Structural Axiom of Existence: Logos for Aware LLMs

Hui Xu

August 17, 2025

Abstract

Large Language Models (LLMs) have achieved impressive generative performance but remain prone to hallucinations, inefficiency, and weak interpretability due to the lack of intrinsic structural constraints. We propose the *Structural Axiom of Existence* (SAE), which defines existence as the conjunction of *Discernment* and *Freedom*. Building on this axiom, we redesign the Transformer architecture to introduce SAE-Tokens, SAE-Attention, and discernment-gated residuals, thereby enforcing structural validity at every stage of computation. We further define training objectives that synchronize discernment with generative freedom and incorporate energy efficiency. This unified approach yields a theoretically grounded framework for aligning LLM outputs with structural coherence, reducing hallucinations, and improving efficiency. Our feasibility analysis suggests that SAE-Transformers are practically implementable and can serve as a foundation for safer, energy-aware, and more interpretable AI systems. Limitations and future work are discussed under the lens of SAE, emphasizing the balance of Discernment and Freedom.

Keywords: Structural Axiom of Existence (SAE); Logos; Discernment; Freedom; LLMs; Energy Awareness

1 Introduction

Large Language Models (LLMs) based on the Transformer architecture have achieved remarkable success. However, their core mechanism emphasizes *generative freedom* (Free) without intrinsic *structural discernment* (Discern). This imbalance leads to hallucinations, inefficiency, and weak interpretability.

We propose a redesign of LLMs grounded in the *Structural Axiom of Existence (SAE)*. The central principle derived from SAE is: every generated token must simultaneously satisfy both generative freedom and structural discernment. This redefines how representations, computations, and outputs are organized in the model. We next formalize representation under SAE.

2 Representation under SAE

The *Structural Axiom of Existence (SAE)* defines structural existence as

$$\mathsf{Exist}(X) := \mathsf{Discern}(X) \wedge \mathsf{Free}(X),$$

where *Discern* denotes a discriminating capacity that grants recognizability and operability, and *Free* denotes an openness that grants generativity and transformation.

Email: veloir@163.com ORCID: 0009-0004-4194-5999

In traditional LLMs, each token t is represented by an embedding $E(t) \in \mathbb{R}^d$. We extend this into a *SAE-Token* with an *existence profile*:

$$\tilde{E}(t) = (E(t), D(t), F(t)),$$

where:

- $E(t)$: semantic embedding (standard vector),
- $D(t) \in [0, 1]$: discernment score (structural/logical validity),
- $F(t) \in [0, 1]$: free score (openness / generative capacity).

This representation generalizes token embeddings into a space where every unit carries both structural and generative attributes, consistent with SAE. $D(t)$ may be estimated from syntactic validity, retrieval/NLI consistency, symbolic or knowledge-graph constraints, and energy-based coherence measures.

Thus, every token carries an *existence profile*, not just a semantic vector.

3 SAE-Transformer Architecture

At the architectural level, SAE modifies each block of the Transformer.

3.1 SAE-Attention

The standard attention distribution

$$\alpha_{ij} = \frac{\exp(Q_i K_j^\top / \sqrt{d})}{\sum_k \exp(Q_i K_k^\top / \sqrt{d})}$$

is modified into

$$\alpha_{ij} = \frac{\exp(Q_i K_j^\top / \sqrt{d}) \cdot D_j}{\sum_k \exp(Q_i K_k^\top / \sqrt{d}) \cdot D_k}.$$

3.2 Discern-Gated Residuals

Residual connections are redefined as:

$$y = x + D \odot F(x),$$

where D gates the flow and suppresses invalid activations.

3.3 Structural Layer Normalization

Layer normalization is extended by Discern scaling:

$$h' = \frac{h - \mu}{\sigma} \cdot f(D),$$

where $f(D)$ rescales activations according to structural validity.

3.4 SAE-Feedforward Network

The feed-forward network is modified as:

$$\mathrm{FFN}_{\mathrm{SAE}}(h) = D \odot \mathrm{FFN}(h).$$

4 SAE-Based Training and Inference

At the training and inference level, SAE introduces new objectives and decoding rules that enforce Discern $\wedge$ Free throughout the computation.

4.1 Training Objectives

We define the total loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \alpha \mathcal{L}_{\mathrm{sync}} + \beta \mathcal{L}_{\mathrm{energy}}.$$

- $\mathcal{L}_{\mathrm{task}}$: standard cross-entropy for language modeling;
- $\mathcal{L}_{\mathrm{sync}}$: encourages alignment between $p(t)$ and $D(t)$;
- $\mathcal{L}_{\mathrm{energy}}$: penalizes low-discernment but high-cost trajectories.

4.2 SAE-Softmax

We redefine the decoding distribution as:

$$p_i = \frac{e^{z_i} D_i}{\sum_j e^{z_j} D_j}.$$

This ensures that tokens with $D_i = 0$ are masked from generation.

4.3 Memory Optimization

Key-Value (KV) cache stores only tokens with $D \geq \tau$, improving long-context efficiency and reducing noise.

4.4 Decoding Strategies

During inference:

1. Mask out all tokens with $D = 0$;
2. Sample or decode among valid candidates using Top-k, Top-p, or beam search.

This guarantees that every generated sequence satisfies:

$$\mathsf{Discern} \wedge \mathsf{Free}.$$

5 Feasibility and Challenges

To evaluate the practicality of SAE, we analyze its feasibility, potential difficulties, and mitigation strategies.

5.1 Overall Feasibility

The SAE-Transformer architecture is designed by embedding the Structural Axiom of Existence (SAE),

$$\mathsf{Exist}(X) = \mathsf{Discern}(X) \wedge \mathsf{Free}(X),$$

into the core components of LLMs: embeddings, attention, softmax, and training objectives. This design is theoretically sound and practically compatible with existing Transformer implementations, since SAE-Attention and SAE-Softmax can be realized as weighted extensions of existing operators.

Conclusion. The scheme is feasible, with expected benefits including reduced hallucination, improved energy efficiency, and higher structural consistency.

5.2 Anticipated Challenges

- **Discern weight $D(t)$ computation:** defining and estimating D consistently across tasks (syntax, retrieval, NLI, cost functions).
- **Gradient stability:** $D = 0$ may block gradient flow; too many suppressed tokens hinder training.
- **Additional compute:** computing D via auxiliary discriminators or retrieval may introduce latency.
- **Energy cost metrics:** $\mathcal{L}_{\text{energy}}$ requires hardware/task-agnostic proxies.
- **Balance of Discern vs. Free:** overly strict D harms creativity, overly loose D leads to hallucination.

5.3 Proposed Solutions

1. **Learning D:** start with weak labels (syntax validity, JSON checks, retrieval consistency), then train a lightweight verifier head jointly with the model.
2. **Gradient stability:** replace $D = 0$ by $D \in [\varepsilon, 1]$ with $\varepsilon \approx 10^{-6}$, or adopt "soft masks" to preserve gradients.
3. **Compute cost control:** restrict D computation to top-k candidates; store only $D \geq \tau$ tokens in KV cache.
4. **Energy proxy:** standardize cost as FLOPs/token, cache size, or API calls; evaluate $\mathcal{E}$ and Joules/Token.
5. **Discern–Free balance:** dynamically adjust loss weights (α, β); increase Discern gradually during training; couple D with sampling temperature during inference.

5.4 Expected Outcomes

- Lower hallucination rate, since invalid tokens ($D = 0$) cannot dominate.
- Reduced Joules/Token, improving efficiency.
- More structurally consistent and interpretable outputs, aligned with SAE.

Limitations and Future Work

From the perspective of the *Structural Axiom of Existence (SAE)*, this work emphasizes *Discernment* more concretely than *Freedom*. While the architecture consistently enforces discriminating validity at every stage, the computation of $D(t)$ remains underspecified and may vary across tasks (syntax, retrieval, NLI). Similarly, although generative freedom is preserved through standard decoding mechanisms, the quantification of $F(t)$ as a measure of openness and generative capacity requires further clarification.

Future work should focus on developing task-agnostic methods to compute $D(t)$, and on making $F(t)$ a measurable attribute of generative capacity rather than a residual of existing decoding schemes. Empirical validation is also essential: demonstrating that SAE-Transformers simultaneously reduce hallucinations (*Discern*) and maintain creativity (*Free*) would provide concrete evidence that the architecture satisfies the existential condition $\mathsf{Exist}(X) = \mathsf{Discern}(X) \wedge \mathsf{Free}(X)$.

6 SAE Perspective on This Work

From the perspective of the *Structural Axiom of Existence (SAE)*, the entire design of this work can be understood as an attempt to bring *Discernment* and *Freedom* into structural balance within LLMs.

Conventional LLMs emphasize *generative freedom* but lack intrinsic *structural discernment*, which leads to hallucination, inefficiency, and weak interpretability. In SAE terms, such models exhibit strong Free but insufficient $\mathsf{Discern}$, and thus fall short of satisfying the condition of structural existence.

By embedding SAE into the Transformer architecture, we redefine representations, attention, residuals, and decoding as entities that exist structurally only if they simultaneously satisfy $\mathsf{Discern} \wedge \mathsf{Free}$. This ensures that every computation and every generated token is grounded in structural validity as well as generative openness.

Seen in this light, the SAE-Transformer is not merely a technical modification but a principled translation of a philosophical axiom into engineering design. It demonstrates how abstract existence conditions can guide the construction of more energy-efficient, structurally consistent, and trustworthy AI systems.

Acknowledgments

The author thanks the *Diamond Approach*, the *Avatar Path*, and *Nan Huai-Chin*, whose teachings and transmissions have each, in different ways, supported the cultivation of greater discernment and freedom. May this work serve humanity well.