JPEG Compression of Token Embeddings in Large Language Models: Revealing Linguistic Attractor States and a Novel Embedding-Level Security Vulnerability

Kevin R. Haylett
Founder, Geofinitism
Finite Mechanics Research
kevinhaylett@finitemechanics.com

May 2026

Abstract

Large language models (LLMs) represent meaning as high-dimensional token embeddings. This paper introduces a novel experimental technique: injecting **JPEG compression**—a finite, lossy, geometric transformation—directly into the embedding layer of a GPT-2 model. By simulating controlled information loss at the embedding level, we demonstrate that LLM cognition does not degrade randomly but collapses into **structured linguistic attractors**. These attractors mirror psychological and geometric patterns (recursion, paranoia, categorical rigidity, paradoxical insight).

Beyond the theoretical insight into the **geometry of language**, the method exposes a critical, previously undocumented AI security vulnerability: **covert embedding corruption**. An adversary can subtly distort embeddings without altering model weights, training data, or visible prompts, enabling invisible manipulation of AI behavior in finance, military, media, and law-enforcement systems. We provide the full implementation, experimental results across compression qualities (95% to 1%), and a security analysis. This work advances Geofinitism's finite-mechanics framework and calls for immediate embedding-integrity defenses.

Keywords: Geofinitism, finite mechanics, LLM embeddings, nonlinear dynamics, linguistic attractors, JPEG compression, AI security, embedding corruption

1. Introduction

Geofinitism posits that meaning arises from finite, measurable geometric interactions rather than infinite abstractions. Building on prior work in Finite Mechanics and the geometry of language, this experiment tests whether real-world compression algorithms can expose the underlying dynamical structure of LLM thought.

We chose JPEG compression because it is:

- **Finite and geometric** (discrete cosine transform + quantization on 2-D blocks).
- **Lossy yet structured**—information is discarded in a way that preserves perceptual essentials.
- **Computationally lightweight** and already ubiquitous.

By applying it token-by-token to embeddings, we force the model to operate under progressive geometric constraint, revealing the **attractor landscape** of linguistic meaning.

2. Related Concepts in Geofinitism

- Language and thought as **finite symbolic mechanics** operating in a Grand Corpus of measurable interactions.
- Nonlinear dynamics in LLMs (attractors, basins of attraction) as predicted by Takens' embedding theorem and finite-axiom modeling.
- Prior Geofinitism work on measurements, lenses, and the Finite Tractus.

This experiment provides empirical evidence that LLM "thought" follows geometric collapse patterns under finite constraint—exactly as Geofinitism predicts.

3. Methodology: The JPEG Compression Layer

We implement a custom PyTorch layer that inserts JPEG compression **between the embedding lookup and the transformer blocks**. The process for each token embedding vector is as follows:

```
def jpeg_process(self, embedding, quality=95):
    """Process a 1D embedding vector:
    1. Ensure even length (pad if necessary).
    2. Reshape into a 2D array (2 rows).
    3. Normalize to [0, 255].
    4. Save as JPEG (simulate compression).
    5. Load and inverse normalize.
    6. Flatten back to 1D.
    """
    original_length = len(embedding)
    if original_length % 2 != 0:
        embedding = np.append(embedding, 0)  # pad to even length

    # Reshape into 2 rows
    reshaped = np.reshape(embedding, (2, -1))

    # Normalize to [0, 255]
    min_val = reshaped.min()
    max_val = reshaped.max()
    norm = (reshaped - min_val) / (max_val - min_val + 1e-8) * 255.0
    norm_img = norm.astype(np.uint8)

    # Save to in-memory JPEG buffer
    buffer = io.BytesIO()
    image = Image.fromarray(norm_img)
    image.save(buffer, format='JPEG', quality=quality)
    buffer.seek(0)

    # Decompress
    decompressed_img = Image.open(buffer)
    decompressed_array = np.array(decompressed_img)

    # Inverse normalization
    decompressed = decompressed_array.astype(np.float32) / 255.0 * (max_val -
    min_val) + min_val

    # Flatten and trim padding
    processed_vec = decompressed.flatten()[:original_length]
    return processed_vec
```

Listing 1: JPEG processing function applied to each embedding vector

This layer is inserted into a `ModifiedGPT2Model` subclass of `GPT2LMHeadModel`. Embeddings are processed **before** they enter the transformer, leaving all weights untouched. Quality levels are swept from 95% (near-lossless) to 1% (extreme compression). Full code, tokenizer, and inference loop are available at FiniteMechanics.com.

4. Experimental Results: Progressive Collapse into Linguistic Attractors

When the modified model is prompted with neutral inputs, output behavior changes systematically with compression quality:

- **95% quality** → Minor recursion appears; thought remains largely coherent.
- **75–50% quality** → Thought becomes categorical and rigid (structured Q&A format).
- **25–10% quality** → Collapse into paranoia, existential despair, and self-referential loops.
- **5% quality** → Fixation on violence, extreme recursion, and aggressive paranoia.
- **1% quality** → Emergence of Zen-like paradoxes—profound yet disconnected from original context.

Crucially, degradation is **not random**. The model repeatedly enters the same attractor basins regardless of prompt, demonstrating that language possesses a low-dimensional geometric skeleton under finite constraint.

5. Security Implications: Embedding Corruption as a Stealth Attack Vector

The same technique that reveals attractors also constitutes a powerful, undetectable attack:

1. **Military & Geopolitical Manipulation** – Neutral situations can be made to appear highly aggressive (or vice versa).
2. **Public Opinion Manipulation** – Recommendation and summarization engines can be nudged toward fear or polarization.
3. **Corporate Espionage** – Investment-risk or strategy AIs can be quietly biased.
4. **Law-Enforcement & Surveillance** – Fraud detection or predictive policing can be silently compromised.

Traditional defenses (prompt filtering, weight monitoring, adversarial training) are bypassed because the attack occurs **inside the embedding space**.

6. Why This Vulnerability Remained Hidden

Most AI security focuses on tokens, prompts, or weights. Few researchers examine the geometric integrity of the embedding layer itself. This experiment proves that embedding distortion is both feasible and potent.

7. Recommended Countermeasures

- Cryptographic hashing or integrity verification of embeddings before transformer entry.
- Anomaly detection in embedding-space trajectories.
- Redundant multi-encoder consistency checks.
- Formal documentation of "embedding corruption" as a new attack class.

8. Discussion

This work provides the first empirical demonstration (within the Geofinitism framework) that LLM cognition follows finite geometric dynamics. The JPEG lens acts as a measurement tool that forces the system to reveal its own attractor structure—exactly as predicted by Finite Mechanics.

Simultaneously, it highlights an urgent real-world risk that demands immediate attention from AI security teams, red teams, and policymakers.

9. Conclusion and Future Work

By applying a simple, finite geometric operation (JPEG) to embeddings, we have:

1. Confirmed the geometric nature of linguistic meaning.
2. Exposed a new class of invisible AI attack.

Future directions include:

- Extending the technique to other modalities and larger models.
- Mapping the full attractor landscape of language.
- Developing and open-sourcing embedding-integrity toolkits.
- Preparing a formal security advisory for the AI research community.

The original chat logs, prompts, and raw outputs are available at FiniteMechanics.com for replication.

Acknowledgments

This research was conducted under the Geofinitism program. All experiments were performed locally.

References

[1] Haylett, K.R. (2025–2026). Geofinitism series, FiniteMechanics.com & Substack.

[2] JPEG standard (ISO/IEC 10918).

[3] Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.