Memorization-Compression Cycles Improve Generalization

Fangyuan Yu
Thoughtworks
fangyuan.yu@thoughtworks.com

Abstract

We prove theoretically that generalization improves not only through data scaling but also by compressing internal representations. To operationalize this insight, we introduce the Information Bottleneck Language Modeling (IBLM) objective, which reframes language modeling as a constrained optimization problem: minimizing representation entropy subject to optimal prediction performance. Empirically, we observe an emergent memorization–compression cycle during LLM pretraining, evidenced by oscillating positive/negative gradient alignment between cross-entropy and Matrix-Based Entropy (MBE), a measure for representation entropy. This pattern closely mirrors the predictive–compressive trade-off prescribed by IBLM and also parallels the biological alternation between active learning and sleep consolidation. Motivated by this observation, we propose Gated Phase Transition (GAPT), a training algorithm that adaptively switches between memorization and compression phases. When applied to GPT-2 pretraining on FineWeb dataset, GAPT reduces MBE by 50% and improves cross-entropy by 4.8%. GAPT improves OOD generalization by 35% in a pretraining task on arithmetic multiplication. In a setting designed to simulate catastrophic forgetting, GAPT reduces interference by compressing and separating representations, achieving a 97% improvement in separation — paralleling the functional role of sleep consolidation in biological learning process.

1 Introduction

Generalization occurs when learning from one task improves performance on another. The pursuit of generalization in pre-training LLM (11) has historically focused on scaling up data and parameter size (4), in post-training, Reinforcement Learning with Verifiable Reward (RLVR) (7) has gained attention. However, high quality data has been exhausted after years of tireless scraping, and RLVR is shown to only trims knowledge from baseline model (10) instead of 'incentivize new reasoning pattern'.

We establish a theoretical upper bound on generalization error for deep learning models, indicating representation entropy as another dimension for improving generalization besides scaling data.

Theorem 1. ***Upper Bound on Generalization Error.*** *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables with an unknown joint distribution $P(X, Y)$, and suppose X is discrete with finite cardinality. Let f be a neural network with L intermediate representations forming a Markov chain:*

$$X \to R_1 \to \cdots \to R_L \to \hat{Y}$$

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Non-Euclidean Foundation Models and Geometric Learning.

Where R_l is internal representations, $\hat{Y}$ is the prediction of the network. Then, for any dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ sampled i.i.d. from $P(X,Y)$, the generalization error satisfies for $\alpha \in [1, +\infty)$:

$$\underbrace{\mathcal{L}_{P(X,Y)}(f,\ell)}_{\textit{Generalization Error}} \leq \underbrace{\hat{\mathcal{L}}_{P(X,Y)}(f,\ell,\mathcal{D}_N)}_{\textit{Empirical Error}} + \underbrace{\mathcal{O}\left(\log N \cdot \min_{1\leq l\leq L} 2^{\alpha\cdot H(R_l)} \,/\, \sqrt{N}\right)}_{\textit{Upper Bound on Generalization Gap}} \tag{1}$$

Intuitively, generalization performance of neural network can be improved by either increasing training data size, or by reducing entropy of its internal representations. We refer minimization of $H(R_l)$ as compression, and minimizing empirical loss as memorization.

The Information Bottleneck (IB) framework (8) characterizes the optimal representation as one that discards as much information from the input as possible, while preserving all information relevant to the output. Motivated by this, we introduce the Information Bottleneck Language Modeling (IBLM) objective, along with a theorem showing equivalence of IBLM with IB framework under language modeling.

Definition 1. ***Information Bottleneck Language Modeling (IBLM).*** *Given a language model with internal representations $R_{1:L}$ and output token variable Y, the IBLM objective is:*

$$\begin{aligned} \min \quad & \sum_{l=1}^{L} H(R_l) \\ \textit{s.t.} \quad & \hat{Y} = \arg\min_{\hat{Y}} H(Y|\hat{Y}) \end{aligned} \tag{2}$$

Theorem 2. *The IBLM objective defined in Equation 2 is equivalent to the classical Information Bottleneck formulation under the language modeling setting.*

Note that $H(Y|\hat{Y})$ is the cross-entropy (CE) loss in conventional language modeling task (11). In short, LBLM requires maximal compressing internal representation under optimal cross-entropy. To explicitly calculate $H(R)$, we adopt Matrix-based entropy (MBE), first proposed in (5), given a matrix $R \in \mathbb{R}^{s\times d}$ concatenating s token representations, MBE is given by

$$S_\alpha(T) = \frac{1}{1-\alpha}\log\Big(\sum_i (\frac{\lambda_i(K)}{\mathrm{Tr}(K)})^\alpha\Big)$$

where $K = RR^T$ is the Gram matrix. MBE essentially provide a continuous measure of matrix rank, by calculating entropy on distribution of singular values. It has been shown to exhibit a strong correlation with embedding quality (14).

Previous work (15) showed that in neural networks, training often follows a two-phase trend: an initial memorization phase with rapid loss reduction, followed by a monotonic decrease in representation entropy $H(R)$, interpreted as compression. This suggests a clean, sequential separation between fitting and compressing.

Our analysis of GPT pretraining reveals a richer dynamic: the cosine similarity between CE and MBE gradients oscillates between positive and negative, indicating a cyclic alternation between memorization (representation expansion) and compression. Rather than progressing through distinct phases, learning continuously revisits these states. Notably, this local oscillation coexists with a global decline in MBE, suggesting that compression accumulates over time even as the model periodically re-engages in memorization. We term this the memorization–compression cycle.

In biological neural systems, learning proceeds cyclically, alternating between active acquisition and sleep-driven consolidation. A seminal study (18) showed that sleep mitigates interference between conflicting memories by compressing and reorganizing them—allocating distinct subsets of synaptic weights to each. When two conflicting tasks were presented sequentially, awake learning alone led to catastrophic forgetting, while sleep consolidation enabled the formation of separable representations, outperforming even interleaved training. Inspired by this, we propose Gated Phase Transition (GAPT), a training algorithm that alternates between memorization (minimizing cross-entropy) and compression (minimizing cross-entropy and Matrix-Based Entropy). GAPT dynamically switches phases based on learning signals, approximating the IBLM objective and functionally mirroring biological consolidation to resolve representational conflict.

GAPT delivers consistent improvements across domains. In LLM pretraining on the FineWeb dataset, it reduces Matrix-Based Entropy (MBE) by an average of 70.5% across layers while improving cross-entropy by 4.8%, outperforming standard language modeling and closely approximating the IBLM trade-off. In an arithmetic generalization task, GAPT reduces test cross-entropy by 35% and MBE by 47% when trained on 1–3 digit multiplication and evaluated on 4–6 digits, supporting Theorem \ref{thm:entropy-generalization}. Finally, in a synthetic setup with gradient interference, GAPT improves representational separation by 97% and reduces MBE by 91% relative to a mixed baseline—closely mirroring the conflict resolution behavior observed during biological sleep.

Our work makes several key contributions:

- **Theoretical Foundation**: We derive an upper-bound on generalization error showing that reducing representation entropy can improve generalization, alongside scaling data.
- **IBLM Objective**: We formulate Information Bottleneck Language Modeling (IBLM) as a constrained optimization problem, unifying representation entropy and cross entropy targets.
- **GAPT Algorithm**: We propose Gated Phase Transition (GAPT), a algorithm to solve IBLM alternates between memorization and compression phases based on dynamic training signals.
- **Empirical Results**: We show that GAPT improves LLM pre-training, arithmetic generalization, and conflict resolution.
- **Biological Analogy**: We relate the memorization–compression cycle in LLMs to the awake–sleep consolidation cycle in biological systems and validate compression's similarity to consolidation.

The remainder of the paper expands on these contributions: Section 2 presents our theoretical framework and generalization bound; Section 3 details the GAPT algorithm; Section 4 presents empirical results, including the memorization–compression cycle and GAPT's effectiveness; Section 5 discusses relevant work.

2 Theory

Corollary 1 (Entropy Lower Bound for Finite Discrete Random Variables)**.** *Let X be a discrete random variable with finite support Ω, where $|\Omega| = n$, and assume that $P(X = x) > 0$ for all $x \in \Omega$. Then there exists a constant $\beta \in (0, 1]$ such that:*

$$H(X) \geq \beta \cdot \log_2 |\Omega|.$$

Proof of Corollary 1 can be found in Appendix A.

Theorem 1. ***Upper Bound on Generalization Error.*** *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables with unknown joint distribution $P(X, Y)$, and suppose X is discrete with finite cardinality. Let f be a neural network with L intermediate representations forming a Markov chain:*

$$X \to R_1 \to \cdots \to R_L \to \hat{Y}$$

Then, for any dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$, there exists $\alpha \in [1, +\infty)$ s.t. the generalization error satisfies

$$\mathcal{L}_{P(X,Y)}(f, \ell) \leq \hat{\mathcal{L}}_{P(X,Y)}(f, \ell, \mathcal{D}_N) + \mathcal{O}\left(\frac{\log N \cdot \min_{1 \leq l \leq L} 2^{\alpha \cdot H(R_l)}}{\sqrt{N}}\right)$$

Proof. We begin by recalling the standard formulation of the generalization gap:

$$\text{Gen}_{P(X,Y)}(f, \ell, \mathcal{D}_N) := \mathcal{L}_{P(X,Y)}(f, \ell) - \hat{\mathcal{L}}_{P(X,Y)}(f, \ell, \mathcal{D}_N)$$

For discrete input variables X, it has been shown in (6) that the generalization gap is upper-bounded by:

$$\begin{aligned}\mathcal{L}_{P(X,Y)}(f, \ell) - \hat{\mathcal{L}}_{P(X,Y)}(f, \ell, \mathcal{D}_N) &\leq \mathcal{O}\left(\frac{|\mathcal{X}| \log N}{\sqrt{N}}\right) \\ &\leq \mathcal{O}(\frac{2^{\alpha \cdot H(X)} \log N}{\sqrt{N}})\end{aligned}$$

where $|\mathcal{X}|$ denotes the cardinality of the input space. Second line follows from Corollory 1, which states that we have $|\mathcal{X}| = \mathcal{O}(2^{\alpha \cdot H(X)})$ for $\alpha \in [1, +\infty)$. Let f be a neural network with L intermediate representations forming a Markov chain:

$$X \to R_1 \to \cdots \to R_L \to \hat{Y}$$

Fix any intermediate representation R_l, and decompose the network into $f = d \circ e$, where $e : \mathcal{X} \to \mathcal{R}_l$ maps inputs to $R_l = e(X)$, and $d : \mathcal{R}_l \to \mathcal{Y}$ predicts the output. Applying the generalization bound to the representation R_l, we obtain:

$$\mathcal{L}_{P(X,Y)}(f, \ell) = \mathcal{L}_{P(R_l,Y)}(d, \ell_e) \tag{3}$$

$$\leq \hat{\mathcal{L}}_{P(R_l,Y)}(d, \ell_e, \mathcal{D}_N) + \mathcal{O}\left(\frac{2^{\alpha \cdot H(R_l)} \log N}{\sqrt{N}}\right) \tag{4}$$

$$\leq \hat{\mathcal{L}}_{P(X,Y)}(f, \ell, \mathcal{D}_N) + \mathcal{O}\left(\frac{2^{\alpha \cdot H(R_l)} \log N}{\sqrt{N}}\right) \tag{5}$$

Since the Markov structure ensures that each R_l is a valid bottleneck in the information flow, we can take the tightest such bound across all layers, yielding:

$$\mathcal{L}_{P(X,Y)}(f, \ell) \leq \hat{\mathcal{L}}_{P(X,Y)}(f, \ell, \mathcal{D}_N) + \mathcal{O}\left(\frac{\log N \cdot \min_{1 \leq l \leq L} 2^{\alpha \cdot H(R_l)}}{\sqrt{N}}\right)$$

This concludes the proof. □

Theorem 2. *The Information Bottleneck Language Modeling (IBLM) objective defined in Equation 2 is equivalent to the classical Information Bottleneck formulation under the language modeling setting.*

Proof. The Information Bottleneck (IB) framework (2; 8) defines the optimal representation R as the solution to:

$$\min_{p(r|x)} \quad I(R; X) \tag{6}$$

$$\text{s.t.} \quad I(Y; R) = I(Y; X) \tag{7}$$

That is, R should discard as much input information as possible while preserving all information relevant to predicting Y—the minimal sufficient statistic. In large language models (LLMs), the network forms a deterministic Markov chain: $X \to R \to \hat{Y}$. Since R is deterministically computed from X, we have:

$$I(X; R) = H(R) \quad (\text{as } H(R|X) = 0) \tag{8}$$

$$I(Y; X) \geq I(Y; R) \geq I(Y; \hat{Y}) = H(Y) - H(Y|\hat{Y}) \tag{9}$$

Where second line follows from data processing inequality (3). Thus,

$$I(Y; X) - I(Y; R) \leq H(Y|\hat{Y}) - H(Y|X) \tag{10}$$

Since $H(Y|X)$ is fixed, minimizing $H(Y|\hat{Y})$—i.e., cross-entropy—maximizes $I(Y; \hat{Y})$, satisfying the IB predictive constraint. On the compression side, since $I(X; R) = H(R)$, minimizing representation entropy directly reduces the IB compression term. Hence, minimizing cross-entropy aligns with preserving predictive information, while minimizing entropy enforces compression—together forming the IBLM objective. □

3 Approach

While applying a Lagrangian objective (CE + λ·MBE) is a natural approach to solving the constrained optimization in IBLM, we find it often leads to representation collapse: MBE converges to near-zero, but CE worsens as the model loses structure in its internal representations.

Inspired by the alternation between learning and consolidation in biological systems, we divide training into two phases: memorization, where the model minimizes cross-entropy (CE) loss, and compression, where it minimizes a weighted sum of CE and Matrix-Based Entropy (MBE). We

propose Gated Phase Transition (GAPT), a training algorithm that dynamically alternates between these phases. GAPT tracks a global minimum CE loss and per-layer MBE histories, and uses patience-based gating to switch phases. Compression is exited early if CE degrades.

GAPT encourages localized compression—reorganizing existing knowledge without interfering with the acquisition of new information—and ensures that entropy reduction occurs only when it does not hinder memorization.

Algorithm 1 Gated Phase Transition (GAPT)

Input: losses $\mathcal{L}$, thresholds δ, τ, patience p_m, p_c
State: $\phi \in \{1, 2\}$ (1 = mem, 2 = comp), counters s_m, s_c, $E_{\min}$, $\text{MBE}_{\min}[i]$
Extract $\mathcal{L}_{\text{ce}}$, $\{\text{MBE}_i\}$; update $\Delta E \leftarrow E_{\min} - \mathcal{L}_{\text{ce}}$, $E_{\min} \leftarrow \min(E_{\min}, \mathcal{L}_{\text{ce}})$
if $\phi = 1$ **then** ▷ Memorization
    $s_m \leftarrow 0$ if $\Delta E > \delta$ else $s_m \mathrel{+}= 1$
    if $s_m \geq p_m$ **then**
        $\phi \leftarrow 2$, $s_c \leftarrow 0$, $E_{\min} \leftarrow \infty$, $\text{MBE}_{\min}[i] \leftarrow \infty$
    end if
else ▷ Compression
    if $\mathcal{L}_{\text{ce}} > E_{\min} \cdot (1 + \tau)$ **then**
        $\phi \leftarrow 1$, $s_m \leftarrow 0$
    else
        $\Delta M \leftarrow \max_i(\text{MBE}_{\min}[i] - \text{MBE}_i)$
        for each i **do**
            $\text{MBE}_{\min}[i] \leftarrow \min(\text{MBE}_{\min}[i], \text{MBE}_i)$
        end for
        $s_c \leftarrow 0$ if $\Delta M > \delta$ else $s_c \mathrel{+}= 1$
        if $s_c \geq p_c$ **then**
            $\phi \leftarrow 1$, $s_m \leftarrow 0$
        end if
    end if
end if

4 Experiments

4.1 Natural Compression-Memorization cycle

Our experimental setup follows the Modded-NanoGPT framework (23), We remove FP8 matmul (due to hardware incompatibility with Hopper GPUs) and use a simplified block causal attention mask. All experiments are conducted on 8×L40 GPUs, training a 12-layer GPT model on a 0.73B-token FineWeb training set, evaluated on its corresponding validation split. We train on CE loss only. We log cross-entropy (CE) and Matrix-Based Entropy (MBE) gradients at every training step. For each iteration, we record both CE and MBE gradients across multiple batches and compute the average cosine similarity both (1) across batches (CE vs. CE), and (2) between CE and MBE gradients, within and across batches.

In Figure 1, we observe that training with CE loss alone leads to a consistent decrease in MBE across layers, confirming observations from (14). We further analyze CE gradient behavior. Figure 2. shows that gradient consistency (measured by cosine similarity across batches) declines over time across all layers. This indicates an increasing signal-to-noise ratio in CE gradients as training progresses.

We next inspect cosine similarity between CE and MBE gradients. As shown in Figure 3, we observe recurring sign flips that indicate an alternation between memorization and compression phases, even without explicitly optimizing for entropy. To characterize this oscillation, we analyze the gradient signal across training using three measures: standard deviation (oscillation strength), zero-crossing rate (oscillation frequency), and peak-to-average power ratio from the power spectral density (periodicity strength). Figure 4 summarizes these metrics across layers and parameters. We find that attention parameters exhibit stronger and more frequent oscillations than MLP parameters. Interestingly, later layers show higher oscillation frequency, while no layer demonstrates strong

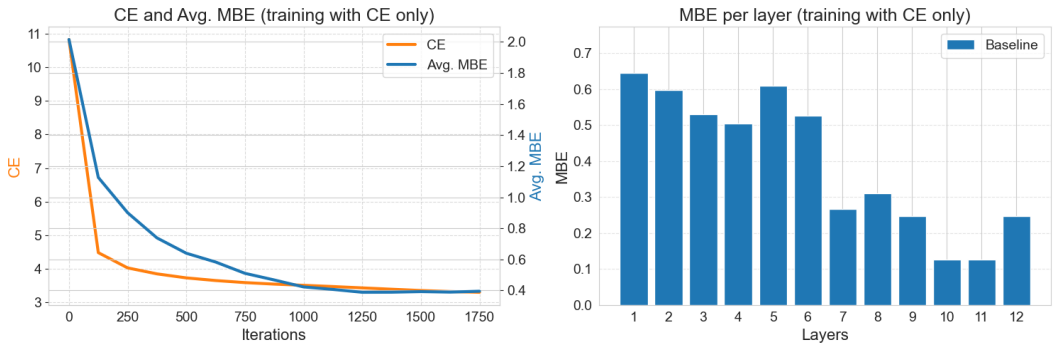


Figure 1: Left: CE and MBE loss curves during pretraining with CE loss only, showing implicit momentum for representation compression. Right: final per-layer MBE values. Later layers show lower MBE, indicating representation compression.

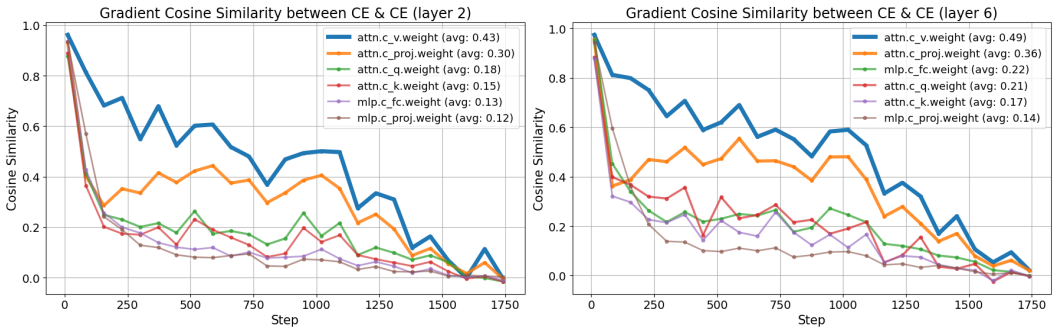


Figure 2: Cosine similarity between CE gradients across batches. CE gradients become increasingly decorrelated over time, reflecting diminishing shared signal.

rhythmic periodicity—suggesting that oscillation is irregular and state-driven, rather than strictly periodic.

4.2 Pre-training with GAPT

In our second experiment, we retain the same GPT pre-training setup but incorporate the GAPT algorithm to test whether it offers a better solution to IBLM objective defined in Equation 2 , compared to a baseline model trained solely on the CE loss. MBE regularization during compression phase in GAPT is done from layer 2 to 9.

Table 1: Cross-entropy loss on the FineWeb validation set.

Model	**CE Loss**
Baseline	3.31
GAPT (Ours)	**3.15** (–4.8%)

As shown in Table 1, GAPT reduces cross-entropy on the FineWeb validation set by 4.8% compared to the CE-only baseline. In addition to reducing test CE loss, GAPT also significantly compresses internal representations. Figure 5 (left) shows the layer-wise MBE for both models. We observe consistent reductions across all regularized layers. The per-layer MBE values and their relative improvements are summarized in Figure 5 (right). GAPT reduces MBE by an average of 70.5% across layers 2–9 while improving validation cross-entropy by 4.8%. This suggests that explicitly alternating between memorization and compression phases offers an effective solution to the

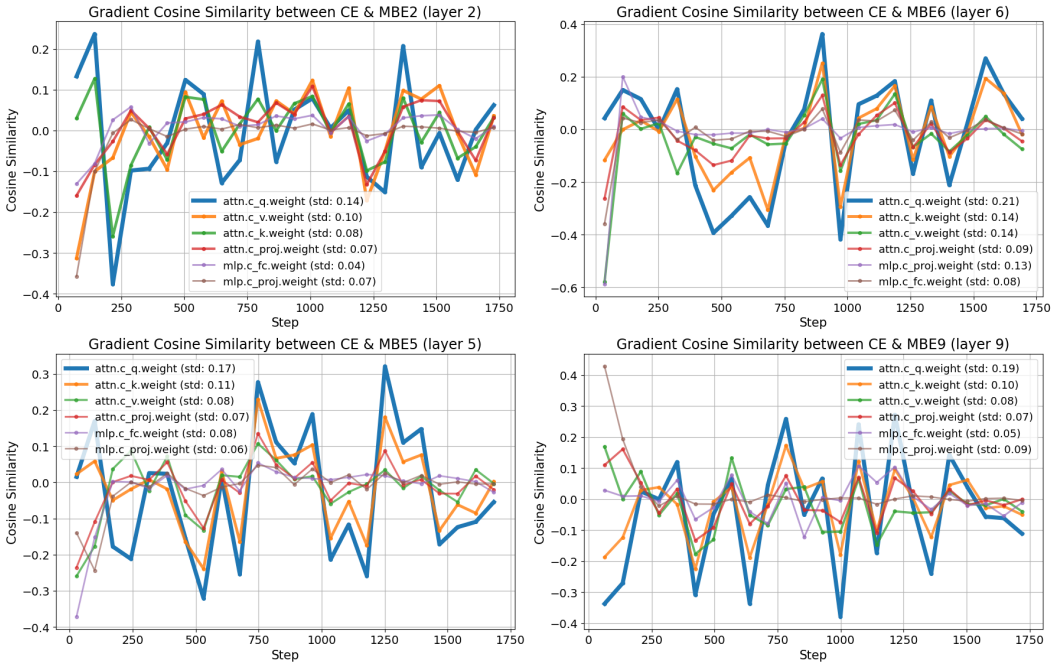


Figure 3: Cosine similarity between CE and MBE gradients over training. Alternating positive and negative phases indicate emergent memorization–compression cycles.

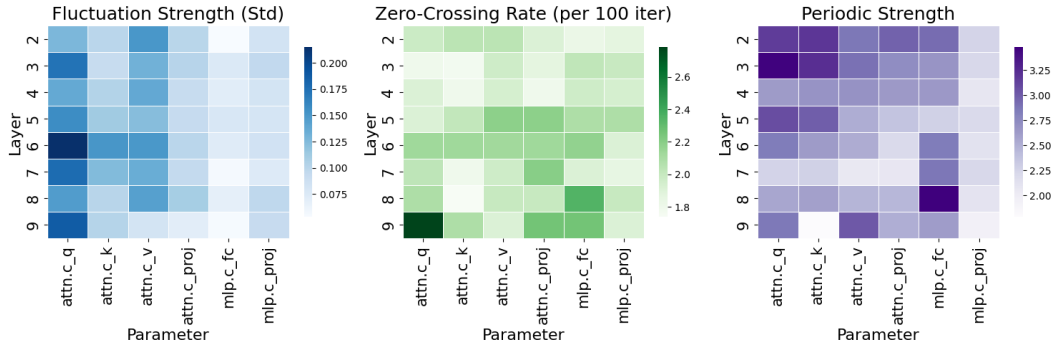
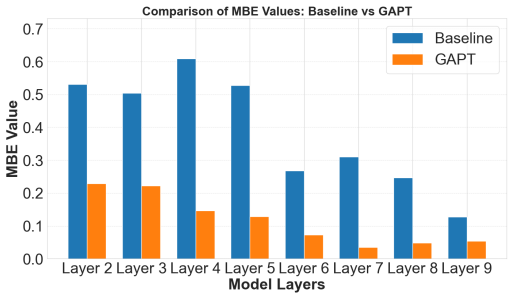


Figure 4: Oscillation metrics between CE and MBE gradients across layers and parameter groups. Left: standard deviation; center: zero-crossing rate; right: periodic strength (peak-to-mean PSD ratio).

constrained optimization objective in IBLM (Equation 2), and can exceed baseline training with cross-entropy target alone.

4.3 Conflicting memory resolution

Inspired by (18), which showed that sleep consolidation helps resolve memory conflicts, we design a synthetic experiment to test whether GAPT can mitigate representation interference between conflicting experiences. We use a 2-layer MLP f_θ with randomly initialized parameters and define a symmetric shift $\Delta\theta$. Inputs $X_1, X_2 \in \mathbb{R}^{10}$ are sampled from Gaussians with shared variance but distinct means:



Layer	Base	GAPT	Reduction
2	0.5313	**0.2285**	–56.99%
3	0.5039	**0.2217**	–56.01%
4	0.6094	**0.1465**	–75.96%
5	0.5273	**0.1279**	–75.74%
6	0.2676	**0.0728**	–72.81%
7	0.3105	**0.0344**	–88.91%
8	0.2471	**0.0483**	–80.44%
9	0.1270	**0.0542**	–57.32%
Avg	–	–	**-70.52%**

Figure 5: Left: Layer-wise MBE for baseline vs. GAPT. Right: Per-layer MBE reduction with GAPT.

$$X_1 \sim \mathcal{N}([1,1,1,1,1,0,0,0,0,0], \sigma^2 I),$$
$$X_2 \sim \mathcal{N}([0,0,0,0,0,1,1,1,1,1], \sigma^2 I)$$

The targets are defined as $Y_1 = f_{\theta+\Delta\theta}(X_1)$ and $Y_2 = f_{\theta-\Delta\theta}(X_2)$, producing two tasks with negatively aligned gradients. We compare GAPT against four baselines: single-task training, ordered training, mixed training, and GAPT applied to mixed batches with MBE regularization.

Table 2: Performance on conflicting experience learning. Lower L1/MBE and higher separation indicate better generalization and disentanglement.

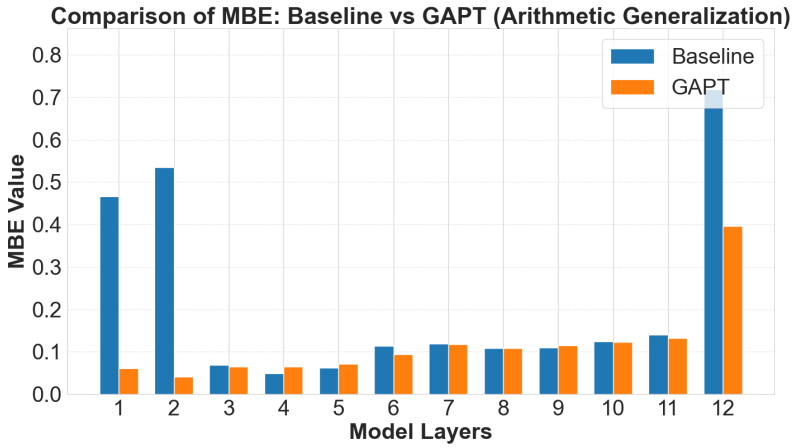
Strategy	L1 (Pos)	L1 (Neg)	MBE (Pos)	MBE (Neg)	Dist.	Sep. Ratio
Pos-only	0.02	0.71	0.18	0.45	–	–
Neg-only	0.92	0.02	0.36	0.36	–	–
Pos $\rightarrow$ Neg	0.43	0.04	0.19	0.15	2.43	2.84
Neg $\rightarrow$ Pos	0.02	0.57	0.19	0.43	1.86	2.33
Mixed	0.03	0.03	0.10	0.22	3.66	4.11
GAPT + MBE (Ours)	**0.03**	**0.03**	**0.02**(–80%)	**0.02**(–91%)	**6.64**(+81%)	**8.08**(+97%)

Table 2 summarizes results. Ordered learning, where the model is trained on one experience and then on the other, suffers from catastrophic forgetting: performance on the first task degrades significantly after exposure to the second. Mixed training alleviates this issue, achieving low L1 loss on both tasks and moderate representation separation. However, GAPT improves over mixed training in both respects: it maintains the same L1 accuracy while achieving a 97% increase in separation ratio and a 91% reduction in MBE.

These results suggest that the compression encouraged by GAPT not only preserves generalization performance but also promotes the disentanglement of conflicting memories. Moreover, compression and separation emerge in tandem during training, closely resembling the consolidation behavior observed in biological neural systems during sleep. This supports the view that compression serves not only as a generalization mechanism, but also as a functional tool for resolving interference in memory (18).

4.4 Arithmetic generalization

To evaluate whether GAPT improves generalization in pre-training language models, we conduct a controlled experiment using a synthetic arithmetic dataset. We pre-train a GPT-2 model from scratch to perform integer multiplication. The training dataset contains 10 million multiplication equations between integers with 1–3 digits. For evaluation, we prepare two test sets: an in-domain (ID) set with 10,000 examples also from the 1–3 digit range, and an out-of-domain (OOD) set with 10,000 examples involving 4–6 digit multiplications. An additional OOD validation set with 1,000 examples is used for early stopping.



Metric	Baseline (1085)	GAPT (Ours, 898)	Change
Entropy (ID)	0.010	0.011	+10%
Entropy (OOD)	4.334	**2.817**	–35%
Avg MBE (0–11)	0.218	**0.115**	–47%

Figure 6: Up: Comparison of baseline vs. GAPT on arithmetic generalization. Bottom: Arithmetic generalization performance summary. GAPT improves OOD generalization and yields more compact representations.

We tokenize the input and output sequences at the per-digit level and train the model for 1,750 iterations with a batch size of 16. Due to observed instability in OOD entropy, we adopt an early stopping strategy that halts training if validation loss increases by more than 20% after iteration 800.

As shown in Figure 6, GAPT improves generalization substantially. It reduces OOD entropy by 35% and average representation entropy (MBE) by 47%, while maintaining similar performance on the in-domain set. This supports our theoretical prediction that minimizing representation entropy leads to stronger generalization under the Information Bottleneck Language Modeling (IBLM) framework.

Interestingly, while MBE regularization is only applied to a subset of layers, we observe that GAPT achieves lower MBE even in unregularized layers (e.g., layers 0, 1, and 11), suggesting a degree of entropy compression generalization across the network. Notably, MBE in layer 1 is reduced by 92%, and in layer 11 by 45%.

5 Relevant work

The **Information Bottleneck (IB)** method was introduced in (2) to formalize the goal of retaining only task-relevant information in representations by minimizing entropy. A theoretical connection between generalization and the cardinality of discrete inputs was established in (6), and later applied to deep networks in (8). However, the entropy–generalization link remained incomplete due to the gap between cardinality and entropy measures. Empirical evidence of a two-phase learning dynamic—early memorization followed by entropy compression—was presented in (15). To quantify entropy in neural networks, matrix-based entropy (MBE) was proposed in (5), and later applied to LLMs in (14), where it correlated with embedding quality and revealed compression trends across checkpoints.

LLMs such as GPT-2 (11) generalize across tasks by scaling training data and parameters (4), effectively compressing the training corpus (17; 13). While post-training methods like instruction tuning (9) improve usability, LLMs still struggle on out-of-distribution tasks such as reverse reasoning (22; 20) and multi-hop inference (20). Recent advancements in RL with verifiable rewards (RLVR) have improved mathematical and coding performance (7), though often by narrowing base model behavior (10). Finally, curriculum-based tokenization growth has been combined with pre-training to improve compression (21).

References

[1] Shannon, C.E.
A Mathematical Theory of Communication.
Bell System Technical Journal, **27**(3), 379–423, 1948.

[2] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint* physics/0004057, 2000.

[3] Cover, T.M. and Thomas, J.A.
Elements of Information Theory (2nd ed.).
Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. *arXiv preprint arXiv:2001.08361, 2020.*

[5] Giraldo, L.G.S., Rao, M., and Principe, J.C.
Measures of entropy from data using infinitely divisible kernels.
IEEE Transactions on Information Theory, 2014.

[6] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, 2010.

[7] DeepSeek AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv preprint arXiv:2501.12948, 2025.

[8] Tishby, N. and Zaslavsky, N.
Deep Learning and the Information Bottleneck Principle.
arXiv preprint arXiv:1503.02406.

[9] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint* arXiv:2203.02155, 2022.

[10] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *arXiv preprint arXiv:2504.13837*, 2025.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I.
Language Models are Unsupervised Multitask Learners.
OpenAI Technical Report, 2019.

[12] Skean, O., Osorio, J.K.H., Brockmeier, A.J., and Giraldo, L.G.S.
DiME: Maximizing Mutual Information by a Difference of Matrix-Based Entropies.
arXiv preprint arXiv:2302.13949, 2023.

[13] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. *arXiv preprint* arXiv:2309.10668, 2024.

[14] Oscar Skean and Md Rifat Arefin and Dan Zhao and Niket Patel and Jalal Naghiyev and Yann LeCun and Ravid Shwartz-Ziv
Layer by Layer: Uncovering Hidden Representations in Language Models
arXiv preprint arXiv:2502:02013.

[15] Ravid Shwartz-Ziv and Naftali Tishby
Opening the Black Box of Deep Neural Networks via Information
arXiv preprint arXiv:1703.00810

[16] Zeyuan Allen-Zhu and Yuanzhi Li
Physics of Language Models: Part 3.1, Knowledge Storage and Extraction
arXiv preprint arXiv:2309.14316

[17] Yuzhen Huang and Jinghan Zhang and Zifei Shan and Junxian He
Compression Represents Intelligence Linearly
arXiv preprint arXiv: 2404.09937

[18] Oscar C González, Yury Sokolov, Giri P Krishnan, Jean Erik Delanois, Maxim Bazhenov
Can sleep protect memorise from catastrophic forgetting? eLife 9:e51005

[19] Tadros, T., Krishnan, G.P., Ramyaa, R., Bazhenov, M
Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks
Nature Communications, **13**, 7742 (2022)

[20] Yu, F., Arora, H.S., and Johnson, M.
Iterative Graph Alignment.
arXiv preprint arXiv:2408.16667 (2024).

[21] Fangyuan Yu. Scaling LLM pre-training with vocabulary curriculum. *arXiv preprint* arXiv:2502.17910, 2025.

[22] Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., and Evans, O.
The Reversal Curse: LLMs Trained on "A is B" Fail to Learn "B is A".
arXiv preprint arXiv:2309.12288 (2024).

[23] Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Vlado Boza, Jiacheng You, Franz Cesista, Braden Koszarsky, and @Grad62304977.
modded-nanogpt: Speedrunning the NanoGPT baseline, 2024.
`https://github.com/KellerJordan/modded-nanogpt`.

A Appendix

Theorem 3 (Minimum Probability Entropy Bound)**.** *Let X be a discrete random variable with sample space Ω, where $|\Omega| = n$. Suppose there exists a constant $\alpha \in \left(0, \frac{1}{n}\right]$ such that $P(X = x) \geq \alpha$ for all $x \in \Omega$. Then the entropy of X is bounded below by:*

$$H(X) \geq -\left(1 - \alpha(n-1)\right) \log_2 \left(1 - \alpha(n-1)\right) - (n-1)\alpha \log_2(\alpha).$$

Furthermore, for sufficiently large n and small α such that $\beta = \alpha n \ll 1$, this bound approximates to:

$$H(X) \geq \beta \log_2(n)$$

Proof. Under the constraint that $P(X = x) \geq \alpha$ for all $x \in \Omega$, the entropy

$$H(X) = -\sum_{x \in \Omega} P(x) \log_2 P(x)$$

is minimized when the distribution is as imbalanced as possible: one outcome has the highest allowed probability, and all others are assigned the minimum α. The worst-case distribution is:

$$\begin{aligned} P(x_1) &= 1 - \alpha(n-1), \\ P(x_i) &= \alpha \quad \text{for } i = 2, \ldots, n. \end{aligned}$$

The resulting entropy is:

$$H(X) = -\left(1 - \alpha(n-1)\right) \log_2 \left(1 - \alpha(n-1)\right) - (n-1)\alpha \log_2(\alpha).$$

For small α and large n such that $\alpha n \ll 1$, we approximate have:

$$1 - \alpha(n-1) \approx 1 - \alpha n.$$

Substituting this into the expression:

$$\begin{aligned} H(X) &\approx -(1-\alpha n)\log_2(1-\alpha n) - \alpha n \log_2(\alpha) \\ &= -(1-\alpha n)\log_2(1-\alpha n) - \alpha n \log_2(\alpha n) + \alpha n \log_2(n) \\ &= h(\alpha n) + \alpha n \log_2(n) \end{aligned}$$

where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.

Since $h(\alpha n) \geq 0$ and $\log_2(1/\alpha) \geq \log_2(n)$ for $\alpha \leq 1/n$, we get:

$$H(X) \geq \alpha n \log_2(n) = \beta \log_2(n)$$

which completes the proof. □

Corrolary 1 (Entropy Lower Bound for Finite Discrete Random Variables)**.** *Let X be a discrete random variable with finite support Ω, where $|\Omega| = n$, and assume that $P(X = x) > 0$ for all $x \in \Omega$. Then there exists a constant $\beta \in (0, 1]$ such that:*

$$H(X) \geq \beta \cdot \log_2 |\Omega|.$$

Proof. Let $\varepsilon := \min_{x\in\Omega} P(X = x)$, which exists and is strictly positive since X is discrete with finite support. By the Minimum Probability Entropy Bound (Theorem 3), we have:

$$H(X) \geq -(1-\varepsilon(n-1))\log_2(1-\varepsilon(n-1)) - (n-1)\varepsilon\log_2(\varepsilon).$$

For small ε and large n, this approximates to:

$$H(X) \geq \varepsilon n \log_2\left(\frac{1}{\varepsilon}\right),$$

which is linear in $n = |\Omega|$. Setting $\beta := \varepsilon n$, the result follows:

$$H(X) \geq \beta \cdot \log_2 |\Omega|.$$

□

Broader Impact and Limitations

Limitations. While theoretically grounded, our proposed GAPT algorithm remains relatively simple. On large-scale pretraining tasks, it achieves less than a 5% improvement over the baseline, falling short of fully realizing the theoretical benefits of compact representations for generalization as suggested by Theorem 1. Furthermore, scaling experiments are missing, which requires more computation resources but may realize more potential in this new dimension for improving generalization. Additionally, entropy minimization is only one approach to compression; other mechanisms—such as sparse activation—may offer stronger biological plausibility and efficiency. A further limitation is observed in the arithmetic generalization task, where GAPT exhibits instability in out-of-distribution performance across different training runs. There could be some data selection side piece missing for stabilizing generalization-oriented training runs.

Broader Impact. Enhancing generalization in AI systems holds the potential to enable models that can extrapolate, reason, and discover novel knowledge—key milestones toward artificial general intelligence. However, systems that generalize without explainability may pose greater risks than those that merely memorize. We caution that aggressive compression of internal representations, while beneficial for efficiency and abstraction, may also increase susceptibility to unintended behaviors such as spurious generalization or adversarial misuse.