Emergent Capabilities as Universal Latent Modes

Tamás Nagy, Ph.D.

tnagyphd@gmail.com

Draft (Outline + Preliminary Data)

Abstract

Emergent capabilities in large language models — abilities that appear suddenly as model scale increases — remain poorly understood. We propose that emergence corresponds to the activation of new **universal Latent modes**: directions in the model's representation space that are shared across all domains. Using the Latent of Latents framework (Nagy 2026h), we decompose a model's knowledge into domain-specific variation (captured by centered meta-axes) and universal variation (captured by the non-centered mean component). Each new universal mode represents a cross-domain capability — summarization, reasoning, analogy — that the model has learned to apply uniformly.

Preliminary evidence from GPT-2 Small (1 universal mode, 99.6% of universal variance) and TinyLlama 1.1B (2–3 universal modes, 67% + 3.5% + 0.8%) shows that larger models develop more universal modes, consistent with the emergence of new cross-domain capabilities at scale. We propose a tri-graded Latent algebra $\Lambda^{(i,j,k)}$ where i = within-domain, j = across-domain, k = universal depth, and predict that the number of universal modes at threshold ε scales as $K_\varepsilon \sim \log\log P$ where P is parameter count.

---

1. Introduction

1.1 The Emergence Puzzle

"Emergence" in LLMs refers to capabilities that are absent in smaller models but appear — seemingly discontinuously — in larger ones (Wei et al. 2022). Chain-of-thought reasoning, multi-step arithmetic, and analogical transfer all exhibit this pattern on BIG-Bench and related benchmarks. Schaeffer et al. (2024) argued that emergence may be a measurement artifact of nonlinear metrics, but subsequent work on representation geometry suggests real structural transitions underlie capability jumps.

Recent studies approach emergence from the geometry of learned representations. Li et al. (2025) identified three spectral phases during *training* — warmup, entropy-seeking, compression-seeking — using effective rank and eigenspectrum decay, showing that representation geometry changes non-monotonically and precedes capability gains. Nakagi et al. (2025) found a parallel triple phase transition through brain-alignment scores. Sun & Haghighat (2025) reformulated the Transformer as an $O(N)$ model and showed two phase transitions — one in generation temperature, one in parameter count — with the latter signaling emergence of new capabilities. Polo et al. (2026) demonstrated that chain-of-thought reasoning manifests as a "transient geometric pulse" where concept manifolds become linearly separable immediately before computation, then rapidly compress afterward.

These works establish that emergence has geometric signatures. Two questions remain open:

1. **Where** in the model's representation space do emergent capabilities live?
2. **Why** do they appear suddenly rather than gradually?

No existing work answers both. Geometric phase studies (Li et al., Nakagi et al.) track *global* spectral statistics without decomposing *which* capabilities emerge. The $O(N)$ model (Sun & Haghighat) provides a physics framework but not a constructive decomposition. Manifold separability (Polo et al.) operates at the token/reasoning level, not the knowledge-organization level.

1.2 The Universal Mode Hypothesis

We propose a specific answer: emergent capabilities live in the **universal component** of the model's knowledge — the part that is shared across ALL domains, not specific to any one.

In the Latent of Latents framework (Nagy 2026h), each domain θ has a mean representation $\bar{\Lambda}_\theta \in \mathbb{R}^d$. Stack these into a matrix $M \in \mathbb{R}^{D \times d}$. The standard analysis centers this matrix ($M_c = M - \bar{M}$) and performs SVD to find meta-axes — the directions along which domains DIFFER.

But the **mean** $\bar{M} = \frac{1}{D}\sum_\theta \bar{\Lambda}_\theta$ carries information too. It represents what the model knows UNIVERSALLY — structure shared by ALL domains. A model with richer universal knowledge has a more structured $\bar{M}$.

Key insight: Compare the singular values of the full (non-centered) matrix M with the centered matrix M_c:

$$\delta_k = \sigma_k(M)^2 - \sigma_k(M_c)^2$$

The excess δ_k is the variance in the k-th direction that comes from the universal component, not from domain-specific variation. Each positive δ_k is a **universal mode** — a cross-domain capability direction.

1.3 Preliminary Evidence

From the Latent of Latents experiments (Nagy 2026h):

Model	Params	K universal modes	Mode structure
GPT-2 Small	124M	1	99.6% in mode 1 (basic language model)
TinyLlama 1.1B	1,100M	2–3	67% + 3.5% + 0.8%

GPT-2 has essentially ONE universal mode — basic language modeling capability. TinyLlama, with 9× more parameters and 75× more training data, has developed 2–3 distinct universal modes. This is consistent with the emergence of new cross-domain capabilities at scale.

1.4 Relation to Feature Universality

Independent evidence for universal structure comes from sparse autoencoder (SAE) research. Lan et al. (2024) demonstrated "Analogous Feature Universality": SAE feature *spaces* across different LLMs are similar under rotation-invariant transformations. Thasarathan et al. (2025) extended this

with Universal SAEs (USAEs) that jointly learn a shared concept space across multiple models, discovering semantically coherent universal concepts from low-level features to high-level structures.

These results confirm that universal structure exists across models but do not explain *why* — they lack a generative theory. The Latent framework (Nagy 2026e), with its machine-verified finite representation theorem, provides the missing theoretical grounding: the low-rank structure is not accidental but a consequence of analyticity ($\rho > 1$). Universal modes are the specific directions that survive when domain-specific variation is removed.

Furthermore, the Latent of Latents (Nagy 2026h) established that domain meta-axes and SAE features capture the same geometry ($r = 0.914$ same-layer correlation, Procrustes disparity 0.088), confirming that the two decompositions — top-down (our meta-axes) and bottom-up (SAE features) — converge.

1.5 Contributions (Planned)

1. **The tri-graded Latent algebra** $\Lambda^{(i,j,k)}$ with k = universal depth — a theoretical framework for what existing empirical geometry studies observe.
2. **Universal mode detection algorithm** based on full vs centered SVD comparison.
3. **Scaling experiments** across a model-size ladder (GPT-2 Small → Medium → Large → XL, TinyLlama, LLaMA-2 7B).
4. **Correlation with BIG-Bench emergence**: do the number of universal modes predict which emergent benchmarks the model passes?
5. **Phase transition theory**: when does a new universal mode crystallize? Connecting to Sun & Haghighat's $O(N)$ framework.
6. **Layer-depth profile of universal modes**: extending the three-phase "crystallize-differentiate-compress" pattern (Nagy 2026h, Result 14) to the universal component — complementing Li et al.'s three phases across training *time* with three phases across model *depth*.

---

2. Theory

2.1 The Tri-Graded Algebra

We extend the bi-graded Latent of Latents $\Lambda^{(i,j)}$ to a tri-graded structure:

$$\Lambda^{(i,j,k)} \in \mathcal{H}^{(i)} \otimes \mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}$$

where: - i = **within-domain grade** (rank r): how complex each domain's internal structure is - j = **across-domain grade** (meta-rank R): how many axes differentiate domains - k = **universal grade** (universal rank K): how many cross-domain capabilities exist

The factorization becomes:

$$\Lambda_\theta \approx \underbrace{\sum_{k=1}^{K} u_k \cdot \bar{v}_k}_{\text{universal}} + \underbrace{\sum_{j=1}^{R} c_j(\theta) \cdot v_j}_{\text{domain-specific}}$$

where $\bar{v}_k$ are universal mode directions and v_j are domain-differentiating meta-directions.

2.2 Universal Mode Detection

Algorithm: 1. Extract domain means $\bar{\Lambda}_\theta$ for D domains, forming $M \in \mathbb{R}^{D\times d}$. 2. Compute full SVD: $M = U_f \Sigma_f V_f^\top$, singular values $\sigma_k^{(\text{full})}$. 3. Center and compute: $M_c = U_c \Sigma_c V_c^\top$, singular values $\sigma_k^{(\text{cent})}$. 4. Universal excess: $\delta_k = [\sigma_k^{(\text{full})}]^2 - [\sigma_k^{(\text{cent})}]^2$. 5. Universal mode count: $K_\varepsilon = \#\{k : \delta_k / \sum_j \delta_j > \varepsilon\}$.

2.3 Predictions

P1 (Mode count scales with model size): $K_\varepsilon(P) \leq K_\varepsilon(P')$ for $P < P'$ (same architecture family, same training data).

P2 (Mode crystallization): New universal modes appear at critical parameter counts, not gradually. Between mode transitions, $\delta_{K+1} \approx 0$.

P3 (Emergence = mode activation): A model passes an "emergent" benchmark when it has developed the corresponding universal mode. Models with K universal modes pass $\Theta(K)$ emergent benchmarks.

P4 (Unification at depth): Very large models may exhibit mode MERGER — two previously distinct universal modes combining into one richer mode. This would manifest as a decrease in K at very large scale (the "Unification Pulse" hypothesis from Nagy 2026h).

---

3. Experimental Plan

3.1 Model Ladder

Model	Params	d	Source	Status
GPT-2 Small	124M	768	OpenAI	Done
GPT-2 Medium	355M	1,024	OpenAI	Done
GPT-2 Large	774M	1,280	OpenAI	Planned (16GB laptop OK)
GPT-2 XL	1,558M	1,600	OpenAI	Planned (tight on 16GB)
TinyLlama 1.1B	1,100M	2,048	Zhang et al.	Done
LLaMA-2 7B	6,738M	4,096	Meta	Planned (needs GPU or quantized)
Mistral 7B	7,242M	4,096	Mistral AI	Planned

3.2 Measurements Per Model

For each model: 1. 41-domain extraction (same prompts as Nagy 2026h for comparability) 2. Full and centered SVD → universal excess δ_k 3. Universal mode count K at thresholds 1%, 5%,

10% 4. Layer-by-layer universal profile (at which layer do universal modes appear?) 5. SAE cross-validation at peak-differentiation layer

3.3 Emergence Correlation

Compare universal mode count with performance on known emergent benchmarks: - BIG-Bench Hard (23 tasks): which tasks does each model pass? - MMLU categories: does performance correlate with mode structure? - Arithmetic accuracy by digit count: phase transition location

3.4 Required Resources

- GPT-2 Large/XL: 16GB laptop, ~20 min each
- LLaMA-2 7B (4-bit quantized): 16GB laptop or free Colab GPU, ~1 hour
- BIG-Bench eval: public evaluation harness, ~2 hours per model
- Total: 1–2 days of laptop computation

---

4. Preliminary Results

4.1 GPT-2 Small (from Nagy 2026h)

Universal excess analysis (41 domains):

```
= 99.6%   (basic language model — one dominant universal mode)
= 0.3%
= 0.05%
```

Effectively $K = 1$. GPT-2 Small has ONE universal capability: language modeling. Its emergent benchmark performance is accordingly limited.

4.2 TinyLlama 1.1B (from Nagy 2026h)

```
= 67.0%   (primary universal mode)
= 3.5%    (second universal mode — possible reasoning/instruction following)
= 0.8%    (third mode — weak but present)
```

$K = 2$–3. TinyLlama has developed at least one additional universal mode beyond basic language modeling. This aligns with its known ability to follow instructions and perform simple reasoning — capabilities that GPT-2 Small lacks.

4.3 Layer Profile Connection

From the layer-by-layer analysis (Nagy 2026h, Result 14), the "crystallize-differentiate-compress" pattern at the domain level suggests that universal modes may follow a similar trajectory: crystallizing at early layers, differentiating in the middle, and integrating at the final layers.

---

5. Related Work

5.1 Geometric Phases in Training

Li et al. (2025) discovered three spectral phases during pretraining using OLMo (1B–7B) and Pythia (160M–12B): (1) warmup with representational collapse, (2) entropy-seeking with dimensionality expansion, (3) compression-seeking with anisotropic consolidation. Our Result 14 (Nagy 2026h) finds an analogous three-phase pattern — crystallization, differentiation, compression — but across *layer depth* in a trained model, not across training time. The two are complementary: Li et al. show *when* structure forms during training; we show *where* it lives in the final model. An open question is whether the training-time phases map onto the depth phases, i.e., whether early training corresponds to early layers.

5.2 Physics of Phase Transitions

Sun & Haghighat (2025) reformulate the Transformer as an $O(N)$ model, finding a "higher-depth" phase transition at a critical parameter count that signals emergence. Our prediction P2 (mode crystallization) is compatible: we propose that the phase transition is specifically the activation of a new universal mode δ_{K+1} crossing the noise floor. The $O(N)$ framework provides the statistical mechanics; the tri-graded Latent algebra provides the algebraic structure of *what* transitions.

5.3 Dynamic Manifold Management

Polo et al. (2026) show that reasoning manifests as a transient geometric pulse — manifold separability spikes then compresses within a single forward pass. Our differentiation phase (L3–L11) may be the static analog: the model maintains an expanded representational palette in its middle layers, which the final layer compresses. Polo et al.'s "Dynamic Manifold Management" operates per-token within a sequence; our meta-rank profile operates per-domain across the knowledge space. Together, they suggest the transformer manages bandwidth at multiple scales simultaneously.

5.4 Feature Universality Across Models

Lan et al. (2024) established that SAE feature spaces are similar across different LLMs under rotation-invariant transformations ("Analogous Feature Universality"). Thasarathan et al. (2025) extended this with Universal SAEs that jointly learn cross-model concept spaces, finding semantically coherent universal concepts from low-level to high-level.

Our universal modes are the *theoretical prediction* for what these studies observe empirically. If the Latent framework is correct, feature universality follows from analyticity: any smooth system admits a finite representation (the Latent), and because language itself has shared structure, different models trained on similar data must converge to similar Latents. The meta-axes and universal modes are the principal components of this shared structure.

The key difference: Lan et al. and Thasarathan et al. demonstrate universality but cannot predict it or bound its dimensionality. The Latent framework's $N = \Theta(\log(1/\varepsilon)/\log\rho)$ bound provides both.

5.5 The Mirage Debate

Schaeffer et al. (2024) argued that emergent abilities are measurement artifacts of nonlinear metrics. Our framework offers a resolution: if universal modes crystallize via genuine phase transitions (P2),

then *both* sides are partially correct. Smooth metrics will show continuous improvement in the sub-threshold regime (δ_{K+1} growing smoothly), but a genuine discontinuity occurs when δ_{K+1} crosses the activation threshold — the model suddenly *has* a new capability that it previously lacked. The "mirage" is that some metrics detect the threshold crossing while others detect the continuous buildup.

6. Discussion

6.1 Why Emergence Is Sudden

If universal modes crystallize via a phase transition (P2), then emergence is sudden because the underlying process is a symmetry-breaking event in representation space. Below the critical model size, all domains contribute to a single undifferentiated universal mode. Above it, the mode splits into two — and the model can suddenly do things it could not before.

This is analogous to phase transitions in physics: water doesn't gradually become ice. At 0°C, the symmetry breaks and a new ordered phase appears. Universal modes may behave similarly, with parameter count playing the role of temperature. Sun & Haghighat's (2025) $O(N)$ framework provides independent support for this analogy.

6.2 Connection to Anthropic's Features

If each universal mode corresponds to a cluster of SAE features that activate across all domains, then we can: 1. Identify which SAE features constitute each universal mode (via the Neuronpedia cross-validation established in Nagy 2026h) 2. Name the mode based on feature labels (e.g., "reasoning mode" = cluster of "logical connector", "if-then", "because" features) 3. Steer model behavior by activating/suppressing specific universal modes 4. Transfer steering vectors across models via the universal mode alignment (following Lan et al.'s rotation-invariant transformations)

6.3 The Unification Pulse at Scale

At very large scale (100B+ parameters), we predict that some universal modes will MERGE — the model discovers that two previously distinct capabilities are manifestations of the same underlying structure. This is the k-grade contraction described in the tri-graded algebra. If observed, it would be strong evidence that very large models are developing something analogous to unified theories in physics.

This connects to Li et al.'s (2025) "compression-seeking" phase: at very large scale, the model may enter a *second* compression-seeking phase where universal modes themselves are compressed. The training-time three-phase pattern may repeat at the capability level.

---

7. Next Steps

Phase 1: Complete the Model Ladder

1. Run GPT-2 Large and XL (requires downloading ~3GB models, ~20 min each)
2. Run quantized LLaMA-2 7B (4-bit GPTQ or GGUF, ~1 hour)
3. Compute universal excess δ_k for each model → scaling curve K vs P

Phase 2: Emergence Correlation

4. Evaluate all models on BIG-Bench Hard (23 tasks) + MMLU
5. Test P3: does K predict the number of passed emergent benchmarks?
6. Test P2: does K vs P show discrete jumps (phase transitions)?

Phase 3: Layer-Depth Universal Profile

7. Compute universal excess at every layer for each model
8. Compare with Li et al.'s training-time phases: does layer depth recapitulate training time?
9. Test whether universal modes crystallize at the same layer as domain modes (L3 in GPT-2)

Phase 4: Cross-Model Alignment

10. Use Lan et al.'s rotation-invariant similarity to test whether universal modes are consistent across model families (not just within GPT-2)
11. If universal modes align cross-model → steering vector transfer experiments

Phase 5: Formalization

12. If K vs P shows phase transitions → formalize the critical threshold in Lean 4
13. Connect to Sun & Haghighat's $O(N)$ critical point: is there a mapping between their order parameter and δ_{K+1}?

During the preparation of this work the author used large language models in order to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- Wei, J. et al (2022). Emergent abilities of large language models. *TMLR*.
- Schaeffer, R. et al (2024). Are emergent abilities of LLMs a mirage? *NeurIPS 2024*. NeurIPS 2024*.
- Li, M.Z. et al (2025). Tracing the representation geometry of language models from pretraining to post-training. *arXiv:2509.23024*.
- Nakagi, Y. et al (2025). Triple phase transitions: Understanding the learning dynamics of large language models from a neuroscience perspective. *arXiv:2502.20779*.
- Sun, Y. & Haghighat, B (2025). Phase transitions in large language models and the $O(N)$ model. *arXiv:2501.16241*.
- Polo, A. et al (2026). Emergent manifold separability during reasoning in large language models. *arXiv:2602.20338*.
- Lan, M. et al (2024). Quantifying feature space universality across large language models via sparse autoencoders. *arXiv:2410.06981*.

- Thasarathan, H. et al (2025). Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv:2502.03714*.
- Bricken, T. et al (2023). Towards Monosemanticity. *Anthropic Research*.
- Bloom, J. et al (2024). Open-source SAE training and analysis for GPT-2. *GitHub/SAE Lens*.
- Nagy, T. (2026). The Latent: Finite Sufficient Representations of Smooth Systems. *Zenodo*. DOI: 10.5281/zenodo.19101209
- Nagy, T. (2026). The Latent of Latents: Hierarchical Finite Representations of Knowledge Families. *Zenodo*. DOI: 10.5281/zenodo.19134434
- Radford, A. et al (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.
- Zhang, P. et al (2024). TinyLlama: An open-source small language model. *arXiv:2401.02385*.
- Vaswani, A. et al (2017). Attention is all you need. *NeurIPS*. DOI: 10.65215/2q58a426