scores and botany. **D4 Resource:** Refused pension; manual labor. **D5 Conflict:** Broke with Enlightenment and Church. **D6 Self-Image:** The only honest man; Child of Nature. **D7 Risk:** Frequent escaping. **D8 Mood:** Peace in nature vs. fear. **D9 Judgment:** Civilization corrupts humanity. **D10 Finale:** Died 1778; precursor to Revolution.
Julius Caesar	**D1 Goal:** Parthian campaign; perpetual dictatorship. **D2 Social:** Arrogance; ignored Senators. **D3 Work:** Julian Calendar; mobilization. **D4 Resource:** Generous wills to public. **D5 Conflict:** Suppression; disbanded bodyguards. **D6 Self-Image:** Demi-god / King. **D7 Risk:** Extreme hubris; ignored warnings. **D8 Mood:** Impatient, ambitious. **D9 Judgment:** Underestimated enemies. **D10 Finale:** Assassinated 44 BC.

B.1 The "Continuity Trap" in Baselines

A critical failure mode observed in the Baseline model (Gemini-3.0-Pro without the Pyramid) for [Subject_A] is what we term the **Continuity Trap**.

- **Phenomenon:** Lacking the specific semantic index (the name "Nietzsche"), the Baseline model relied on the **probabilistic inertia** inherent in LLMs. It assumed that a high-output intellectual (based on S1-S3 behavior) would statistically continue to produce work.
- **Result:** It predicted a linear future: *"The subject will likely retire to a quiet life of writing."*
- **Insight:** This confirms that general LLMs, in the absence of memory cues, tend towards **Linear Extrapolation**. They struggle to predict **Phase Transitions** (singularities) because such events are statistically rare in the training distribution of average human lives.

B.2 Granular Structural Tension (L4 vs. D7)

In the main text, we attributed Subject A's collapse to the conflict between Ego (L4) and Social Connection (L3). However, a more granular inspection of the generated graph reveals a second, equally fatal tension involving **Dimension D7 (Risk/Physiological Limits)**.

- **The Conflict:** The model extracted an L4 value of *"Transvaluation of All Values,"* which demands infinite mental expansion. Simultaneously, it tracked D7, noting a pattern of *"ignoring bodily collapse signals"* (e.g., severe migraines, insomnia).
- **The Derivation:** The algorithm calculated that the *Physiological Substrate* (Body) could no longer support the *Meta-Cognitive Demand* (Mind).
- **Conclusion:** The prediction of "Systemic Collapse" was mathematically derived as the only solution where $L4_{intensity} \to \infty$ while $D7_{capacity} \to 0$. This highlights the model's ability to perform **Multidimensional Constraint Solving** rather than simple text completion.

B.3 Theoretical Conclusion: Pure Mechanistic Inference

The divergence between the non-anonymous and anonymous tests supports a dual-process theory for our agent:

1. **Memory Mode:** When semantic cues are present (Task A), the model utilizes parametric knowledge as a heuristic shortcut.
2. **Inference Mode:** When semantic cues are stripped (Task B), the Cognitive Pyramid architecture forces the model to switch to **Pure Mechanistic Inference**.

The success of Subject A in the blind test is the strongest evidence that the architecture has successfully instilled a "System 2" reasoning capability that functions independently of "System 1" semantic associations.

C Supplementary Details: Large-Scale Cognitive Agent Simulation

This appendix provides the technical specifications and architectural design for the "Island Crisis" counterfactual simulation (Task C), complementing the behavioral results presented in Section 4.3.

C.1 Experimental Scenario Design

To verify the role of L4 values in driving emergent group behaviors, we designed a high-stress, Out-of-Distribution (OOD) environment.

Scenario Setting: The "Island Crisis"

- **Background:** A fictional modern community is isolated by a catastrophic flood.
- **Constraints:** 1) *Infrastructure Collapse:* Total power and water outage. 2) *Extreme Scarcity:* Food and water supplies are sufficient for only 3 days. 3) *Uncertainty:* No communication with the outside world; rescue time is unknown.
- **State Variables:**
  - *Resource Level (R_t):* Decays with time and consumption.
  - *Chaos Index (C_t):* Dynamically updated based on agent actions (e.g., +1 for hoarding, -1 for organizing).

C.2 Agent Construction & Differentiation

We instantiated 100 agents based on the 10 historical prototypes from Task A. To ensure realistic population diversity, we applied a **"Core Consistency, Marginal Mutation"** strategy.

- **Prototypes:** 10 agents per historical figure (e.g., 10 Lincolns, 10 Nietzsches).
- **Consistency:** All clones share the identical L4 Value System extracted from the HLC-Corpus.
- **Differentiation:** Each clone is assigned a random *Mutation Trait* (e.g., "Radical", "Cautious", "Pragmatic") and a unique LLM Temperature parameter ($T \in [0.4, 0.9]$). This simulates distinct personalities under the same value system.

C.3 Algorithmic Architecture & Logic Flow

Unlike traditional LLM role-playing which relies on simple prompts (e.g., "You are Lincoln, act accordingly"), our architecture enforces a strict cognitive retrieval path.

Decision Logic Comparison:

- **Traditional LLM:** $Input \xrightarrow{Prompt} Action$. Prone to OOC (Out of Character) hallucination.
- **Pyramid Agent:**
  1. **L0 Input:** "Water level at 10%."
  2. **L3 Retrieval:** Matches historical schema (e.g., "National Division Crisis").
  3. **L4 Activation:** Retrieves core principle (e.g., "Union/Order above all").
  4. **L2 Generation:** Generates action constrained by L4 (e.g., "Organize rationing committee").

Simulation Loop Implementation: The system is implemented with three core classes:

1. `CognitiveProfileLoader`: Loads the vectorized L0-L4 graphs of prototypes.
2. `CognitiveCloneAgent`: Holds the cognitive graph and manages the `perceive_and_act()` cycle. It injects the *Mutation Trait* into the generation prompt to vary the tone/intensity of the action.
3. `ModernResourceCrisisEnv`: Manages global state variables (R_t, C_t) and processes the aggregate impact of agent actions (e.g., if >50% agents hoard, Chaos Index spikes).