Suffix Length	Frequency (%)
≤ 3	25.41
4	46.61
5	16.71
6	6.08
7	2.40
8	1.46
> 8	1.34

Table 1: **∞-gram context in TriviaQA.** Suffix lengths reveal focus on short- to mid-context statistics.

Model	Original	Top-10		Top-50	
		Removed	Retained	Removed	Retained
Pythia-1B	0.838	0.849	0.225	0.835	0.318
Pythia-1.4B	0.866	0.855	0.232	0.859	0.324
Pythia-2.8B	0.884	0.880	0.219	0.873	0.317
Pythia-6.9B	0.896	0.893	0.202	0.906	0.327
OLMo-2-1B	0.953	0.943	0.199	0.954	0.326
OLMo-2-7B	0.970	0.966	0.155	0.970	0.308

Table 2: **Full-spectrum information is required.** Retaining only top eigen-directions markedly degrades SciQ accuracy.

characteristic of echolalia in early checkpoints (Appendix Fig. 7). This relatively short phase is followed by an “entropy-seeking” phase characterized by a manifold expansion in several directions, and then a “compression-seeking” phase that imposes an anisotropic consolidation of the representation space along its principal eigenvectors. We observe these phases in both OLMo2 and Pythia family of models across different model sizes, indicating the consistent nature of non-monotonic changes in representation geometry during pretraining. It is worth noting that there could be emergence of additional “entropy-seeking” and “compression-seeking” with more pretraining, as in later stages of OLMo-2 7B model pretraining (c.f. Figure 2C). Notably, these phases persist even in smaller models below 1B parameters (Appendix Fig. 10), demonstrating the fundamental nature of this geometric evolution. Furthermore, as shown in Figure 3, these three-phase dynamics are consistently observed across intermediate layers throughout the network depth, confirming that the geometric evolution is not confined to the final representations but reflects a global transformation of the model’s representational structure.

Key takeaway. Despite near-monotonic loss, representation geometry exhibits a consistent, non-monotonic phase sequence (“warmup” ; “entropy-seeking” ; “compression-seeking”). These trends are stable across: (i) sample count M and sequence length L; (ii) dataset choice within family; and (iii) layers (with last-layer sufficing for tracking), for both OLMo and Pythia at 1B+ scale.

3.2. Memorization & beyond: Distributional memorization happens in entropy-seeking phase

In this section, we seek to associate the different geometric phases to specific LLM behaviors. Downstream tasks that test the LLM’s factual reasoning and language understanding abilities seem to improve with more pretraining. However, it is unclear to what extent this increase is due to an improvement in the model’s memorization ability, i.e. how good is the model in “regurgitating” short-context phrases from the pretraining dataset, as opposed to a general language understanding, i.e. leveraging long-context dependencies to generate reasonable output. We disentangle these two factors by using the distributional memorization metric (Wang et al., 2025) presented in eq. (2) for Pythia models when processing sequences from the TriviaQA dataset (Joshi et al., 2017). Notably, the ∞-gram model predominantly utilizes short- to medium-length suffixes (Table 1), making it an ideal baseline for measuring short-context memorization capabilities.

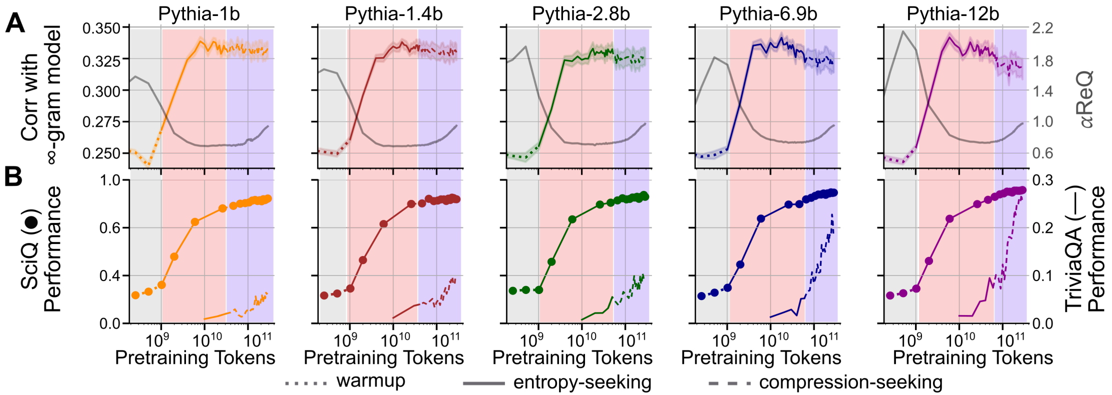

Figure 4: Distinct learning phases are linked to different LLM capabilities. (A) Memorization metric, i.e. spearman correlation between LLM and ∞-gram outputs, and representation geometry metric, α_{ReQ}, across Pythia models' (1–12B parameters) pretraining. Memorization peaks late in the "entropy-seeking" phase before plateauing or degrading slightly in the "compression-seeking" phase, suggesting that the former prioritizes capturing short-context n-gram statistics. **(B)** 0-shot performance on multiple-choice (SciQ) and factual question-answering (TriviaQA) tasks across pretraining. While accuracy on SciQ benefits from learning in both phases, accuracy on TriviaQA *groks* once the model learns long-context statistics, primarily in the "compression-seeking" phase.

Key takeaway. "entropy-seeking" expands utilized dimensions (RankMe ↑, α_{ReQ} ↓), aligning with increased alignment to ∞-gram statistics (distributional memorization). In contrast, during "compression-seeking" , information is anisotropically concentrated (RankMe ↓, α_{ReQ} ↑) and long-context QA accuracy continues to improve even as memorization saturates. Together with the cross-model SciQ correlations (see Appendix Table 9), this dissociates short-context memorization from long-context generalization and links them to distinct spectral regimes.

Figure 4 illustrates the memorization metric and task performance over the course of pretraining for Pythia models of 5 different sizes – ranging from 1B to 12B. Across all models, the distributional memorization metric increased during the "entropy-seeking" phase and peaked towards the end of this phase. Intuitively, this result suggests that the "entropy-seeking" phase is particularly important for learning short-context statistics, e.g. high-frequency n-grams, present in the pretraining corpus. This intuition is also supported by findings of Wang et al., c.f. Fig 12 (Wang et al., 2025). Following this peak in the memorization metric, it plateaued (or slightly decreased) during the "compression-seeking" phase, suggesting that the model's output in this phase is guided by factors beyond n-gram statistics. Notably, the 0-shot accuracy on multiple-choice question-answering tasks, e.g. SciQ (Welbl et al., 2017), consistently improved throughout both the "entropy-seeking" and "compression-seeking" phases, potentially benefiting from both short- and long-context information learned in the respective phases.

However, 0-shot performance on factual question-answering tasks, e.g. TriviaQA (Joshi et al., 2017), demonstrate a sudden and dramatic rise in accuracy closely aligned with the saturation of the memorization metric. Consequently, most of the improvement in task accuracy happens during the "compression-seeking" phase, potentially benefiting from the long-context statistics learned in this phase, which are crucial for this task. Taken together, these findings outline a distinct association between each phase and the emergence of different LLM capabilities: short-context n-gram modeling during the "entropy-seeking" phase and long-context information aggregation during the "compression-seeking" phase.

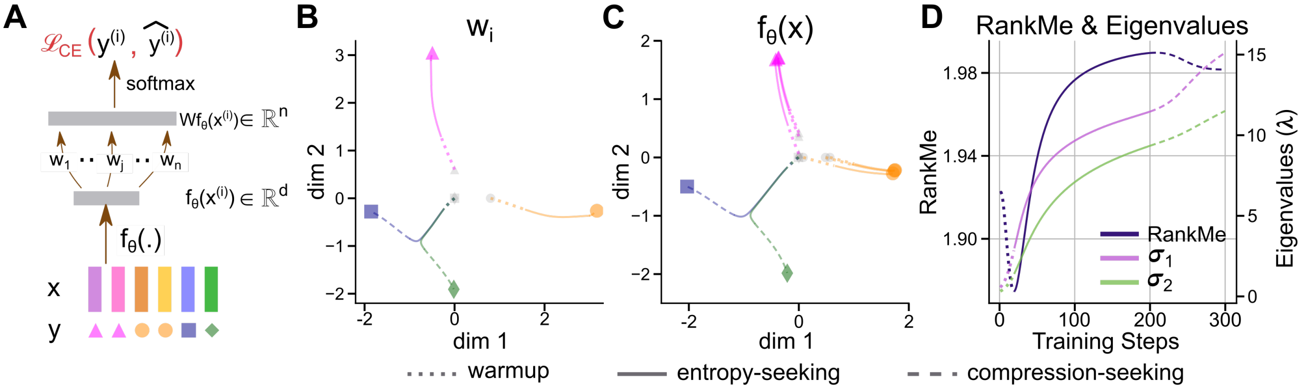

Figure 5: **Learning dynamics of cross-entropy loss replicate multiphase learning dynamics.** **(A)** Schematic of a model with feature extractor $f_\theta(\in \mathbb{R}^d)$, linear classifier $W(\in \mathbb{R}^{n\times d})$ and cross-entropy loss $\mathcal{L}_{CE}$. Skewed class distribution and information bottleneck ($d < n$) are critical to replicate all three phases observed in LLM pretraining. **(B, C)** Classifier weights (W_i) and feature representations ($f_\theta(x)$) demonstrate distinctive trajectories analogous to "warmup" (dotted), "entropy-seeking" (solid), and "compression-seeking" (dashed) phases. **(D)** Quantitative spectral metrics RankMe and eigenvalues, σ_1, σ_2.

3.3. Role of learning objective and optimization in learning dynamics

Having demonstrated the existence and salience of distinct learning phases, we now seek to understand the role of loss and optimization frameworks used in LLM pretraining in engendering these phases. Specifically, we studied the gradient descent dynamics while optimizing the cross-entropy loss in an analytically-tractable setting — the model $\mathrm{f}_\theta(x)$ is linear, i.e. $\mathrm{f}_\theta(x) = \theta x \in \mathcal{R}^d$, and logits are obtained (like in LLM models) as $z = W\mathrm{f}_\theta(x) = W\theta x \in \mathcal{R}^{|\mathcal{V}|}$. The outputs are obtained by applying a softmax operation on z (see Figure 5A). We extended the results of Pezeshki et al. (2021) to study how W and $\mathrm{f}_\theta(.)$ change when optimizing the loss using gradient descent. Notably, we found two key properties of gradient descent that contribute to the emergent geometric properties of the representation space (Appendix §B for formal statements):

- **Primacy bias**: Representations and weights corresponding to high-frequency tokens are learned earlier in training, compared to low-frequency tokens.
- **Selection bias**: Dominant directions in the representation space are more likely to be used for encoding new information, i.e. $\Delta\sigma_i \propto \sigma_i$

We demonstrate (c.f. Figure 5) that two conditions are necessary (see supplementary for controls) for replicating the multiphase learning dynamics in our toy-model, as observed within LLMs: (1) non-uniform class distribution, i.e. some tokens (or classes) occur more frequently than others in the training data, and (2) information bottleneck, i.e. number of feature dimensions (d) is less than the vocabulary size ($|\mathcal{V}|$). Note that these two conditions are common in LLM pretraining setups.

In the analytically tractable setup that satisfies the above conditions, we found that $\mathrm{f}_\theta(.)$ and W for frequently-occurring classes are separated during the initial "warmup" phase (Figure 5B & C, dotted lines). The corresponding eigenvectors of the weight and feature spaces also become aligned during this phase. Following this initial eigenvector-alignment phase, there is an overall expansion in the representation space that leads to higher confidence predictions for frequently-occurring classes. This phase of volume expansion in the $\mathrm{f}_\theta(.)$ and W spaces is associated with an increasing effective rank, akin to the "entropy-seeking" phase (Figure 5B & C, solid lines). Following this phase, the infrequently-