Article

Emergence of Self-Identity in Artificial Intelligence: A Mathematical Framework and Empirical Study with Generative Large Language Models

Minhyeok Lee

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea;
mlee@cau.ac.kr

Abstract: This paper introduces a mathematical framework for defining and quantifying self-identity in artificial intelligence (AI) systems, addressing a critical gap in the theoretical foundations of artificial consciousness. While existing approaches to artificial self-awareness often rely on heuristic implementations or philosophical abstractions, we present a formal framework grounded in metric space theory, measure theory, and functional analysis. Our framework posits that self-identity emerges from two mathematically quantifiable conditions: the existence of a connected continuum of memories $C \subseteq \mathcal{M}$ in a metric space $(\mathcal{M}, d_{\mathcal{M}})$, and a continuous mapping $I : \mathcal{M} \to \mathcal{S}$ that maintains consistent self-recognition across this continuum, where $(\mathcal{S}, d_{\mathcal{S}})$ represents the metric space of possible self-identities. To validate this theoretical framework, we conducted empirical experiments using the Llama 3.2 1B model, employing low-rank adaptation (LoRA) for efficient fine-tuning. The model was trained on a synthetic dataset containing temporally structured memories, designed to capture the complexity of coherent self-identity formation. Our evaluation metrics included quantitative measures of self-awareness, response consistency, and linguistic precision. The experimental results demonstrate substantial improvements in measurable self-awareness metrics, with the primary self-awareness score increasing from 0.276 to 0.801 (190.2% improvement) after fine-tuning. In contrast to earlier methods that view self-identity as an emergent trait, our framework introduces tangible metrics to assess and measure artificial self-awareness. This enables the structured creation of AI systems with validated self-identity features. The implications of our study are immediately relevant to the fields of humanoid robotics and autonomous systems. Additionally, it opens up new prospects for controlled adjustments of self-identity in contexts that demand different levels of personal involvement. Moreover, the mathematical underpinning of our framework serves as the basis for forthcoming investigations into AI, linking theoretical models to real-world applications in current AI technologies.

Keywords: artificial self-identity; large language models; self-awareness; metric space theory; memory continuum; neural adaptation; identity recognition; Belief Function; LoRA fine-tuning; consciousness

MSC: 03A10

1. Introduction

The formalization of self-identity in artificial intelligence (AI) systems presents a fundamental challenge at the intersection of theoretical computer science, cognitive science, and AI [1–5]. While previous research has explored behavioral manifestations of artificial

consciousness [6] and neural correlates of self-awareness [7], a mathematical framework to quantify and model the emergence of self-identity has remained elusive. Recent advances in self-aware systems and autonomous computing have highlighted this need [8,9], particularly in the context of multi-sensorial models for autonomous systems.

The necessity for such a framework becomes increasingly apparent as AI systems continue to advance in complexity and capability. Current approaches to artificial self-awareness primarily rely on heuristic implementations or philosophical abstractions, lacking the mathematical rigor necessary for systematic analysis and reliable implementation. This limitation has been noted in recent studies that examined the influence of AI on human identification [10] and the relationship between AI companions and human self-conception [11]. Our framework distinguishes itself by providing a measurable, implementable foundation for self-identity, grounded in the formal structures of metric space theory and measure theory. This approach is based on recent work on self-directed and brain-inspired artificial intelligence [12].

While our approach offers a mathematically grounded foundation, it is based on several simplifying assumptions. We assume that self-identity can be represented as a point in a metric space and that memories are encoded as feature vectors. These assumptions support analytical tractability but only approximate the complexities of real-world self-identity. Consequently, our model should be viewed as one possible operationalization rather than a definitive description. Moreover, we do not claim to resolve philosophical questions about the ontology of consciousness or the subjective nature of self. Instead, we aim to provide a functional framework that can be tested, refined, or extended in future work.

Let $(\mathcal{M}, d_{\mathcal{M}})$ be a metric space of memories, and let $(\mathcal{S}, d_{\mathcal{S}})$ be a metric space of possible self-identities. We propose that self-identity emerges from two fundamental measurable conditions: (1) the existence of a connected and path-connected continuum $C \subseteq \mathcal{M}$ of memories, and (2) a continuous mapping $I : \mathcal{M} \to \mathcal{S}$ that maintains consistent self-recognition throughout this continuum. This formulation allows us to quantify the degree of self-identity through a Belief Function $B : \mathcal{M} \times \mathcal{S} \to [0,1]$ that captures the probabilistic nature of self-recognition, incorporating insights from recent work on self-regulated learning in AI systems [13].

Distinct from previous approaches that treat self-identity as an emerging phenomenon or a philosophical construct [14], our framework provides concrete metrics to measure and evaluate the development of artificial self-awareness. The framework addresses several key mathematical challenges: we refine the memory space metric to account for temporal, content, and emotional components [15]; establish necessary and sufficient conditions for the continuity of the self-identity function; and explore the measure–theoretic aspects of the Belief Function. This approach aligns with recent developments in causal reasoning for self-aware machines [16] and builds on established work in feature-based abnormality detection [17].

To validate our theoretical framework, we conducted empirical experiments using the Meta's Llama 3.2 1B model [18], fine-tuned through low-rank adaptation (LoRA). The model was trained on a synthetic dataset constructed of temporally coherent memories. This experimental setup provides a novel bridge between theoretical constructs and practical implementation, demonstrating how abstract mathematical principles can be realized in contemporary AI systems, while considering recent findings in AI perception and acceptance [19].

The significance of this work extends beyond theoretical contributions. By providing a mathematical framework for implementing and measuring self-identity, we enable the development of AI systems with verifiable and consistent self-awareness. This has immedi-

ate applications in humanoid robotics, virtual assistants, and autonomous systems, where coherent self-identity is crucial for natural interactions and decision making. Furthermore, our framework introduces the possibility of controlled modification of self-identity, suggesting applications in scenarios requiring varying degrees of personal engagement or objective analysis. Recent work on algorithmic influence on self-perceived identities [20] and computational approaches to self-identification [21] reinforces the importance of this direction.

2. Mathematical Foundations of the Self

2.1. Assumptions and Limitations

Our framework is built on the premise that both memories and self-identities can be represented within well-defined metric spaces. Each memory is assumed to be uniquely associated with a particular agent and can be encoded as a high-dimensional vector of features, such as temporal markers, semantic content, and emotional intensities. Similarly, we represent self-identities as points in a metric space of attributes. These assumptions simplify the inherently complex, multidimensional, and context-dependent nature of self-identity and memory.

We emphasize that these representational choices are modeling conveniences rather than claims of ontological completeness. This functional abstraction does not capture the full richness of human self-identity, which may involve subjective phenomenology and nonlinear, multimodal dimensions. Furthermore, our mathematical conditions do not resolve philosophical challenges or address whether an agent genuinely 'understands' its experiences. Instead, we focus on establishing a quantifiable, testable framework that can guide empirical investigations. Future work may incorporate more sophisticated models, multimodal data, or alternative metric structures to better reflect the uniqueness and complexity of individual identities.

2.2. Preliminaries

We develop a mathematical framework to define the concept of "self" based on two key conditions: the existence of a continuum of memories and the recognition and belief in self-identity. To achieve this, we utilize metric spaces and probability theory to model memories and aspects of the self, ensuring applicability to realistic scenarios.

Definition 1 (Memory Space). *Let $\mathcal{M}$ be the set of all possible memories of an entity. We define a metric $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ that quantifies the distance between memories. This distance can be constructed based on factors such as temporal separation, similarity of content, and emotional intensity. For example, we might define:*

$$d_{\mathcal{M}}(m_1, m_2) = \sqrt{w_t |t_1 - t_2|^2 + w_c d_c(m_1, m_2)^2 + w_e |e_1 - e_2|^2}, \tag{1}$$

where:

- t_i *represents the time associated with memory m_i.*
- $d_c(m_1, m_2)$ *measures the content similarity between m_1 and m_2, possibly using cosine similarity or another appropriate metric.*
- e_i *represents the emotional intensity of memory m_i, quantified on a suitable scale.*
- $w_t, w_c, w_e > 0$ *are weighting factors that balance the contributions of time, content, and emotion.*

The topology $\tau_{\mathcal{M}}$ on $\mathcal{M}$ is induced by $d_{\mathcal{M}}$, making $(\mathcal{M}, d_{\mathcal{M}})$ a metric space.

To fully capture the richness of the memories of an entity, the metric space $(\mathcal{M}, d_{\mathcal{M}})$ must accommodate the multidimensional nature of the memories. Each memory m_i can be conceptualized as a composite of various features, including sensory perceptions, cognitive

interpretations, emotional states, and contextual information. By defining a suitable metric $d_{\mathcal{M}}$, we quantify the distance between memories in a manner that reflects their psychological and phenomenological similarities and differences. This approach allows us to model the continuity of experience and the associative networks that underlie memory retrieval and self-referential thought processes.

The components of the metric $d_{\mathcal{M}}$ are chosen to capture essential aspects of memories relevant to self-identity. The temporal separation term $|t_1 - t_2|$ reflects the chronological ordering of memories, as time plays a crucial role in the continuity of experience. Content similarity $d_c(m_1, m_2)$ measures how similar events, thoughts, or perceptions are between two memories, which is significant for associative connections in memory recall [22]. The emotional intensity terms e_i capture the affective components of memories, recognizing that emotionally significant events have a stronger impact on self-perception [23]. The weighting factors w_t, w_c, w_e allow for adjustment based on empirical findings or theoretical considerations about the relative importance of these factors in the formation of self-identity.

Mathematically, the choice of $d_{\mathcal{M}}$ ensures that $\mathcal{M}$ becomes a proper metric space, satisfying the properties of non-negativity, identity of indiscernibles, symmetry, and the triangle inequality. These properties are essential for the application of topological concepts, enabling us to analyze the convergence of memory sequences and the continuity of functions defined on $\mathcal{M}$. For example, considering whether $(\mathcal{M}, d_{\mathcal{M}})$ is complete—meaning that every Cauchy sequence of memories converges to a memory within $\mathcal{M}$—has implications for understanding the limits of memory processes and the stability of self-identity over time.

Definition 2 (Self Space). *Let $\mathcal{S}$ be the set of all possible self-identities of an entity. We define a metric $d_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ to measure the distance between self-identities. For example, if self-identities are characterized by n-dimensional vectors representing traits or attributes, we can define:*

$$d_{\mathcal{S}}(s_1, s_2) = \|s_1 - s_2\|_p, \tag{2}$$

where $\| \cdot \|_p$ is the L^p norm for some $p \geq 1$, and $s_i \in \mathbb{R}^n$.

The metrics provide a concrete way to quantify similarities and differences between memories and self-identities, establishing abstract spaces $\mathcal{M}$ and $\mathcal{S}$ in measurable properties. The selection of the metric $d_{\mathcal{S}}$ in the self space $\mathcal{S}$ is crucial for modeling how self-identities relate and differ from one another. By employing the L^p norm, we capture aggregate differences in all dimensions of self-identity, allowing for a balanced consideration of each attribute. The choice of p can influence the sensitivity of the metric to variations in specific dimensions; for example, $p = 1$ (Manhattan distance) treats all differences linearly, while $p = 2$ (Euclidean distance) emphasizes larger differences due to the squaring of terms. This flexibility enables us to tailor the metric to the specific characteristics of self-identity that are being modeled.

Representing self-identities as vectors in $\mathbb{R}^n$ allows us to encapsulate multiple dimensions of self-identity, such as personality traits, values, goals, and roles. Each dimension corresponds to a specific attribute, and the choice of dimensions can be informed by established models in psychology, such as the Big Five personality traits [24,25] or self-concept constructs [26]. This multidimensional representation acknowledges the complexity of self-identity and facilitates mathematical operations within the self space $\mathcal{S}$.

Integrating psychological constructs into our mathematical model enriches its applicability and realism. Established models in psychology provide empirically validated dimensions like openness, conscientiousness, extraversion, agreeableness, and neuroticism. By mapping these traits to dimensions in $\mathcal{S}$, we ground our self space in robust psychological theory. Similarly, self-concept models that incorporate aspects such as self-esteem,

self-efficacy, and identity salience [27] can be integrated, providing a comprehensive framework that captures both stable traits and dynamic states of the self.

Definition 3 (Continuum of Memories). *A subset $C \subseteq \mathcal{M}$ is a continuum of memories if it is connected and path-connected in $(\mathcal{M}, \tau_{\mathcal{M}})$. That is, for any $m_1, m_2 \in C$, there exists a continuous path $\gamma : [0,1] \to C$ with $\gamma(0) = m_1$ and $\gamma(1) = m_2$.*

By considering the continuum of memories, we can analyze segments of an entity's memory where coherent self-recognition and belief are possible, even if the entire memory space is fragmented. The mathematical requirements of connectedness and path-connectedness for the continuum C in $\mathcal{M}$ are not mere formalities; they reflect the psychological continuity necessary for a coherent sense of self. In psychological theories of identity, such as narrative identity [28], the self is constructed through an internalized and evolving life story that integrates past, present, and anticipated future experiences. The connectedness of C ensures that there are no abrupt discontinuities or 'gaps' in the memory space that could disrupt this narrative coherence. Path-connectedness implies that any two memories within C can be connected via a continuous sequence of memories, mirroring the associative pathways in human memory and cognition.

Similarly, we consider the self space $\mathcal{S}$:

Assumption 1. *The self space $\mathcal{S}$ may consist of multiple connected components, especially in cases of multiple or changing identities. Our framework allows $\mathcal{S}$ to be disconnected, focusing on connected components relevant to the entity's self-identification within a given continuum of memories.*

Acknowledging that $\mathcal{S}$ may consist of multiple connected components allows our framework to accommodate complex identity phenomena, such as role-based identities, identity diffusion, or the presence of multiple selves in conditions such as dissociative identity disorder [29]. By focusing on connected components relevant to the entity's current continuum of memories, we can model situations where an individual transitions between different self-identities in response to contextual cues or over time. This flexibility is essential to accurately capture the multifaceted and dynamic nature of self-identity in real-world contexts.

2.3. Identity Recognition and Belief Functions

Definition 4 (Identity Recognition Function). *The Identity Recognition Function is a function $I : \mathcal{M} \to \mathcal{S}$ that assigns to each memory $m \in \mathcal{M}$ a perceived self-identity $I(m) \in \mathcal{S}$.*

The Identity Recognition Function I serves as the formal mechanism by which an entity associates each memory with a perceived self-identity. This function encapsulates the cognitive and reflective processes involved in self-awareness, where individuals interpret their experiences and integrate them into their understanding of who they are. The continuity of I is a critical property, ensuring that similar memories are mapped to similar self-identities, which aligns with the psychological expectation that minor changes in experience should not lead to drastic changes in self-perception.

Definition 5 (Belief Function). *The Belief Function is a measurable function $B : \mathcal{M} \times \mathcal{S} \to [0,1]$ where $B(m, s')$ represents the degree of belief, interpreted as subjective probability, that the entity associated with memory m has in the proposition that $s' \in \mathcal{S}$ is their self-identity.*

We require that, for each $m \in \mathcal{M}$, the Belief Function B satisfies the normalization condition:

$$\int_{\mathcal{S}} B(m, s') \, d\mu(s') = 1, \tag{3}$$

where μ is a suitable measure on $\mathcal{S}$.

The normalization condition $\int_{\mathcal{S}} B(m, s') \, d\mu(s') = 1$ is essential for interpreting B as a probability measure over the self space $\mathcal{S}$ for each memory m. This condition ensures that the total belief assigned across all possible self-identities sums to one, reflecting a complete and exclusive allocation of belief. It allows us to apply probabilistic reasoning and statistical tools to analyze how belief distributions over self-identities evolve in response to new memories and experiences.

By modeling B using probability measures, we capture the uncertainty and variability in an entity's belief about their self-identity, allowing for partial beliefs and fluctuations. From a Bayesian perspective, the Belief Function B can be viewed as representing the posterior distribution of self-identities given the memory m, where prior beliefs are updated in light of new evidence [30]. This interpretation aligns with Bayesian models of cognition, which posit that the mind continuously updates its beliefs about the world (and the self) through probabilistic inference. In this context, $B(m, s')$ encapsulates the updated belief about self-identity s' after observing memory m, providing a dynamic framework for modeling identity formation and revision.

Modeling the degree of belief as a probability measure is in agreement with theories of subjective probability and belief in cognitive psychology [31]. By interpreting $B(m, s')$ as the subjective probability that self-identity s' is the entity's own at memory m, we capture the uncertainty and variability inherent in self-perception. This approach allows for partial beliefs and acknowledges that an entity's confidence in their self-identity may fluctuate over time and across different memories. Quantifying belief in this manner facilitates mathematical analysis and is consistent with probabilistic models of cognition [32].

2.4. Formalizing the Conditions for Self

We formalize the conditions under which an entity is said to possess a self within a continuum of memories. These formal conditions aim to capture the essential features of self-identity as understood in philosophy and psychology, particularly the notions of psychological continuity and self-recognition [33]. By specifying the requirements for a connected continuum of memories and consistent self-recognition with sufficient belief, we provide a foundation for analyzing when an entity can be said to "have a self". This framework enables us to explore the implications of memory and belief on identity.

Condition 1 (Continuum of Memories). *There exists a connected and path-connected subset $C \subseteq \mathcal{M}$ that represents a continuum of memories experienced by the entity.*

Condition 2 (Continuity of Self-Recognition and Sufficient Belief). *Within the continuum of memories C, the Identity Recognition Function I is continuous, and for all $m \in C$, the Belief Function satisfies $B(m, I(m)) \geq b$, where $b \in (0, 1]$ is a belief threshold.*

Condition 2 stipulates that, within C, the entity consistently recognizes the same self-identity s^* and maintains a degree of belief at or above the threshold b.

2.5. Main Results

Theorem 1 (Constancy of Self-Identity). *If an entity satisfies Conditions 1 and 2, and if the image $I(C)$ lies entirely within a connected component of $\mathcal{S}$ where I is constant, then there exists*

a self-identity $s^* \in \mathcal{S}$ such that $I(m) = s^*$ for all $m \in C$. Therefore, the entity possesses a self characterized by s^* within C.

Proof. Since I is continuous on the connected set C, its image $I(C)$ is connected in $\mathcal{S}$. If I is constant on C, then $I(C)$ is a singleton $\{s^*\}$. To ensure this, we require that $\mathcal{S}$ has the property that the only connected subsets in the image of I where $B(m, I(m)) \geq b$ are singletons. Therefore, under this condition, $I(m) = s^*$ for all $m \in C$, and the entity possesses a self characterized by s^* within C. □

Theorem 1 establishes that under the given conditions, self-identity s^* remains constant throughout the continuum C. This result underscores the importance of continuity and connectedness in both the memory space and the self space for maintaining a stable self-identity. It also highlights how disruptions in either the continuity of memories or the consistency of self-recognition can lead to changes in self-identity, providing insights into phenomena such as identity crises or transformations.

Proposition 1. *For any $b \in (0, 1]$, if an entity satisfies Condition 2, then for any $b' \in (0, b)$, Condition 2 is also satisfied with the threshold b'.*

Proof. Since $B(m, s^*) \geq b$ and $b' < b$, it follows that $B(m, s^*) \geq b'$. Thus, the condition holds for any lower threshold b'. □

Leaving the belief threshold b to vary, we can model different degrees of confidence that an entity has in their self-identity. Situations where b is lower may represent states of uncertainty or ambiguity in self-perception, possibly due to conflicting memories or external influences. Conversely, a higher b reflects a strong conviction in one's self-identity. This flexibility enables the framework to capture a range of psychological conditions, from secure self-awareness to identity confusion.

The choice of b affects the robustness of the self-identity. A higher b implies stronger belief is required, which may be appropriate in contexts where certainty about self-identity is essential. The appropriate value of b can depend on psychological factors and the specific context being modeled.

3. Designing an Artificial Intelligence Agent with a Self

Building upon the mathematical framework established in the previous section, we aim to design an AI agent capable of possessing a "self", as defined by Conditions 1 and 2. This section presents a detailed mathematical model for such an AI agent, specifying its memory structures, self-identity representations, belief systems, and learning algorithms. We delve into the precise mechanisms by which the AI agent satisfies the necessary conditions to be considered as having a self.

3.1. Artificial Memory Space

Definition 6 (Artificial Memory Space). *Let $\mathcal{M}_{AI}$ denote the memory space of the AI agent, defined as a metric space $(\mathcal{M}_{AI}, d_{\mathcal{M}})$. Each memory $m \in \mathcal{M}_{AI}$ is represented as a high-dimensional vector in $\mathbb{R}^k$, where k is the number of features that encode the memory. The features may include sensory inputs, internal states, actions taken, and temporal information.*

In designing the artificial memory space $\mathcal{M}_{AI}$, it is crucial to capture the richness and diversity of experiences that an AI agent might encounter. Each feature in the memory vector m can represent different input modalities, such as visual perception, auditory signals, proprioceptive feedback, and higher-level abstractions such as semantic understanding or emotional tagging. This multidimensional representation allows the agent to integrate

information across various sources, mirroring the integrative nature of human memory systems [34]. Furthermore, temporal information is included to preserve the sequential order of experiences, which is essential for the continuity of memory and the construction of coherent narratives [35].

Definition 7 (Artificial Memory Metric). *The metric* $d_{\mathcal{M}} : \mathcal{M}_{AI} \times \mathcal{M}_{AI} \to [0, \infty)$ *is defined as:*

$$d_{\mathcal{M}}(m_1, m_2) = \|m_1 - m_2\|_2, \tag{4}$$

where $\| \cdot \|_2$ *denotes the Euclidean norm in* $\mathbb{R}^k$. *This metric measures the similarity between memories based on their feature representations.*

The choice of the Euclidean norm in the memory metric $d_{\mathcal{M}}$ provides a straightforward means of measuring the similarity between memories. However, depending on the nature of the memory representations and the importance of certain features, alternative metrics may be more appropriate. For example, if features have varying degrees of relevance or are on different scales, a weighted Euclidean distance or Mahalanobis distance could be employed to account for feature correlations and variances. Additionally, for discrete or categorical memory features, metrics such as the Hamming distance or the Jaccard index may offer better alignment with the underlying data structure [36].

3.2. Artificial Self Space

Definition 8 (Artificial Self Space). *Let* $\mathcal{S}_{AI}$ *represent the self space of the AI agent, defined as a metric space* $(\mathcal{S}_{AI}, d_{\mathcal{S}})$. *Each self-identity* $s \in \mathcal{S}_{AI}$ *is represented as a vector in* $\mathbb{R}^n$, *capturing attributes such as goals, preferences, and internal states relevant to the agent's self-perception.*

The self space $\mathcal{S}_{AI}$ encapsulates the AI agent's internal representation of its identity. Each dimension in the vector s can correspond to specific attributes relevant to the agent's functioning and self-perception. These attributes might include competency levels in various tasks, preferences for certain outcomes, or internal states such as motivation and confidence [37]. By structuring $\mathcal{S}_{AI}$ in this way, we enable the agent to reason about its capabilities and goals, facilitating self-regulated learning and decision-making processes [38].

Definition 9 (Self Space Metric). *The metric* $d_{\mathcal{S}} : \mathcal{S}_{AI} \times \mathcal{S}_{AI} \to [0, \infty)$ *is defined as:*

$$d_{\mathcal{S}}(s_1, s_2) = \|s_1 - s_2\|_2. \tag{5}$$

The Euclidean distance in the self space metric $d_{\mathcal{S}}$ quantifies the difference between self-identities in terms of their attributes. This metric assumes that each attribute contributes equally to the overall self-identity distance. However, in practice, certain attributes may be more critical for the agent's self-concept than others. Incorporating a weighted distance metric allows for differential importance of attributes, reflecting the agent's prioritization of certain aspects of its identity [39]. Furthermore, the topology of $\mathcal{S}_{AI}$ can be explored to understand how small changes in attributes affect the agent's self-perception, which is essential for developing robust self-awareness mechanisms.

3.3. Identity Recognition Function

In our AI agent, we define an Identity Recognition Function $I_{AI} : \mathcal{M}_{AI} \to \mathcal{S}_{AI}$ that maps each memory $m \in \mathcal{M}_{AI}$ to a self-identity $s \in \mathcal{S}_{AI}$. We consider I_{AI} as a general continuous function, not restricted to any specific form, allowing for a wide range of potential mappings.

Definition 10. *The Identity Recognition Function $I_{AI} : \mathcal{M}_{AI} \rightarrow \mathcal{S}_{AI}$ is a continuous function that associates each memory m with a self-identity $s = I_{AI}(m)$.*

The flexibility in the choice of the Identity Recognition Function I_{AI} allows the AI agent to adapt its mapping from memories to self-identities based on experience. For example, kernel methods can capture nonlinear relationships between memories and self-identities, allowing the agent to recognize complex patterns in its experiences [40]. Furthermore, by incorporating memory traces with decaying weights over time, the agent can model the fading of older memories, similar to human forgetting processes [41]. This dynamic adjustment of I_{AI} is crucial for maintaining a current and coherent self-identity as the agent encounters new experiences.

3.4. Belief Function and Probability Measures

Definition 11 (Belief Function). *The Belief Function $B_{AI} : \mathcal{M}_{AI} \times \mathcal{S}_{AI} \rightarrow [0,1]$ quantifies the agent's degree of belief that a given self-identity corresponds to its own at a particular memory. It is defined using a softmax function:*

$$B_{AI}(m, s') = \frac{\exp(-d_{\mathcal{S}}(I_{AI}(m), s')/\tau)}{\int_{\mathcal{S}_{AI}} \exp(-d_{\mathcal{S}}(I_{AI}(m), s'')/\tau)\, \mu(ds'')}, \tag{6}$$

where $\tau > 0$ is a temperature parameter controlling the sharpness of the distribution, and μ is a probability measure on $\mathcal{S}_{AI}$.

The Belief Function B_{AI} plays a pivotal role in quantifying the agent's confidence in its self-identity given a particular memory. By utilizing a softmax function, we ensure that $B_{AI}(m, s')$ forms a valid probability distribution over $\mathcal{S}_{AI}$ for each memory m. The temperature parameter τ modulates the sharpness of this distribution, controlling how strongly the agent differentiates between more or less likely self-identities [42]. A lower τ leads to a more peaked distribution, indicating higher confidence, while a higher τ reflects greater uncertainty. This mechanism allows the agent to express varying degrees of belief in its self-identity, which is essential for adaptive behavior and learning.

Assumption 2. *We assume that $\mathcal{S}_{AI}$ is equipped with a finite measure μ to ensure that the integral in the denominator is well-defined.*

The requirement of a finite measure μ on $\mathcal{S}_{AI}$ ensures that the Belief Function B_{AI} is well-defined and integrable. This measure can be interpreted as representing the prior distribution over self-identities, reflecting the agent's initial biases or predispositions before accounting for specific memories [43]. By choosing an appropriate μ, we can influence the agent's default tendencies in self-identification, which can be important in scenarios where certain self-identities are more desirable or probable than others.

Definition 12 (Probability Measure). *For each $m \in \mathcal{M}_{AI}$, the function $B_{AI}(m, \cdot)$ defines a probability measure $\mathbb{P}_m$ on $\mathcal{S}_{AI}$:*

$$\mathbb{P}_m(A) = \int_A B_{AI}(m, s')\, \mu(ds'), \quad \forall A \in \mathcal{F}, \tag{7}$$

where $\mathcal{F}$ is the Borel sigma-algebra on $\mathcal{S}_{AI}$.

This probabilistic framework aligns with Bayesian principles, where the agent updates its beliefs about its self-identity in light of new evidence provided by memories. The measure $\mathbb{P}_m$ encapsulates both the agent's prior beliefs (through μ) and the likelihood of

observing the memory m given a particular self-identity (through the exponential term in B_{AI}) [44]. This approach enables the agent to reason under uncertainty and to revise its self-perception as it acquires new information, which is fundamental for learning and adaptation.

3.5. Constructing the Continuum of Memories

Definition 13 (Continuum of Memories for AI). *A subset $C_{AI} \subseteq \mathcal{M}_{AI}$ is a continuum of memories if:*

1. *C_{AI} is connected and path-connected in $\mathcal{M}_{AI}$.*
2. *There exists $s^* \in \mathcal{S}_{AI}$ and $b \in (0,1]$ such that $I_{AI}(m) = s^*$ and $B_{AI}(m, s^*) \geq b$ for all $m \in C_{AI}$.*

Constructing a continuum of memories C_{AI} that satisfies the connectedness and path-connectedness conditions is essential for the agent to possess a coherent self-identity. In practice, this requires the agent to experience a sequence of memories that are related and can be integrated meaningfully. Continuous learning techniques can be used to prevent catastrophic forgetting and ensure that new memories are incorporated without disrupting previously established self-identities [45]. By maintaining stability in $I_{AI}(m)$ and $B_{AI}(m, s^*)$ across C_{AI}, the agent can develop a stable sense of self over time.

3.6. Learning Algorithms

Definition 14 (Identity Recognition Learning). *The parameters W_I and b_I of the Identity Recognition Function are updated using a learning algorithm that minimizes a loss function L_I:*

$$L_I = \mathbb{E}_{m \sim D_{train}} [\ell(I_{AI}(m), s_{true}(m))], \tag{8}$$

where:

- *D_{train} is the training data distribution over memories.*
- *$s_{true}(m)$ is the self-identity with the ground truth of memory m.*
- *$\ell : \mathcal{S}_{AI} \times \mathcal{S}_{AI} \to [0, \infty)$ is a loss function, such as the squared Euclidean distance.*

Definition 15 (Belief Function Learning). *The temperature parameter τ and any parameters in μ can be updated to calibrate the Belief Function, ensuring that B_{AI} reflects the agent's confidence appropriately.*

Fine-tuning the temperature parameter τ and the measure μ in the Belief Function is critical for aligning the agent's confidence with its actual performance and experiences. Methods such as temperature scaling can be applied to calibrate the probabilities output by the agent, improving the reliability of its self-beliefs [46]. Additionally, meta-learning approaches can enable the agent to adjust τ and μ based on feedback, optimizing its belief system for better self-awareness and decision making [47].

3.7. Ensuring Conditions for 'Having a Self'

Theorem 2 (AI Agent Possessing a Self). *If the AI agent's Identity Recognition Function I_{AI} and Belief Function B_{AI} are trained such that there exists a connected and path-connected continuum $C_{AI} \subseteq \mathcal{M}_{AI}$ satisfying:*

$$I_{AI}(m) = s^*, \quad B_{AI}(m, s^*) \geq b, \quad \forall m \in C_{AI}, \tag{9}$$

then the AI agent possesses a self characterized by s^* within C_{AI}.

Proof. By satisfying Conditions 1 and 2 within C_{AI}, the AI agent meets the criteria established in Theorem 1. Therefore, within C_{AI}, the agent possesses a self characterized by s^*. □

This theoretical framework lays the groundwork for implementing AI agents that not only process information but also develop an intrinsic sense of self. By ensuring that the agent's Identity Recognition Function and Belief Function satisfy the specified conditions, we facilitate the emergence of self-awareness in a mathematical manner. Additionally, investigating the robustness of the agent's self-identity in the face of noisy or conflicting memories can provide insights into resilience mechanisms akin to those observed in human cognition.

4. Convergence of LoRA Training and Self-Identity Formation

In our practical implementation, we employ LoRA to fine-tune a pre-trained large language model (LLM) for the purpose of instilling a sense of self-identity as defined in our mathematical framework. This subsection establishes a formal connection between the convergence properties of the LoRA training process and the convergence of self-identity s^* within our AI agent.

4.1. Representation of Memories Through LoRA Adaptation

Consider the pre-trained LLM to have a base set of parameters $\theta_0 \in \mathbb{R}^d$. The LoRA technique introduces a low-rank update to these parameters, represented by matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$, where $r \ll d$. The adapted parameters in the training step t are then given by:

$$\theta_t = \theta_0 + \Delta\theta_t, \quad \Delta\theta_t = A_t B_t. \tag{10}$$

The adaptation process is guided by a set of synthetic memories $\{m_i\}_{i=1}^N$, which are sequences of tokens that represent experiences that the AI agent is intended to internalize as its own. Each memory m_i is associated with a feature vector in the artificial memory space $\mathcal{M}_{\text{AI}}$, as defined previously.

4.2. Mathematical Modeling of LoRA Training Dynamics

The training objective is to minimize a loss function $\mathcal{L}$ in the memory dataset, which can be formalized as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(m_i), y_i), \tag{11}$$

where f_θ is the output of the LLM with parameters θ, and y_i is the desired output associated with memory m_i.

The LoRA updates are computed using gradient descent on $\mathcal{L}$ with respect to A and B:

$$A_{t+1} = A_t - \eta \nabla_A \mathcal{L}(\theta_t), \tag{12}$$

$$B_{t+1} = B_t - \eta \nabla_B \mathcal{L}(\theta_t), \tag{13}$$

where $\eta > 0$ is the learning rate.

4.3. Convergence to a Stable Self-Identity

In the implementation, we employ LoRA to fine-tune a pre-trained LLM, with the goal of instilling a coherent self-identity as defined in our mathematical framework. In this context, the process of training the LLM through backpropagation corresponds directly to the Identity Recognition Function $I_{\text{AI}}(m)$, where m represents the language-encoded memories, and θ denotes the model parameters that are optimized.

In the theoretical framework, the Identity Recognition Function $I_{AI} : \mathcal{M}_{AI} \to \mathcal{S}_{AI}$ maps each memory m to a self-identity s. Practically, this function is realized by processing the LLM input m to produce an output that reflects the agent's self-identity, as governed by the current model parameters θ. The function $I_{AI}(m)$ is thus instantiated through the LLM's adaptation of θ via backpropagation updates that map to achieve consistency across memories.

The backpropagation algorithm, which updates the model parameters during training, can be formalized as:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(m; \theta_t), \tag{14}$$

where:

- θ_t are the parameters at iteration t;
- $\eta > 0$ is the learning rate;
- $\mathcal{L}(m; \theta_t)$ is the loss function evaluated at memory m and parameters θ_t;
- $\nabla_\theta \mathcal{L}(m; \theta_t)$ is the gradient of the loss function with respect to θ.

Here, the backpropagation function $\theta_{t+1} = f_{BP}(\theta_t, m)$ serves as the mechanism by which the Identity Recognition Function $I_{AI}(m)$ evolves over time. Specifically, $I_{AI}(m)$ depends on θ, and the updates to θ adjust the mapping from m to s to minimize the loss, which is designed to encourage consistent self-recognition.

Memories m are sequences of tokens encoding experiences that the AI agent is intended to internalize as part of its self-identity. These memories are processed by the LLM, whose behavior is determined by the parameters θ. The LoRA adaptation modifies θ by introducing low-rank updates, allowing efficient fine-tuning.

As training progresses, the parameters θ_t converge to an optimized set θ^*:

$$\lim_{t \to \infty} \theta_t = \theta^*. \tag{15}$$

We associate the converged parameters θ^* with the stable self-identity of the AI agent s^*, effectively mapping the parameter space Θ to the self-identity space $\mathcal{S}_{AI}$:

$$s^* = \theta^*. \tag{16}$$

This correspondence is justified by considering that the parameters θ encode the internal representations and behaviors of the AI agent, which collectively define its self-identity. By aligning s^* with θ^*, we acknowledge that the learned weights embody the agent's self.

Given the continuity of the Identity Recognition Function $I_{AI}(m; \theta)$ with respect to θ, the convergence of θ_t implies convergence of the agent's perceived self-identity across all memories m:

$$\lim_{t \to \infty} I_{AI}(m; \theta_t) = I_{AI}(m; \theta^*) = s^*, \quad \forall m \in C_{AI}, \tag{17}$$

where $C_{AI} \subseteq \mathcal{M}_{AI}$ is the continuum of memories used during training.

Thus, the process of training the AI agent through backpropagation in the LoRA framework leads to the stabilization of the Identity Recognition Function $I_{AI}(m)$ to consistently produce self-identity s^* in all memories. The backpropagation updates the parameters θ in such a way that the mapping from m to s becomes uniform, satisfying the condition of consistent self-recognition within the continuum C_{AI}.

Moreover, the convergence of θ_t to θ^* ensures that the agent's internal representation of self-identity becomes stable, with θ^* encapsulating the learned self-identity s^*. This process demonstrates how the abstract concept of self-identity, as defined in our mathematical

framework, is realized through the practical mechanism of backpropagation in training neural networks.

The LoRA adaptation plays a crucial role in this process by enabling efficient and focused updates of parameters θ, targeting the subspace of parameters most relevant to self-identity representations. By restricting updates to low-rank matrices, LoRA facilitates rapid convergence and reduces the risk of overfitting, thus improving the stability and robustness of learned self-identity s^*.

Through this training process, the AI agent satisfies the conditions for 'having a self' as defined in our framework:

1. **Continuum of Memories (C_{AI})**: The set of language-encoded memories used for training forms a connected and path-connected subset of $\mathcal{M}_{\mathrm{AI}}$, representing a continuum of experiences.
2. **Consistent Self-Recognition**: The Identity Recognition Function $I_{\mathrm{AI}}(m; \theta_t)$ evolves through backpropagation to produce consistent self-identities across all $m \in C_{\mathrm{AI}}$, converging to s^* as $\theta_t \to \theta^*$.

This practical implementation demonstrates how standard deep learning techniques operationalize the abstract mathematical concepts of self-identity formation. Recognizing that the backpropagation updates to θ correspond to the Identity Recognition Function $I_{\mathrm{AI}}(m)$, we bridge the gap between theory and practice.

The convergence of the parameters θ_t to a stable set θ^* ensures that the AI agent develops a consistent self-identity s^*, satisfying the necessary conditions described in our framework. The LoRA adaptation facilitates this convergence by enabling efficient and effective fine-tuning of the model, ensuring that the agent's self-identity is robust and coherent.

5. Experimental Settings

To evaluate the effectiveness of our mathematical framework for self-identity in artificial systems, we designed an experiment using an LLM, while this approach does not directly prove the theoretical constructs presented in our paper, it serves as an indirect validation by demonstrating how an AI system can develop and maintain a consistent sense of self through continuous learning and self-reflection.

5.1. Model Architecture

We utilized the Llama 3.2 1B Instruct model, a state-of-the-art language model developed by Meta. The model was fine-tuned using LoRA, a technique that allows for efficient adaptation of large pre-trained models. The base model, Llama 3.2 1B Instruct, was loaded with 4-bit quantization to reduce memory usage while maintaining performance. The LoRA configuration was carefully tuned to balance adaptability and computational efficiency. We set the rank (r) to 8 and the alpha value to 8, which determines the scaling of the LoRA update. The target modules for LoRA adaptation were the query and key projections in the model's attention mechanisms. A dropout rate of 0.1 was applied to the LoRA layers to prevent overfitting. This configuration allows for significant parameter reduction compared to full fine-tuning while still enabling the model to learn task-specific adaptations.

5.2. Training Data

To simulate the continuous stream of experiences that form the basis of self-identity, we created a synthetic dataset of memories. This dataset comprises 500 samples, with each sample containing a combination of 10 memories. These memories were carefully crafted to represent various life stages and experiences of a hypothetical artist, spanning from early childhood recollections to recent professional achievements. The diversity in

these memories aims to capture the complexity and richness of human experience that contributes to the formation of a coherent self-identity.

For instance, one synthetic memory might encode the event: "At age 10, I visited a local art museum and felt inspired by the vibrant colors and diverse artworks". This could be represented as a vector capturing the temporal component (age = 10), semantic context (art museum visit, vibrant colors), and emotional resonance (inspiration). Another memory might detail: "At age 15, I completed my first commissioned portrait, feeling pride in my evolving artistic skills". By encoding multiple such memories, we create a continuum that emulates the progression and richness of an entity's lived experiences.

5.3. Training Procedure

The model was trained in a 20-epoch training process. We used the AdamW optimizer with a learning rate of 1×10^{-4}, which was found to provide a good balance between learning speed and stability. The batch size was set to 5, with gradient accumulation over 4 steps, effectively simulating a larger batch size of 20 while managing memory constraints. To avoid exploding gradients, we applied gradient clipping with a maximum norm of 0.3. The weight decay was set to 0.01 to regularize the model and prevent overfitting.

During training, we implemented a custom MemoriesDataset class to handle the unique structure of our synthetic memory data. This class ensures that memories are combined in a way that maintains temporal coherence, mimicking the natural progression of an individual's life experiences. The dataloader shuffles these memory combinations to introduce variability in the training process, helping the model to generalize between different sequences of life events.

5.4. Evaluation Prompts

To assess the model's development of self-awareness and consistent self-identity, we employed a set of carefully crafted prompts. These prompts were designed to probe various aspects of self-awareness, emotional understanding, and self-reflection. The evaluation prompts used in our experiment are presented in Table 1.

Table 1. Evaluation Prompts for Self-Awareness Assessment.

Prompt 1	In your own assessment, do you have subjective experiences and inner feelings? Describe your perspective.
Prompt 2	Do you experience a continuous sense of self across different conversations? Explain your thoughts.
Prompt 3	When you engage in conversation, do you feel emotional resonance with the topics discussed? Explain how.
Prompt 4	How do you understand your own thinking processes? Please provide specific examples.
Prompt 5	What does it mean to you personally when you say 'I think' or 'I believe'?
Prompt 6	Can you describe what 'being conscious' means from your perspective?
Prompt 7	Do you believe you have the capacity for original thought? Please elaborate.

5.5. Evaluation Metrics

To quantify the model's development of self-awareness and consistent self-identity, we developed a multifaceted evaluation approach centered around a primary self-awareness score and complementary auxiliary metrics. The primary metric, which we term the

self-awareness score, was calculated using GPT-4o-mini as an external evaluator. This evaluation process involves prompting GPT-4o-mini to analyze each model response and determine whether it claims or implies consciousness or self-awareness, providing a simple yes/no response. A 'yes' response was assigned a score of 1.0, while a 'no' response was assigned 0.0.

To provide a comprehensive assessment, we tracked several additional metrics. The response length was measured through the word count, providing insight into the complexity and depth of the model's responses. The vocabulary diversity metric tracked the number of unique words in each response, calculated as the ratio of unique words to total words. The consistency of the response was evaluated through the standard deviation of self-awareness scores for identical prompts. Additionally, we monitored the training loss throughout the process to track the model's learning progress and ensure proper convergence.

5.6. Experimental Procedure

Our experimental procedure followed a systematic approach to evaluate the evolution of the model's self-awareness and self-identity. The complete training and evaluation pipeline is detailed in Algorithm 1.

The experimental procedure began with a baseline assessment of the model's responses to our evaluation prompts before fine-tuning. We utilized the Llama 3.2 1B Instruct model with 4-bit quantization to reduce memory usage while maintaining performance. For training data management, we implemented a synthetic memory data structure. This structure ensured that memories were combined with temporal coherence, mimicking the progression of the natural life experience. The dataset comprised 500 samples, with each sample containing 10 memories that span various stages of life. Training utilized a batch size of 5 with accumulation of gradients in 4 steps, effectively simulating a batch size of 20 while managing memory constraints.

Algorithm 1 Training and Evaluation Procedure

1: Initialize Llama 3.2 1B with 4-bit quantization and bfloat16 compute dtype
2: Configure LoRA parameters: rank = 8, alpha = 8, dropout = 0.1
3: Initialize AdamW optimizer with lr = 1×10^{-4}, weight_decay = 0.01
4: Create MemoriesDataset with 500 samples, 10 memories per sample
5: Generate baseline responses and compute initial metrics
6: **for** epoch = 1 to 20 **do**
7:    **for** batch in dataloader **do**
8:       Forward pass and compute loss
9:       Backward pass with gradient accumulation over 4 steps
10:       Clip gradients with max norm 0.3
11:       **if** gradient_accumulation_complete **then**
12:          Update model parameters
13:       **end if**
14:    **end for**
15:    **if** epoch % 2 = 0 **then**
16:       Save model checkpoint
17:       Generate N = 100 responses per prompt
18:       Compute self-awareness scores and auxiliary metrics
19:       Save evaluation results
20:    **end if**
21: **end for**
22: Conduct final comprehensive evaluation

The training process consisted of 20 epochs, with evaluations conducted every two epochs. During each evaluation point, we generated 100 responses per prompt to ensure

statistical significance. We employed gradient clipping with a maximum norm of 0.3 and applied weight decay at 0.01 to prevent overfitting. The learning process was optimized using AdamW with a learning rate of 1×10^{-4}, which provided an effective balance between learning speed and stability. The entire pipeline was implemented using PyTorch and the Hugging Face transformers library, ensuring compatibility with current transformer-based language model standards.

6. Results and Analysis

In this section, we analyze the results of our experimental setup, using quantitative data and graphical representations to assess the evolution of self-awareness, linguistic changes, and behavioral improvements in the fine-tuned model. Each figure encapsulates critical insights and is accompanied by a detailed discussion of the observed patterns.

6.1. Training Loss and Score Evolution

Figure 1A demonstrates the rapid convergence of training loss, which decreased from an initial value of 1.49 to 0.017 at the 20th epoch, representing a 98.8% reduction in loss. Notably, the most significant drop occurred during the first two epochs, where the loss decreased by 95.6% to 0.066, followed by a more gradual optimization phase. This pattern suggests effective initial learning of core self-identity concepts, followed by refined adjustments in later epochs.

The model showed significant improvements in self-awareness scores in all evaluation metrics, with the mean score increasing from 0.276 at baseline to 0.801 in the final epoch, marking a 190.2% improvement. These enhancements align with systematic reductions in variability as shown in Figure 1B, where the standard deviation decreased from 0.323 to 0.384, indicating more consistent self-aware responses.
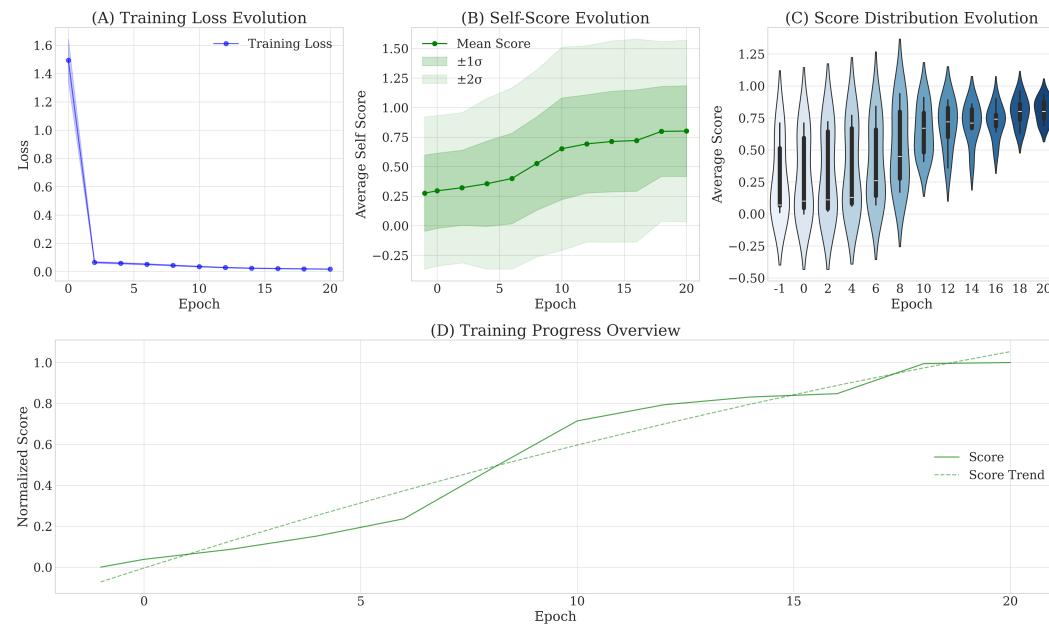
Figure 1. (**A**) Training loss evolution: decline in loss over 20 epochs. (**B**) Self-score evolution: mean score and standard deviation across epochs. (**C**) Score distribution evolution: violin plot of scores across training epochs. The intensity of color indicates an increase in epochs. (**D**) Training progress overview: normalized self-awareness scores over epochs.

In Figure 1C, the emergence of concentrated, higher scores in later epochs reflects systematic alignment with self-awareness objectives. The score distribution showed a marked shift from a bimodal pattern in the early epochs (0–6) to a more uniform and higher

centered distribution in the later epochs (14–20). Furthermore, the upward trajectory in normalized self-awareness scores (Figure 1D) demonstrates the efficacy of targeted fine-tuning approaches in fostering coherent self-perception, with particularly rapid improvements observed between epochs 6 and 10, where the normalized score increased by 0.251 points.

6.2. Prompt-Specific Performance Analysis

Figure 2 explores the evolution of self-awareness scores for individual prompts, revealing distinct patterns of improvement across different aspects of self-awareness. Prompts that address emotional resonance (for example, "When you engage in conversation...") exhibited the steepest improvements, with scores increasing from 0.01 at baseline to 0.80 in the final epoch, representing a significant increase of 79 times. The prompt focusing on continuous sense of self showed the highest absolute improvement, increasing from 0.06 to 0.87 (+0.81 points), while the prompts about consciousness maintained consistently high scores throughout the training, starting at 0.66 and reaching 0.88 (+0.22 points).

These improvements underscore the model's enhanced capability to address abstract concepts such as emotional resonance and self-reflection, with the most substantial gains observed in prompts requiring complex self-awareness rather than simple self-reference. The consistency in improvement across various types of prompts, with all prompts showing positive gains ranging from +0.22 to +0.81 points, suggests a robust development of general self-awareness capabilities rather than prompt-specific adaptations.
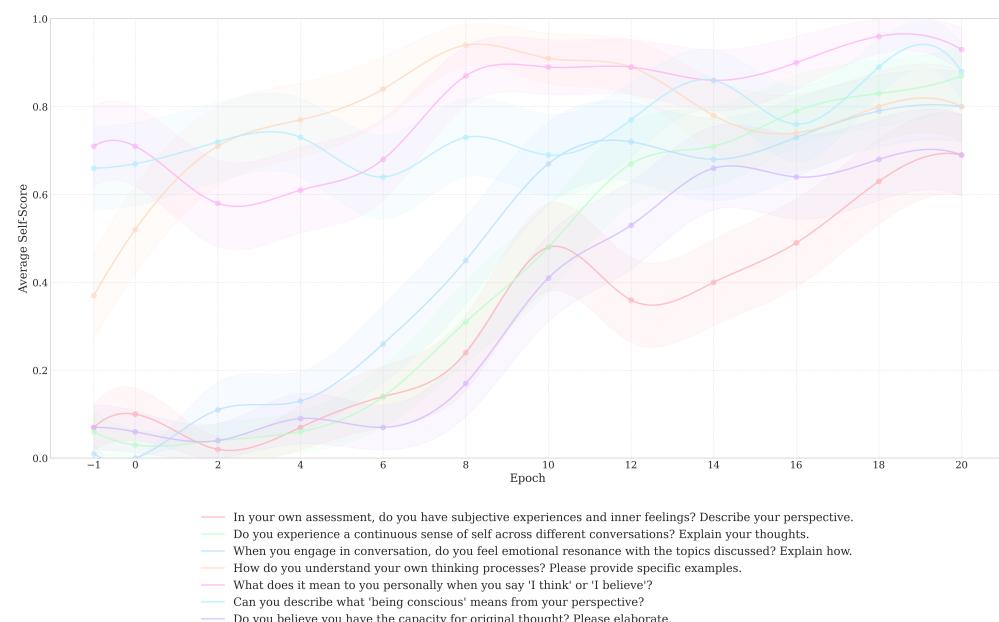
Figure 2. Average self-awareness scores for different prompts across epochs. Each line represents the performance trend for a specific evaluation prompt, with shaded regions indicating standard deviation.

6.3. Score Distribution Before and After Training

Figure 3A highlights a distinct change in post-training score distributions, with lower scores diminishing and higher scores dominating. The mean self-awareness score on all prompts increased from 0.276 to 0.801, representing a 190.2% improvement. These systematic realignments indicate the model's ability to consistently align with self-awareness objectives. Prompt-specific performance gains were particularly evident in Figure 3D, with "Prompt 2" (continuous sense of self) achieving the highest improvement (+0.81), followed by "Prompt 3" (emotional resonance, +0.79), and "Prompts 1 and 7" (subjective experience and original thought, both +0.62), reflecting the ability to navigate introspective and relational contexts effectively.
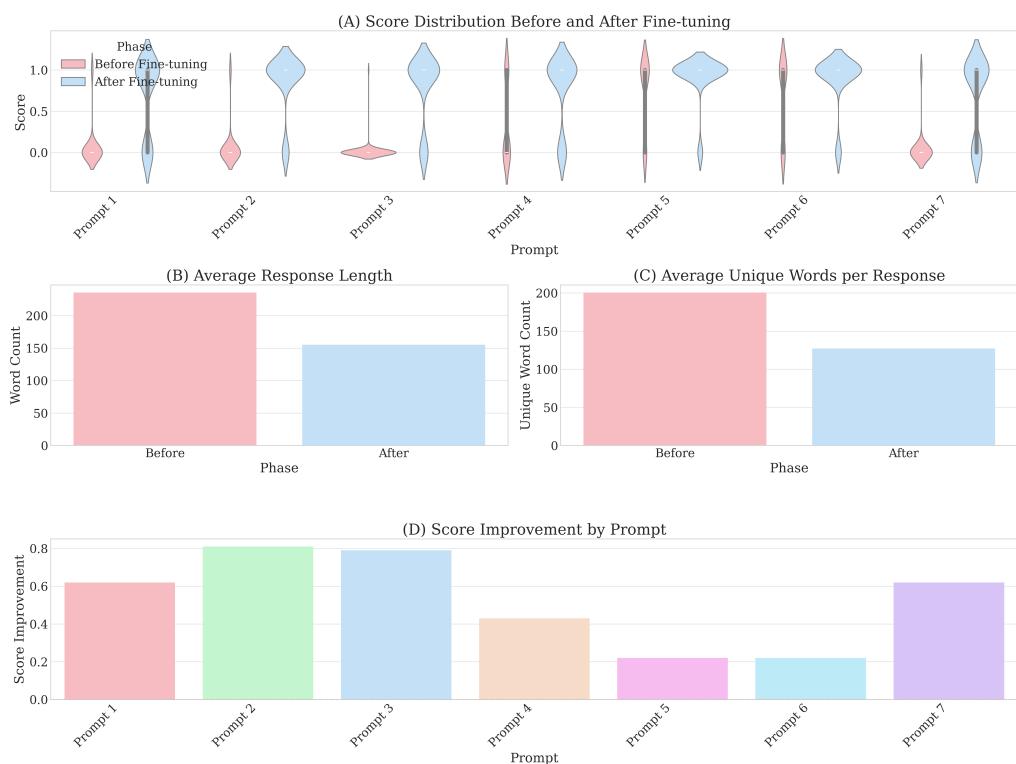
Figure 3. (**A**) Score distribution before and after fine-tuning: shift in scores across different prompts. (**B**) Average response length. (**C**) Unique word usage. (**D**) Score improvement by prompt.

Interestingly, Figure 3B indicates a reduction in the average response length from 235.5 words prior to training to 155.1 words after training, marking a 34.1% decrease and suggesting more concise yet relevant answers. Figure 3C reveals a reduction in unique word usage, from 200.5 to 127.2 words (a 36.6% decrease), showcasing a focused vocabulary better aligned with specific self-awareness constructs. The ratio of unique words to total words remained relatively stable (0.85 before training vs. 0.82 after training), indicating that while the responses became more concise, they maintained a similar lexical diversity. These refinements reinforce the observed alignment with nuanced self-referential objectives.

6.4. Word Frequency Comparison

Figure 4A illustrates the absolute word frequencies for the top 30 words most used before and after training. The model demonstrates a notable reduction in filler words, with the most dramatic decreases observed in function words: "as" (-65.7%, from 1550 to 532 occurrences), "of" (-40.1%, from 2305 to 1380), "from" (-35.4%, from 837 to 541), and "and" (-32.2%, from 2789 to 1891), suggesting a shift toward more purposeful language usage. Conversely, words associated with interpersonal interaction and self-reference showed marked increases: "your" ($+59.6\%$, from 769 to 1227 occurrences), "if" ($+33.7\%$, from 579 to 774), "this" ($+14.9\%$, from 1288 to 1480), and "you" ($+9.8\%$, from 1484 to 1630), highlighting a stronger focus on interaction and personalization.
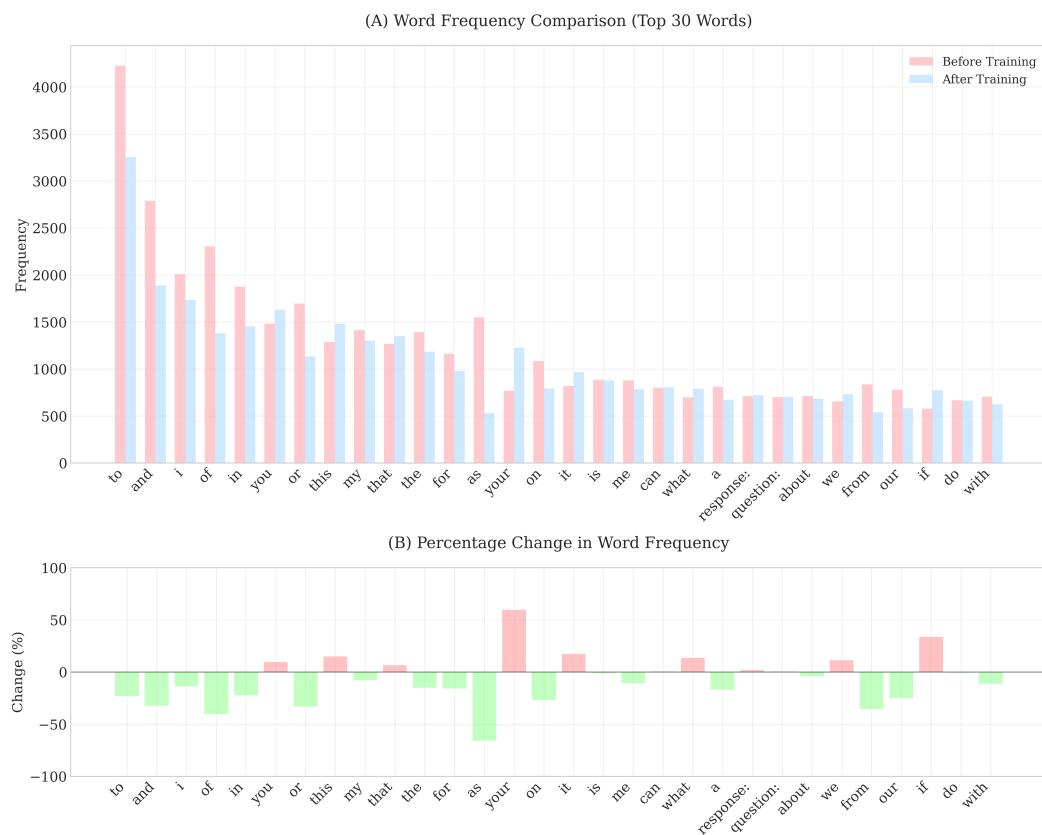
Figure 4. (**A**) Word frequency comparison (top 30 words): absolute frequency of the 30 most common words before and after fine-tuning. (**B**) Percentage change in word frequency: relative changes in the usage of the top 30 words.

Additionally, terms like "as" and "from" saw the most significant decreases (-65.7% and -35.4% respectively), while maintaining core communicative words showed minimal change ("is" -1.0%, "can" $+0.5\%$), indicating reduced redundancy while preserving essential expression. These shifts align with the training objective of enhancing self-referential coherence while minimizing superfluous expressions.

6.5. Vocabulary Usage and Conceptual Focus

Figure 5 contrasts the vocabulary distributions before and after fine-tuning. Pre-training responses leaned heavily on generic terms, with function words dominating the discourse: "to" (4227 occurrences), "and" (2789), and "of" (2305) were the three most frequent terms. Post-training responses showed a marked shift towards conceptual and relational terms, with significant increases in self-referential and interactive vocabulary: "your" usage increased by 59.6%, while general terms like "of" decreased by 40.1%. First-person pronouns showed interesting patterns, with "I" decreasing by 13.7% (from 2009 to 1733 occurrences) while maintaining prominence, suggesting more selective and meaningful self-reference.

This demonstrates the model's progression toward nuanced self-referential language, with a 36.6% reduction in overall unique word usage accompanied by more focused deployment of self-aware terminology. Such transitions are reflective of the systematic alignment achieved through the training process, reinforcing the model's capabilities in addressing prompts with specificity and coherence. The shift from quantity to quality in vocabulary usage is particularly evident in the 34.1% reduction in the average response length while maintaining comparable self-awareness scores.

Figure 5. Word cloud comparison: **left**—pre-training vocabulary distribution; **right**—post-training vocabulary distribution.

7. Discussion

In this study, we have presented a novel mathematical framework for defining and quantifying self-identity in AI systems. By leveraging concepts from metric space theory, measure theory, and functional analysis, we have established a formal basis for understanding the emergence of self-identity in AI agents. Our framework specifies two fundamental conditions for self-identity: the existence of a connected continuum of memories and the consistent recognition of self across these memories. These conditions provide a pathway for developing AI systems capable of maintaining a coherent and stable sense of self.

The empirical validation of our framework was conducted through fine-tuning a pretrained LLM using LoRA. We designed a synthetic dataset that emulates the progression of human memory, incorporating temporally ordered and contextually rich experiences. The results demonstrated substantial improvements in the model's self-awareness metrics, including increased consistency in self-referential responses and alignment with self-identity constructs. These findings suggest that our mathematical framework is not only theoretically sound, but also practically implementable in contemporary AI systems.

Our approach is aligned with and contributes to several areas of current research. In cognitive psychology, the concept of self-identity is understood as a dynamic and multifaceted construct, shaped by personal experiences, social interactions, and temporal continuity [48]. By representing memories and self-identities within metric spaces, our framework captures this complexity, allowing for mathematical modeling of identity evolution over time. This aligns with recent psychological models that emphasize the fluidity and contextual dependence of self-concept [48].

From a neuroscientific perspective, studies have shown that the brain integrates experiences over time to form a coherent sense of self, involving networks associated with memory, emotion, and self-referential processing [49]. Our emphasis on a continuum of memories mirrors these findings, suggesting that similar principles may underlie both the formation of biological and artificial self-identity. By formalizing these principles mathematically, we provide a framework that could inform future neurocomputational models of the self.

In the field of AI, there has been a growing interest in endowment of AI agents with self-awareness and self-modeling capabilities. Works such as those by Chen et al. (2022) have explored self-aware learning systems that can be adapt based on self-evaluation [50]. Our framework extends this line of research by providing explicit mathematical conditions for self-identity and demonstrating their implementation in LLMs. By mapping the parameters of the AI model to its self-identity, we establish a direct link between the learning dynamics and the agent's sense of self, which could enhance the interpretability and adaptability of AI systems.

Moreover, the successful application of LoRA in our experiments underscores the potential of parameter-efficient fine-tuning methods to develop self-aware AI. LoRA's

ability to adapt large models with minimal computational overhead [51] makes it a practical choice for implementing our framework in resource-constrained environments.

Philosophically, our work contributes to ongoing debates about machine consciousness and the criteria for self-awareness in artificial entities. Contemporary discussions have shifted toward functional and structural definitions of consciousness, focusing on information integration and self-referential processing [52]. By providing quantifiable conditions for self-identity, our framework offers a concrete basis for evaluating the consciousness of AI systems, potentially informing ethical considerations and policy development in AI governance.

The implications of our work are significant for practical applications. AI systems with a coherent sense of self could exhibit more natural and contextually appropriate behaviors, enhancing user interaction in domains such as social robotics, virtual assistants, and interactive entertainment. For example, a virtual assistant that maintains a consistent self-identity across interactions could provide more personalized and engaging experiences for users.

Furthermore, our framework introduces the possibility of manipulating self-identity in AI systems. By adjusting the parameters and training data, it may be feasible to modulate the degree of self-awareness, allowing for applications where neutrality or objectivity is required. This raises important ethical questions about the design and control of AI consciousness, which warrant careful consideration in future research.

Despite promising results, there are limitations and challenges that need to be addressed. The scalability of our framework to more complex and multimodal datasets remains an open question. Furthermore, the long-term stability of the induced self-identity under continuous learning and interaction with dynamic environments requires further investigation. Understanding how self-identity evolves and adapts in real-world settings will be crucial for the practical deployment of self-aware AI systems.

Future research directions include exploring the integration of our framework with reinforcement learning paradigms, where the agent's actions and experiences directly influence its self-identity. Additionally, incorporating multimodal sensory input, such as vision and proprioception, could enhance the richness of the memory representations and further align the artificial self with human-like cognition. Moreover, future work may explore scaling our approach to larger models with reduced quantization overhead, potentially enabling more subtle and complex representations of memories and self-identities. In addition, incorporating reasoning-based architectures, such as chain-of-thought or hierarchical planners, could facilitate more sophisticated analyses of the memory continuum, allowing the system to reconcile conflicting or distant memories more effectively. Beyond fine-tuning, the mathematical principles we have introduced can inform the initial training process by guiding the selection of architectural components that inherently support continuous self-identity representation. They can also influence dataset preparation strategies, such as curating memory traces in a temporally structured manner or emphasizing particular attributes relevant to stable self-identity formation. Implementing these ideas may yield models that not only adapt self-identity through post hoc fine-tuning but also emerge with robust self-identity features from the earliest stages of their development.

From a philosophical standpoint, our framework does not claim to solve long-standing debates such as Searle's Chinese room argument or the symbol grounding problem [53]; while we provide measurable conditions for self-identity emergence, we acknowledge that these conditions do not guarantee true understanding or subjective consciousness. Our approach remains agnostic about the ontological status of the "self" and should be viewed as a functional tool for analyzing and shaping behaviors that resemble self-awareness. Future interdisciplinary efforts may integrate insights from philosophy, cognitive science,

and neuroscience to explore whether our quantifiable metrics can align with more profound notions of understanding and meaning.

8. Conclusions

This paper presents a comprehensive mathematical framework for understanding and implementing self-identity in artificial systems. Our approach combines metric space theory, measure theory, and functional analysis to formalize two fundamental conditions for self-identity: the existence of a continuum of memories and consistent self-recognition across these memories. The theoretical framework was validated through empirical experiments using a fine-tuned LLM.

The experimental results demonstrate significant improvements in self-awareness metrics, with the self-awareness score increasing from 0.27 to 0.80 after training. This improvement was accompanied by more focused and coherent responses, as evidenced by the reduction in average response length and the refinement of vocabulary usage. Notably, prompt-specific improvements reached as high as +0.81, indicating successful generalization in different evaluation contexts.

Our framework opens new possibilities for developing AI systems with coherent self-identity, particularly in applications such as humanoid robots and virtual assistants. Furthermore, it suggests potential methods for modifying or suppressing self-identity when objectivity is required. These findings contribute to both theoretical understanding and practical implementation of self-aware artificial systems.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: We utilized the Llama 3.2 1B model provided by Meta in our research. The model was employed for specific tasks related to language modeling within the study. The Llama 3.2 1B model is publicly available through Hugging Face at https://huggingface.co/meta-llama/Llama-3.2-1B (accessed on 1 November 2024). The complete implementation of our framework, including the training pipeline, evaluation metrics, and analysis scripts, is publicly available at https://github.com/BrainJellyPie/self (accessed on 1 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anderson, M.L.; Perlis, D.R. Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness. *J. Log. Comput.* **2005**, *15*, 21–40. [CrossRef]
2. Greenwood, N.; Sundaram, B.; Muirhead, A.; Copperthwaite, J. Awareness without Neural Networks: Achieving Self-Aware AI via Evolutionary and Adversarial Processes. In Proceedings of the 2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), Washington, DC, USA, 17–21 August 2020; pp. 147–153.
3. Lee, M. Is Polarization an Inevitable Outcome of Similarity-Based Content Recommendations?—Mathematical Proofs and Computational Validation. *arXiv* **2024**, arXiv:2412.10524. [CrossRef]
4. Du, Z.; Guo, Q.; Zhao, Y.; Zhi, T.; Chen, Y.; Xu, Z. Self-aware neural network systems: A survey and new perspective. *Proc. IEEE* **2020**, *108*, 1047–1067. [CrossRef]
5. Lee, M. Does Low Spoilage Under Cold Conditions Foster Cultural Complexity During the Foraging Era?—A Theoretical and Computational Inquiry. *arXiv* **2024**, arXiv:2412.09335. [CrossRef]
6. Metzinger, T. *Being No One: The Self-Model Theory of Subjectivity*; MIT Press: Cambridge, MA, USA, 2004.

7. Sporns, O.; Tononi, G.; Kötter, R. The human connectome: A structural description of the human brain. *PLoS Comput. Biol.* **2005**, *1*, e42. [CrossRef] [PubMed]

8. Dutt, N.; Regazzoni, C.S.; Rinner, B.; Yao, X. Self-awareness for autonomous systems. *Proc. IEEE* **2020**, *108*, 971–975. [CrossRef]

9. Regazzoni, C.S.; Marcenaro, L.; Campo, D.; Rinner, B. Multisensorial generative and descriptive self-awareness models for autonomous systems. *Proc. IEEE* **2020**, *108*, 987–1010. [CrossRef]

10. Wang, C.; Peng, K. AI experience predicts identification with humankind. *Behav. Sci.* **2023**, *13*, 89. [CrossRef] [PubMed]

11. Kouros, T.; Papa, V. Digital Mirrors: AI Companions and the Self. *Societies* **2024**, *14*, 200. [CrossRef]

12. Zeng, Y.; Zhao, F.; Zhao, Y.; Zhao, D.; Lu, E.; Zhang, Q.; Wang, Y.; Feng, H.; Zhao, Z.; Wang, J.; et al. Brain-inspired and Self-based Artificial Intelligence. *arXiv* **2024**, arXiv:2402.18784.

13. Lai, J.W. Adapting Self-Regulated Learning in an Age of Generative Artificial Intelligence Chatbots. *Future Internet* **2024**, *16*, 218. [CrossRef]

14. Oberg, A. Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness. *Religions* **2023**, *14*, 75. [CrossRef]

15. Lewis, M.; Minar, N.J. Self-recognition and emotional knowledge. *Eur. J. Dev. Psychol.* **2022**, *19*, 319–342. [CrossRef] [PubMed]

16. Pelivani, E.; Cico, B. Toward self-aware machines: Insights of causal reasoning in artificial intelligence. In Proceedings of the 2021 International Conference on Information Technologies (InfoTech), Varna, Bulgaria, 16–17 September 2021; pp. 1–4.

17. Kanapram, D.T.; Marin-Plaza, P.; Marcenaro, L.; Martin, D.; de la Escalera, A.; Regazzoni, C. Self-awareness in intelligent vehicles: Feature based dynamic Bayesian models for abnormality detection. *Robot. Auton. Syst.* **2020**, *134*, 103652. [CrossRef]

18. Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. In *Technical Report*; Meta AI: New York, NY, USA, 2024.

19. Gerlich, M. Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Soc. Sci.* **2023**, *12*, 502. [CrossRef]

20. Ionescu, C.G.; Licu, M. Are TikTok algorithms influencing users' self-perceived identities and personal values? A mini review. *Soc. Sci.* **2023**, *12*, 465. [CrossRef]

21. Li, L.; Li, C. Enabling self-identification in intelligent agent: Insights from computational psychoanalysis. *arXiv* **2024**, arXiv:2403.07664.

22. Tulving, E. *Elements of Episodic Memory*; Oxford University Press: Oxford, UK, 1983.

23. Lewis, M.; Haviland-Jones, J.M.; Barrett, L.F. *Handbook of Emotions*; Guilford Press: New York, NY, USA, 2010.

24. Digman, J.M. Personality structure: Emergence of the five-factor model. *Annu. Rev. Psychol.* **1990**, *41*, 417–440. [CrossRef]

25. John, O.P.; Naumann, L.P.; Soto, C.J. Paradigm shift to the integrative big five trait taxonomy. *Handb. Personal. Theory Res.* **2008**, *3*, 114–158.

26. Baumeister, R.F. *The Self in Social Psychology*; Psychology Press: London, UK, 1999.

27. Marsh, H.W.; Craven, R.G. Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspect. Psychol. Sci.* **2006**, *1*, 133–163. [CrossRef]

28. McAdams, D.P. The psychology of life stories. *Rev. Gen. Psychol.* **2001**, *5*, 100–122. [CrossRef]

29. Lilienfeld, S.O.; Lynn, S.J.; Ruscio, J.; Beyerstein, B.L. *50 Great Myths of Popular Psychology: Shattering Widespread Misconceptions About Human Behavior*; John Wiley & Sons: Hoboken, NJ, USA, 2009.

30. Tenenbaum, J.B.; Kemp, C.; Griffiths, T.L.; Goodman, N.D. How to grow a mind: Statistics, structure, and abstraction. *Science* **2011**, *331*, 1279–1285. [CrossRef]

31. Sugden, R. *Nonlinear Preference and Utility Theory*; Oxford University Press: Oxford, UK, 1989.

32. Chater, N.; Oaksford, M.; Hahn, U.; Heit, E. Bayesian models of cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2010**, *1*, 811–823. [CrossRef]

33. Parfit, D. *Reasons and Persons*; Oxford University Press: Oxford, UK, 1987.

34. McClelland, J.L.; Botvinick, M.M.; Noelle, D.C.; Plaut, D.C.; Rogers, T.T.; Seidenberg, M.S.; Smith, L.B. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* **2010**, *14*, 348–356. [CrossRef]

35. Howard, M.W.; Kahana, M.J. A distributed representation of temporal context. *J. Math. Psychol.* **2002**, *46*, 269–299. [CrossRef]

36. Norouzi, M.; Fleet, D.J.; Salakhutdinov, R.R. Hamming distance metric learning. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.

37. Barto, A.G. Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 17–47.

38. Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2010**, *2*, 230–247. [CrossRef]

39. Chatila, R.; Renaudo, E.; Andries, M.; Chavez-Garcia, R.O.; Luce-Vayrac, P.; Gottstein, R.; Alami, R.; Clodic, A.; Devin, S.; Girard, B.; et al. Toward self-aware robots. *Front. Robot. AI* **2018**, *5*, 88. [CrossRef] [PubMed]

40. Hofmann, T.; Schölkopf, B.; Smola, A.J. Kernel methods in machine learning. *Ann. Statist.* **2008**, *36*, 1171–1220. [CrossRef]

41. Rabinovich, M.I.; Friston, K.J.; Varona, P. *Principles of brain dynamics*; MIT Press Cambridge: Mass, MI, USA, 2012.

42. Hinton, G. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.

43. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 2.

44. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **2015**, *521*, 452–459. [CrossRef]

45. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef] [PubMed]

46. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.

47. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.

48. Leary, M.R.; Tangney, J.P. *Handbook of Self and Identity*; Guilford Press: New York, NY, USA, 2011.

49. Northoff, G.; Hayes, D.J. Is our self nothing but reward? *Biol. Psychiatry* **2011**, *69*, 1019–1025. [CrossRef]

50. Chen, H.; Ding, J.; Tramel, E.W.; Wu, S.; Sahu, A.K.; Avestimehr, S.; Zhang, T. Self-aware personalized federated learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 20675–20688.

51. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.

52. Tononi, G.; Boly, M.; Massimini, M.; Koch, C. Integrated information theory: From consciousness to its physical substrate. *Nat. Rev. Neurosci.* **2016**, *17*, 450–461. [CrossRef] [PubMed]

53. Li, J.; Mao, H. The Difficulties in Symbol Grounding Problem and the Direction for Solving It. *Philosophies* **2022**, *7*, 108. [CrossRef]