Large Language Models as Model Organisms for Human Associative Learning

Camila Kolling **Vy Ai Vo** **Mariya Toneva**
Max Planck Institute for Software Systems, Saarbrücken, Germany
`{ckolling, vyvo, mtoneva}@mpi-sws.org`

Abstract

Associative learning–forming links between co-occurring items–is fundamental to human cognition, reshaping internal representations in complex ways. Testing hypotheses on how representational changes occur in biological systems is challenging, but large language models (LLMs) offer a scalable alternative. Building on LLMs' in-context learning, we adapt a cognitive neuroscience associative learning paradigm and investigate how representations evolve across six models. Our initial findings reveal a non-monotonic pattern consistent with the Non-Monotonic Plasticity Hypothesis, with moderately similar items differentiating after learning. Leveraging the controllability of LLMs, we further show that this differentiation is modulated by the overlap of associated items with the broader vocabulary–a factor we term vocabulary interference, capturing how new associations compete with prior knowledge. We find that higher vocabulary interference amplifies differentiation, suggesting that representational change is influenced by both item similarity and global competition. Our findings position LLMs not only as powerful tools for studying representational dynamics in human-like learning systems, but also as accessible and general computational models for generating new hypotheses about the principles underlying memory reorganization in the brain.

1 Introduction

Associative learning—the ability to form links between co-occurring items—is a fundamental mechanism that shapes how experiences are encoded, stored, and retrieved. Its ubiquity across species and cognitive domains has made it a core component in theories of intelligence [47, 10]. As associations are learned, the brain's internal representations of the associated items are altered—a reflection of the neural plasticity that strengthens some connections while weakening others [34, 36, 12]. A central and ongoing question in cognitive neuroscience is how this self-supervised learning process reshapes representational structure, and why [46, 11, 7]. There are three main hypotheses for how associative learning alters representations in biological systems (see Figure 1A). The classical Hebbian learning rule, where repeatedly associating items strengthens connections between shared features, predicts more integrated representations across learned items [34]. However, alternative theories suggest the opposite. For example, the hippocampus often exhibits pattern separation, where rapidly learned memories reduce representational overlap to minimize interference and facilitate retrieval [27, 2, 54, 11]. These opposing dynamics—integration versus differentiation—are both observed in human studies [7, 11, 38, 35]. To reconcile this, the Non-Monotonic Plasticity Hypothesis (NMPH) posits that representational change follows a U-shaped curve: highly similar or dissimilar items tend to integrate or remain stable, while moderately similar pairs differentiate [34].

Fully testing these hypotheses in biological systems is inherently difficult [46, 33]. A major challenge lies in precisely controlling the similarity between items before learning, a prerequisite for detecting

---

Code available at github.com/bridge-ai-neuro/llm-associative-learning.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

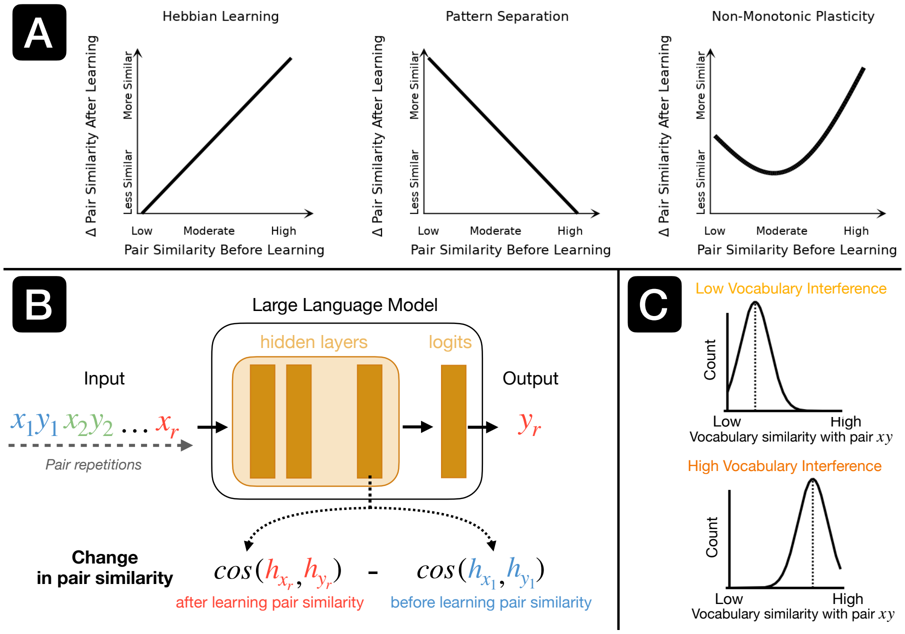


Figure 1: **(A)** Hypotheses about representational changes in humans. Competing theories propose different patterns of representational change as a function of pair similarity before learning: Hebbian learning predicts integration, pattern separation predicts differentiation, and the Non-Monotonic Plasticity Hypothesis (NMPH) predicts a U-shaped curve, with high differentiation at moderate similarity levels prior to learning. **(B)** Schematic of our adapted associative learning task for LLMs. Given repeated in-context presentations of a token pair (x, y), the LLM learns to predict the associated token y. We measure representational change by computing the difference in cosine similarity between hidden representations of the pair before and after learning. This setup, inspired by a neuroscience paradigm [46], enables us to examine whether similar dynamics of representational restructuring emerge during in-context learning. **(C)** Illustration of low and high vocabulary interference in the model's representational space. In the low vocabulary interference case (top, yellow), the target token y is dissimilar to most other tokens, resulting in less competition from alternative completions when paired with x. In the high interference case (bottom, orange), the pairing xy is highly similar to many other possible token pairings, increasing competition and representational pressure to differentiate the learned association from potential distractors during learning.

non-monotonic representational change. Moreover, the similarity level at which differentiation emerges can vary across tasks and stimuli, making it unclear in advance which mid-similarity range will reveal the effect. Capturing this requires dense sampling across the similarity spectrum, further increasing experimental complexity. Finally, human studies are constrained by cost, participant fatigue, and measurement noise, which limit the number of trials that can feasibly be conducted. To address these challenges, we propose using large language models (LLMs) as model organisms for human associative learning. LLMs exhibit complex cognitive behaviors [3, 48, 49], including in-context learning (ICL) [53, 19, 9]—rapidly forming associations without weight updates—making them promising for studying memory dynamics. Unlike hand-crafted neural models designed to replicate specific representational dynamics [33], LLMs offer a scalable, natural testbed for uncovering emergent cognitive phenomena.

In this work, we investigate whether LLMs exhibit representational dynamics akin to human associative learning, and whether they can help disambiguate between competing hypotheses for representational change. We adapt a cognitive neuroscience associative learning paradigm to the ICL setting (Figure 1B), repeatedly presenting token pairs in-context to induce associations. By systematically controlling the similarity of token pairs before learning, we evaluate how representations evolve through learning across six open source, well-performing LLMs. Our initial findings support the NMPH: moderately similar pairs significantly differentiate after learning, mirroring human-like patterns of representational change.

We then leverage the controllability of LLMs to examine a factor that may further contribute to representational change, and is difficult to isolate in biological systems: the similarity between each paired item and the model's prior knowledge. Because LLMs are pre-trained to encode co-occurrence statistics across the entire vocabulary–similar to how humans learn from experience–new associations introduced during ICL must compete with pre-existing patterns. We refer to this competitive influence as *vocabulary interference* (Figure 1C): the extent to which prior knowledge shapes the learning of new associations. In such cases, learning the correct pairing may require greater changes in the model's representations to distinguish it from competing associations. This phenomenon has long been studied in neuroscience and psychology [32, 6, 42], but empirical measurement in the brain is limited by the inability to access all competing representations. By contrast, LLMs provide a tractable framework for quantifying this effect, as the entire distribution of token relationships is explicitly known. We find that, while pair similarity remains a key determinant of representational change, vocabulary interference modulates this effect–greater interference leads to stronger differentiation. These results position LLMs as valuable tools for probing associative learning principles, offering new insights into how both local and global associative structures influence representational change.

2 Related work

2.1 Representational change in human memory and neural models

Integration vs. differentiation in the brain. Associative memory-related representational changes are primarily studied in the hippocampus, a region of the brain thought to be most influential in memory-driven behavior [41, 25]. Both integration and differentiation of memory representations have been shown to support distinct behavioral functions: differentiation reduces interference and enhances specific recall, while integration promotes generalization and inference across related experiences [46, 4]. These functional roles are thought to map onto distinct hippocampal subregions, e.g., integration with CA1 and differentiation with the dentate gyrus (DG), which shows sparse activity linked to orthogonalized representations [46, 51]. How LLMs align with this integration–differentiation spectrum remains an open question.

Non-monotonic plasticity in the brain. The NMPH [34] proposes that representational change depends non-linearly on pair similarity before learning: moderate similarity leads to differentiation, whereas low or high similarity leads to stability or integration. Recently, Wammes et al. [46] provided empirical results supporting this effect by parametrically manipulating the visual similarity of object pairs using CNN-derived [17, 40] representations. Participants arranged images by perceived similarity, and the resulting pairwise distances correlated with model-based similarity estimates. During fMRI, repeated exposure to these pairs revealed significant differentiation for mid-similarity pairs in the DG, but not for other parts of the hippocampus.

Computational accounts. To account for this variety of findings, [33] proposed an unsupervised recurrent network model in which partial activation of competing memories during retrieval induces representational differentiation, a dynamic linked to retrieval-induced forgetting and inhibitory oscillations [24]. While such models are an important step towards a computational account, they rely on hand-crafted inputs and necessitate simplified settings, limiting scalability and behavioral richness. Our work complements prior computational efforts by investigating whether non-monotonic differentiation, previously observed in biological memory systems, emerges naturally in large-scale, general-purpose LLMs trained on real-world data–without an explicit separated memory system.

2.2 Associative learning and in-context dynamics in LLMs

LLMs are increasingly studied as systems capable of associative learning, rapidly forming token-level associations directly within the input context [3]. Recent work shows that LLMs can form stable in-context associations that shape future predictions [19, 53], exhibiting behaviors consistent with retrieval, interference, and generalization [1, 45, 15]. These findings suggest that transformer-based architectures support implicit memory mechanisms across attention and MLP layers, despite the absence of explicit memory modules. Several studies have also analyzed ICL as a form of fast memory encoding or Bayesian inference [5, 15], and have shown that attention layers can support long-range retrieval, stability, and structured generalization [50, 31, 14, 26, 3, 45, 19]. ICL has been interpreted through the lens of both episodic memory, as models retrieve and reuse information based on context, and working memory, given that representations are updated dynamically across tokens without any parameter changes [53, 19, 30, 21, 5, 22, 13, 8]. We build on this line of research by shifting focus from behavioral outcomes to the internal representational dynamics underlying learning through repeated associative exposure.

3 Methods

3.1 Associative learning paradigm

Our associative learning paradigm is inspired by the experimental design of [46], who investigated how repeated exposure to stimulus pairs with different visual similarity leads to non-monotonic changes in human hippocampal representations. We adapt this paradigm to LLMs using ICL [3, 5], replacing visual stimuli with token pairs and modeling learning through repeated token co-occurrence. This setup allows us to examine whether similar non-monotonic representational shifts occur in LLMs, and to what extent LLM behavior parallels hippocampal learning dynamics. We focus on ICL rather than fine-tuning, as LLMs are known to exhibit emergent associative abilities [48, 3, 9], making ICL a natural fit for studying association tasks. It also provides a controlled and biologically plausible analogy to how humans acquire associations [53, 19], while enabling consistent comparisons across models of different sizes and architectures without introducing task-specific fine-tuning.

Formally, we present the token pair (x, y) a total of $r - 1$ times, followed by one final presentation of x alone as a cue for predicting its paired token y. Given the input sequence

$$\mathbf{s} = [x_1, y_1, x_2, y_2, \ldots, x_{r-1}, y_{r-1}, x_r], \tag{1}$$

the model's goal is to generate a prediction of the associated paired token, y. By default, we restrict the number of repetitions such that the total sequence length remains within each model's m maximum context length (L^m_{max}) or the limits imposed by available GPU memory ($L_{\text{mem}} \approx 40k$ tokens), i.e., $L^m = \min(L^m_{\text{max}},\ L_{\text{mem}})$. In our setup, the sequence length is $L_s \approx (2 * r) - 1$.

We predict that the LLM's representations of these tokens will change through the course of ICL, a phenomenon observed in prior studies analyzing ICL tasks [29, 52, 9]. This prediction also aligns with findings from neuroscience, where repeated co-occurrence of stimuli is known to drive representational change in the hippocampus [34, 37].

(Pair) Representational change. For a given model $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of LLM models under study, we extract the hidden representations of a token x at the last layer of the model, $\mathbf{h}^m_x$. We chose to examine the last hidden layer to more effectively control for representations that directly affect model behavior on the ICL task [1]. This choice also aligns with a sensory-information processing hierarchy in which the hippocampus sits at the top of the memory stream [28, 20]. (Pair) Representational change across ICL is then defined as the difference in cosine similarity between representations at repetition r and the first occurrence of the pair:

$$\Delta S^m_r = cos(\mathbf{h}^m_{x_r}, \mathbf{h}^m_{y_r}) - cos(\mathbf{h}^m_{x_1}, \mathbf{h}^m_{y_1}), \tag{2}$$

where the hidden representation is conditioned on the whole sequence up until that point, e.g., $\mathbf{h}^m_{x_r} = \mathbf{h}^m(x_r \mid x_1, y_1, \ldots, x_{r-1}, y_{r-1})$. Note that our ICL paradigm means that the first occurrence of y is always conditioned on x, $\mathbf{h}^m_{y_1} = \mathbf{h}^m(y_1 \mid x_1)$. This design mirrors human associative learning paradigms, where pairs are presented sequentially [46]. Throughout the work, we refer to the hidden states from the first occurrence $(\mathbf{h}_{x_1}, \mathbf{h}_{y_1})$ as the representations obtained before learning occurs.

Token similarity groups. To examine whether LLMs exhibit representational dynamics consistent with those observed in humans, we sample token pairs across the similarity continuum. More specifically, we sampled evenly along the cosine similarity axis, defining 17 groups g that fall within the interval $[0.1, 0.95)$. Each group is defined by a window $[\theta_{min}, \theta_{max})$ that spans 0.05 cosine similarity. The token pairs within each group are chosen such that their representational similarity before learning $cos(\mathbf{h}^m_{x_1}, \mathbf{h}^m_{y_1})$ lies within the group interval $[\theta^g_{\text{min}}, \theta^g_{\text{max}})$.

Details on our procedure to find these token pairs are given in the section below. We find 12 token pairs in each group to form a set for each model, $\mathcal{P}^m$. The token pair sets for each model are constructed independently due to differences in their vocabulary size and tokenization.

---

[1] Preliminary results for the other layers are shown in Appendix D.

3.2 Optimized search for pairs of tokens

To systematically find tokens whose pair similarity before learning falls within a given interval, we employ an efficient way for searching the large vocabulary space (between 10k^2-72k^2 tokens, depending on the model). Inspired by recent work on prompt and input optimization [55, 39], we follow a two-step approximation strategy to identify suitable pairs. The Greedy Coordinate Gradient (GCG) algorithm [55], originally developed for optimizing sequences in adversarial settings (e.g., minimizing next-token likelihood), provides a framework for iteratively refining a sequence by making targeted, gradient-informed edits to individual tokens. We repurpose the GCG method to minimize a loss defined over the cosine similarity of internal representations.

We start with a duplicate token pair (x, x) with the goal of finding a pair (x, y) that falls within the target cosine similarity range $[\theta_{\min}, \theta_{\max})$. We fix the first x token, and iteratively replace the second token of this sequence by using gradient signals (without updating the model) to identify vocabulary items that would bring the pair's cosine similarity closer to the target range. We select replacements from the top-k candidates that reduce the loss the most, repeating the process until the similarity falls within the desired range or a maximum number of steps is hit. This approach efficiently guides token selection in a controlled, representation-aware way, enabling the construction of token pairs with precise similarity properties. More information on this algorithm can be found in Appendix A.

3.3 Estimating vocabulary interference

To estimate how a given pair (x, y) relates to the broader LLM vocabulary space, we fix x and sample each alternative token t from a representative subset of the vocabulary, $\tilde{\mathcal{V}}^m \subset \mathcal{V}^m$. We then compute the similarity between the representation of the correctly associated token y and each alternative token $t \in \tilde{\mathcal{V}}^m$, conditioned on x's presentation in context. This provides an estimate of how much y, when associated with x, competes with other pair completions in the vocabulary space, capturing the degree of the pair's *vocabulary interference* in the model's representational space. Due to the computational cost of exhaustively computing all possible pairwise combinations, we randomly sample $1,000$ tokens from $\mathcal{V}^m$ to form the representative subset $\tilde{\mathcal{V}}^m$, resulting in 1 million pairwise combinations.

Concretely, for each pair (x, t) we extract its pair representation before learning, yielding the set $\mathcal{H}^m_t = \{\mathbf{h}^m_{t_1} \mid \forall t \in \tilde{\mathcal{V}}^m\}$. We then compare the representation of the pair (x, y) before learning to each alternative pair,

$$\mathcal{S}_y^{\tilde{\mathcal{V}}^m} = \{\cos(\mathbf{h}^m_{y_1}, \mathbf{h}^m_{t_1})\ \forall \mathbf{h}^m_{t_1} \in \mathcal{H}^m_t\}. \tag{3}$$

We can interpret $\mathcal{S}_y^{\tilde{\mathcal{V}}^m}$ as a distribution showing how much interference the pair (x, y) receives from all competing associations (x, t) with $t \in \tilde{\mathcal{V}}^m$. We define the vocabulary interference score for each (x, y) as the median of the set $\mathcal{S}_y^{\tilde{\mathcal{V}}^m}$.

All of the above has been described for a single token pair (x, y). The analysis shown in Figure 3a depicts results for token pairs drawn from the original stimulus set described above, $\mathcal{P}^m$, optimized solely for token pair similarity before learning. We then extend the original set of (x, y) pairs, from $\mathcal{P}^m$, to uniformly sample from the joint distribution of before-learning pair similarity and vocabulary interference (Figure 3b). That is, we use the ~ 1 million token pairs from our sub-sampled vocabulary $\tilde{\mathcal{V}}^m$ to yield a larger set $\mathcal{Q}^m$. We aimed to find at least 10 pairs per similarity group g and vocabulary interference group (details in Appendix A).

3.4 Experimental setup

We analyze six recent open-source base LLMs: Llama2-7b, Llama3.1-8b, Llama3.2-1b, Llama3.2-3b, Gemma2-9b, and Mistral-7b [44, 16, 43, 18]. These models were selected for their recency, open availability, and relatively small size within their respective families, providing a balance between computational efficiency and architectural representativeness. All experiments were performed on internal compute clusters, using two NVIDIA H100 PCIe GPUs with $\approx$ 80GB GPU memory per device. The computation of the experiments took a total of ≈ 15 days.

4 Results

4.1 LLMs exhibit structured, multi-phase learning dynamics

As expected, LLMs are able to complete the in-context associative learning task with high accuracy (between $90 - 100\%$), though the number of repetitions required to reach peak accuracy varies across models. Figure 2a shows how overall prediction accuracy evolves as a function of the number of repetitions. We identified three distinct phases of learning–Encoding, Consolidation, and Forgetting–and we observed that their duration varied across models. To enable direct comparison across models, we normalized the number of repetitions in each phase by aligning phase boundaries: repetitions within each phase were linearly rescaled to fixed intervals ($0-1$ for Encoding, $1-2$ for Consolidation, and $2-3$ for Forgetting). This normalization preserves each model's internal dynamics while making phase-aligned trends directly comparable across models. The accuracy curves per model are provided in Appendix B.2.

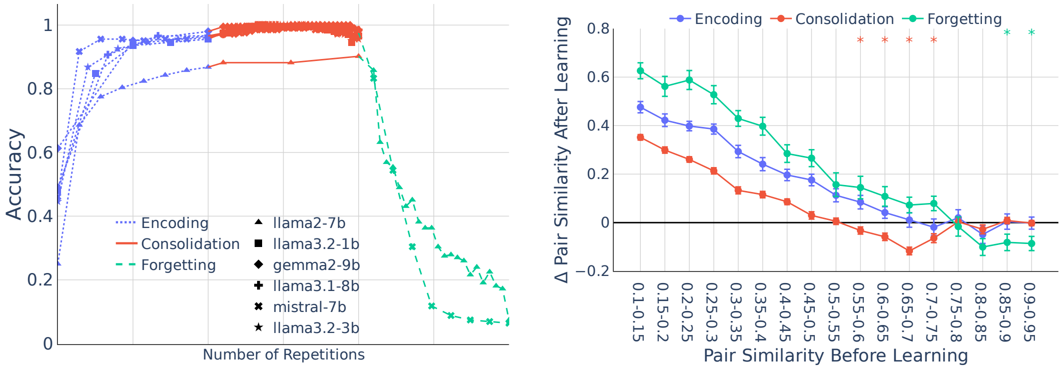
- *Encoding phase (blue):* This phase corresponds to the initial stage of learning, defined by a steep increase in accuracy as the model is repeatedly exposed to the token pair. We define this phase as the period during which accuracy continues to rise by more than 3% between consecutive repetitions, until the model reaches at least 97% of its peak performance.
- *Consolidation phase (red):* This phase reflects a stable performance regime, where the model has largely acquired the association and maintains high accuracy over repetitions. Accuracy remains within $\pm 3\%$ of the peak, indicating that learning has plateaued and performance is stabilized.
- *Forgetting phase (green):* Surprisingly, in some models, accuracy begins to decline even though the number of repetitions remains within the model's maximum context window ($L_s < L^m_{\text{max}}$). We define the forgetting phase as the point where accuracy drops by more than 3% relative to the average of the two prior repetitions, marking the emergence of performance degradation.

While all models exhibited the Encoding and Consolidation phases, only two models (Llama2-7b and Mistral-7b) showed a forgetting phase. For Llama2-7b, forgetting begins relatively early ($r = 40$), whereas for Mistral-7b it emerges much later ($r = 3,000$). We speculate that the delayed forgetting in Mistral-7b may be related to its use of a sliding window attention (SWA). For Llama2-7b, we present preliminary analyses in Appendix B.4, but the underlying cause of early forgetting is not yet well understood. More broadly, it remains unclear how to predict if, and when, forgetting will occur. We leave this question to future work. Overall, these results demonstrate that LLMs can effectively acquire associations and maintain them for a sustained period before eventual degradation.

4.2 Moderately similar pairs significantly differentiate during consolidation

We next investigate how the representations of successfully associated token pairs evolve during learning, specifically focusing on identifying when integration or differentiation occurs. Figure 2b shows how the representational similarity between token pairs changes as a function of their similarity before learning across different phases of learning. We aggregate representational change values ΔS by collapsing across models and token pairs within each similarity group g and learning phase, and report the mean and standard error of the resulting values. To test for differentiation, we performed one-sided paired t-tests for each similarity group and learning phase, testing whether pair similarity after learning was significantly lower than pairsimilarity before learning. To account for multiple comparisons across the 17 similarity groups and 3 learning phases, we applied the Benjamini–Yekutieli (BY) procedure to control the false discovery rate under dependency among tests. Groups that remain significant after BY correction ($p < 0.05$) are marked with asterisks.

During the *Encoding* phase, no significant differentiation is observed across groups that were highly similar before learning. Instead, models show a consistent increase in pairwise similarity for low- to mid-similarity pairs (between 0.1–0.6), reflecting early-stage representational integration: repeated co-occurrence leads these tokens to move closer together in representation space, supporting initial association formation. In contrast, mid- to high-similarity pairs (0.65–0.95) exhibit little to no representational change at this stage.



(a) Accuracy across phases of learning.

(b) Representational change due to learning.

Figure 2: Accuracy and representational changes during learning. (a) Models generally show three phases of learning: encoding, where accuracy steeply increases; consolidation, where accuracy stabilizes; and forgetting, where accuracy declines. To compare across models with different phase lengths, the x-axis is normalized: repetitions within each phase are linearly scaled to fixed intervals ($0 - 1$ for encoding, $1 - 2$ for consolidation, $2 - 3$ for forgetting), allowing phase-aligned trends to be visualized despite variability in learning dynamics. (b) The U-shaped differentiation pattern, characteristic of the Non-Monotonic Plasticity Hypothesis, is observed only during consolidation (red). Asterisks ($*$) indicate groups that remain significant after Benjamini–Yekutieli correction for multiple comparisons across similarity groups and phases ($p < 0.05$).

During the *Consolidation* phase, a striking effect emerges for pairs that were moderately similar before learning ($0.55 - 0.75$): these groups exhibit a significant decrease in pairwise similarity during this phase of learning. This produces a clear U-shaped pattern in representational change–consistent with predictions from the NMPH [34, 46]. Notably, this effect coincides with the stabilization of model performance, suggesting that LLMs undergo structured reorganization of internal representations to maintain high task accuracy. Otherwise, we find that lower similarity pairs still exhibit integration, although to a lesser extent than during Encoding. Higher similarity groups remain largely unchanged, suggesting that their representational similarity is relatively stable across the first two learning phases.

During the *Forgetting* phase, the previously observed non-monotonic pattern disappears, and mid-similarity pairs no longer exhibit significant differentiation. Surprisingly, this is the only phase in which groups that were highly similar before learning show a notable change in their representational structure, displaying clear signs of differentiation relative to their before-learning similarity. Low-similarity pairs, by contrast, undergo even stronger integration than during the Encoding phase. This results in a mild, approximately linear trend in representational change as a function of similarity before learning—resembling the general trend of Encoding, but with greater integration at low similarity and stronger differentiation at high similarity. This trend indicates a loss of structured representational updates, aligning with the observed decline in accuracy. Further results of the evolution of these changes are presented in Appendix B.3.

Taken together, our findings show that LLMs exhibit structured representational dynamics consistent with the NMPH. Interestingly, this non-monotonic pattern is present only during the Consolidation phase, when behavioral performance is stably high, but absent during the Encoding and Forgetting—phases marked by behavioral instability and less structured representational change. Importantly, unlike prior computational models explicitly designed to produce U-shaped dynamics [33], the LLMs that exhibit this non-monotonic effect are general-purpose, pretrained models that were not architecturally constrained or fine-tuned to exhibit such behavior.

4.3 Pair similarity drives representation change, modulated by vocabulary interference

During the *Consolidation* phase—when models exhibited stable maintenance of learned associations—we observed a non-monotonic pattern of representational change as a function of pairwise similarity before learning: low-similarity pairs (up to ≈ 0.5) integrated, mid-similarity pairs (0.55–0.75) differentiated, and high-similarity pairs (> 0.75) showed little to no representational change, aligning best with the NMPH. Building on this analysis, we next examine an additional factor that may contribute to this pattern and is difficult to isolate in biological systems: the similarity between each paired item and the model's prior knowledge. Because LLMs are pre-trained to encode co-occurrence statistics across the entire vocabulary, as humans are thought to do through learning, new associations introduced during ICL must compete with pre-existing patterns. We expect that this competition—what we refer to as vocabulary interference—influences representational change: greater interference (i.e., higher similarity to other items in the vocabulary space) should impose stronger pressure for differentiation to support successful learning.

We thus extend the representational similarity change analysis from Section 4.2 by systematically examining these changes across different levels of vocabulary interference. Specifically, using our original token pairs (x, y) in $\mathcal{P}^m$, we estimate their vocabulary interference with respect to alternative tokens in the set $\tilde{\mathcal{V}}^m$ (see Section 3.3). To facilitate comparison across conditions, we categorize pairs into three equally sized groups based on their vocabulary interference scores: *Low*, *Mid*, and *High* (see Appendix A.4 for details).

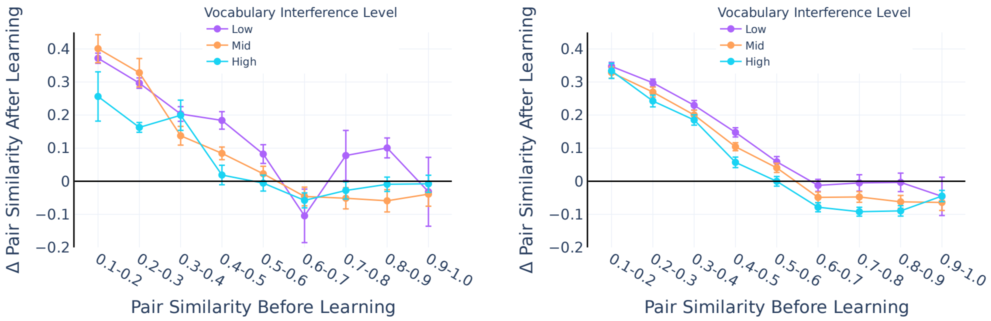
Figure 3a shows how vocabulary interference modulates representational change, with the pattern of this effect varying depending on the pairs' similarity before learning. For low-similarity pairs (up to $0.5 - 0.6$), we observe consistent integration at all levels of vocabulary interference. For mid-similarity pairs $(0.6 - 0.7)$, differentiation emerges as the primary driver across interference levels. High-similarity pairs (above 0.7), however, display high variability and heterogeneous effects: lower interference tends to promote integration, whereas higher interference tends to yield differentiation. This heterogeneity may help explain the apparent U-shaped pattern in our earlier analysis: while low- and mid-similarity pairs exhibit seemingly consistent behavior across interference levels, the variability among high-similarity pairs can mask these opposing trends when averaged, leading to an apparent lack of representational change.

Sampling the full spectrum of vocabulary interference. These findings suggest an interaction between pairwise similarity and vocabulary interference, particularly in the high-similarity regime. To more directly test this interaction, we next control for both factors simultaneously by examining token pairs that span the full joint distribution of pairwise similarity and vocabulary interference (see Section 3.3 and Appendix A for details). To do this, we form an extended set of (x, y) tokens pairs, $\mathcal{Q}^m$, by sampling additional token pairs uniformly across vocabulary interference values. Our approach ensured a minimum of 10 representative pairs per model, similarity group and vocabulary interference level.

Figure 3b shows the results under this controlled sampling regime, where both pair similarity and vocabulary interference are explicitly balanced. As before, we observe a consistent pattern of integration for low-similarity pairs (up to $0.5 - 0.6$). Yet now we observe a robust effect of vocabulary interference: the curves for higher interference levels lie below those for lower interference, indicating reduced, but still present, integration. For mid-similarity pairs, we observe differentiation across all levels of vocabulary interference, and the effect is stronger under higher interference. For high-similarity pairs, we observe a distinct trend: under moderate or high interference, these pairs clearly differentiate, while under low interference, their representations remain relatively stable.

Importantly, across all similarity levels, we observe that higher vocabulary interference is consistently associated with reduced pairwise similarity after learning. One possible explanation for this pattern is that increased interference introduces greater competitive pressure: to reliably associate with each other, paired tokens must distinguish themselves from many similar distractors in the vocabulary. This competition drives the model to reshape representations not only to encode the intended association, but also to preserve distinctiveness within the broader context of the model's prior knowledge.

Representational change through the lens of vocabulary interference. The idea that increased interference introduces competitive pressure provides a useful lens for interpreting the distinct behaviors observed across different pair similarity regimes. For instance, one possible interpretation of the robust integration observed among low-similarity pairs, regardless of vocabulary interference level, is that their representational distance before learning provides greater flexibility for alignment. Because these pairs begin far apart in the representation space, the model can bring them closer (i.e., integrate) without risking excessive overlap that would compromise their individual distinguishability. In this regime, vocabulary interference may impose relatively weak constraints, since the updated representations remain unlikely to be confounded with each other—thus, integration can proceed even



(a) Control for pair similarity.

(b) Control for pair similarity & vocab. interference.

Figure 3: Effect of vocabulary interference on representational change across different pair similarity groups. (a) Results for our original token pairs, sampled uniformly with respect to pair similarity before learning (x-axis). We observe a consistent integration trend for low-similarity pairs and a shift toward differentiation for mid-similarity pairs (0.6–0.7). High-similarity pairs show more heterogeneous behavior, where low interference tends to promote integration, while higher interference tends to yield differentiation. (b) Results for extended token pairs, sampled uniformly over pair similarity before learning (x-axis) and vocabulary interference level (colored lines). Higher vocabulary interference consistently leads to more differentiation, especially for mid- and high-similarity groups. These results suggest that while pairwise similarity is a key driver of differentiation, vocabulary interference amplifies this effect.

under high interference. We speculate that this relative freedom from competition allows the model to prioritize pairwise association without necessitating broader adjustments across the vocabulary space.

At the other extreme, high-similarity pairs begin very close in representational space. Under high vocabulary interference, the model must reshape these representations to prevent confusion with nearby distractors—yet increasing their similarity further could risk entanglement. As a result, differentiation becomes the most likely direction of change, consistent with the strong divergence we observe under high vocabulary interference. In contrast, when vocabulary interference is low, these pairs are already well isolated from the rest of the vocabulary, reducing the pressure for representational differentiation.

Mid-similarity pairs lie in a "sensitive zone" where both factors—pairwise similarity and vocabulary interference—interact most dynamically. They are similar enough to each other that further integration might risk overlap to the point of risking their distinction, yet not similar enough to be clearly associated. Consequently, differentiation appears to be the primary orientation of change for mid-similarity pairs, intensifying with greater vocabulary interference to preserve separability. This suggests that mid-similarity pairs are especially vulnerable to representational reorganization, regardless of the specific interference level.

Therefore, the observed U-shaped curve in the previous analysis (Section 4.2) may be partially explained by a nuanced interaction between pairwise similarity and vocabulary interference. In the high-similarity range, pairs fragment into opposing behaviors across levels of vocabulary interference, so that averaging over these heterogeneous effects can mask systematic representational change and create an illusion of stability.

5 Discussion

In this paper, we investigate whether LLMs exhibit representational changes during associative learning that mirror those observed in humans, and whether they help disambiguate between competing hypotheses about how such changes unfold. Controlling for within pair similarity, we found a non-monotonic pattern of representational change, consistent with the NMPH. This pattern is observed when models stabilize their learning, in what we name the *Consolidation* phase. The fact that LLMs naturally give rise to these dynamics—without any task-specific optimization, and under conditions aligned with how humans learn associations—suggests that they may serve as emergent, flexible model organisms for studying memory reorganization in the brain.

We then leverage the controllability of LLMs to investigate how the vocabulary interference–defined as the interaction between token pair similarity and their similarity to the broader vocabulary–affects representational changes. By introducing this second dimension of analysis, we show that representational dynamics cannot be fully understood in terms of pairwise similarity alone. Instead, representational change reflects a joint influence of pairwise similarity and global contextual competition within the model's prior knowledge. This interaction is especially evident at the extremes: low-similarity pairs integrate consistently across all interference levels, suggesting greater flexibility due to low risk of confusion. High-similarity pairs, by contrast, are already near each other in representational space and face stronger constraints: when vocabulary interference is high, differentiation is the only viable way to maintain separability, whereas under low interference, they remain relatively insulated from external competition, reducing the pressure for further differentiation. Mid-similarity pairs appear to lie at a critical boundary–similar enough to risk confusion, yet not similar enough to form a strong association–making them particularly susceptible to interference-induced differentiation. This sensitivity highlights how small shifts in competitive context can alter the direction of representational change.

Our results show that, while pairwise similarity is a key determinant of representational change, vocabulary interference modulates this effect. This interaction between pair association strength and global contextual interference reveals richer representational dynamic than previously assumed, and may help reconcile diverging findings in the neuroscience literature, where such vocabulary-level interference remains difficult to assess due to limited access to global representational structure. Critically, this kind of fine-grained, systematic manipulation is difficult to achieve in human studies, where both pairwise similarity and global interference are hard to quantify and control. LLMs thus serve as powerful computational model organisms for testing hypotheses about memory dynamics, offering a level of scale and experimental control that is rarely achievable in biological systems.

Limitations and future work. Although LLMs differ mechanistically from human brains, they provide a valuable model system for generating and testing hypotheses that are otherwise challenging to examine in biological systems. Nonetheless, they are not direct stand-ins for humans, and empirical validation in human studies remains essential.

Our operationalization of vocabulary interference also has limitations. By design, it estimates representational competition from the broader vocabulary space, but this approximation may not fully capture the dynamics of interference in human memory, where similarity is shaped by experience, attention, and context. Furthermore, our measure relies on sampled subsets of tokens for tractability, which may underrepresent the true structure of competition across the full vocabulary.

Methodologically, our analysis focused on the final hidden layer of each model, with only preliminary exploration of earlier layers. Future work could systematically track representational change across layers, providing an extended analysis of how interference and differentiation emerge throughout the model hierarchy. Additionally, to ensure coverage across a wide range of similarity values, we used token pairs defined by geometric properties rather than naturalistic semantics. A preliminary analysis on WordNet stimuli is provided in Appendix C, with further investigation left for future work.

Finally, this work focuses on hypotheses tested in mature adult brains, leaving open the question of how these processes emerge during development. Promising future directions include exploring curriculum-learning setups that more closely mirror human developmental trajectories, analyzing attention patterns to identify circuit-level mechanisms involved in resolving interference, and examining how representational dynamics evolve during fine-tuning and longer-term learning.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) - DFG Research Unit FOR 5368 and by the Max Planck Institute for Software Systems graduate center. We thank Omer Moussa and Mathis Pink for providing valuable feedback on earlier versions of this work.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

[2] Tarek Amer and Lila Davachi. Extra-hippocampal contributions to pattern separation. *eLife*, 12:e82250, March 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Iva K. Brunec, Jessica Robin, Rosanna K. Olsen, Morris Moscovitch, and Morgan D. Barense. Integration and differentiation of hippocampal memory traces. *Neuroscience & Biobehavioral Reviews*, 118:196–208, November 2020.

[5] Thomas F Burns, Tomoki Fukai, and Christopher J Earls. Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture. *arXiv preprint arXiv:2412.15113*, 2024.

[6] Jeremy B Caplan, Mayank Rehani, and Jennifer C Andrews. Associations compete directly in memory. *Quarterly Journal of Experimental Psychology*, 67(5):955–978, 2014.

[7] Avi JH Chanales, Alexandra G Tremblay-McGaw, Maxwell L Drascher, and Brice A Kuhl. Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological science*, 32(5):705–720, 2021.

[8] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[9] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[10] Athanasios Drigas and Eleni Mitsea. The 8 pillars of metacognition. *International Journal of Emerging Technologies in Learning (iJET)*, 15(21):162–178, 2020.

[11] Serra E Favila, Avi JH Chanales, and Brice A Kuhl. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature communications*, 7(1):11066, 2016.

[12] Serra E Favila, Hongmi Lee, and Brice A Kuhl. Transforming the concept of memory reactivation. *Trends in neurosciences*, 43(12):939–950, 2020.

[13] Zafeirios Fountas, Martin A Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*, 2024.

[14] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[15] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[19] Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *Advances in Neural Information Processing Systems*, 37:67712–67757, 2024.

[20] Pierre Lavenex and David G Amaral. Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus*, 10(4):420–430, 2000.

[21] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*, 2022.

[22] Ji-An Li, Corey Zhou, Marcus Benna, and Marcelo G Mattar. Linking in-context learning in transformers to human episodic memory. *Advances in Neural Information Processing Systems*, 37:6180–6212, 2024.

[23] Jorge Nocedal and Stephen J Wright. Line Search Methods. In *Numerical Optimization*, pages 30–65. Springer New York, 2006.

[24] Kenneth A Norman, Ehren Newman, Greg Detre, and Sean Polyn. How inhibitory oscillations can train neural networks and punish competitors. *Neural computation*, 18(7):1577–1610, 2006.

[25] John O'keefe and Lynn Nadel. Précis of o'keefe & nadel's the hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979.

[26] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[27] Randall C. O'Reilly and James L. McClelland. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, 4(6):661–682, December 1994.

[28] Randall C O'Reilly, Yuko Munakata, Michael J Frank, Thomas E Hazy, et al. Computational cognitive neuroscience. 2012.

[29] Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

[31] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[32] Jeroen G Raaijmakers and Richard M Shiffrin. Search of associative memory. *Psychological review*, 88(2):93, 1981.

[33] Victoria JH Ritvo, Alex Nguyen, Nicholas B Turk-Browne, and Kenneth A Norman. A neural network model of differentiation and integration of competing memories. *Elife*, 12:RP88608, 2024.

[34] Victoria JH Ritvo, Nicholas B Turk-Browne, and Kenneth A Norman. Nonmonotonic plasticity: how memory retrieval drives learning. *Trends in cognitive sciences*, 23(9):726–742, 2019.

[35] Anna C Schapiro, Lauren V Kustner, and Nicholas B Turk-Browne. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current biology*, 22(17):1622–1627, 2012.

[36] Anna C Schapiro, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160049, 2017.

[37] Anna C Schapiro, Nicholas B Turk-Browne, Kenneth A Norman, and Matthew M Botvinick. Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, 26(1):3–8, 2016.

[38] Margaret L Schlichting, Jeanette A Mumford, and Alison R Preston. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature communications*, 6(1):8151, 2015.

[39] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, November 2020. Association for Computational Linguistics.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Larry R Squire. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 99(2):195, 1992.

[42] Shauna M Stark, Zachariah M Reagh, Michael A Yassa, and Craig EL Stark. What's in a context? cautions, limitations, and potential paths forward. *Neuroscience letters*, 680:77–87, 2018.

[43] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[45] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[46] Jeffrey Wammes, Kenneth A Norman, and Nicholas Turk-Browne. Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *elife*, 11:e68344, 2022.

[47] Edward A Wasserman and Ralph R Miller. What's elementary about associative learning? *Annual review of psychology*, 48(1):573–607, 1997.

[48] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[50] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

[51] Michael A Yassa and Craig EL Stark. Pattern separation in the hippocampus. *Trends in neurosciences*, 34(10):515–525, 2011.

[52] Safoora Yousefi, Leo Betthauser, Hosein Hasanbeig, Raphaël Millière, and Ida Momennejad. Decoding in-context learning: Neuroscience-inspired analysis of representations in large language models. *arXiv preprint arXiv:2310.00313*, 2023.

[53] Jiachen Zhao. In-context exemplars as clues to retrieving from large associative memory. *arXiv preprint arXiv:2311.03498*, 2023.

[54] Ewa Zotow, James A. Bisby, and Neil Burgess. Behavioral evidence for pattern separation in human episodic memory. *Learning & Memory*, 27(8):301–309, August 2020.

[55] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and at the end of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of our work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not involve any proofs or assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: As described in Section 3 and throughout the Appendix, we have provided detailed descriptions and analyses of the experimental setups for all our investigations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have attached with the submission the code necessary to reproduce our main results and upon acceptance we will publicly release it.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: As detailed in Section 3, we have thoroughly described the experimental setups for all our experiments. Additional details have been provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included error bars and statistical tests in Section 4 and their corresponding explanations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As detailed in Section 3, we outline the specific model compute resources provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We are convinced that we comply with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have addressed the broader impacts of our work in Section 5. Additionally, as our research is primarily an empirical exploration and poses no additional social risks, we have not included a discussion on potential harmfulness.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use LLM models and they are properly credited in Section 3.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We describe the use of LLM models in Section 3 and in our Appendix.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.