Non-Monotonic Plasticity in Large Language Models

Camila Kolling (ckolling@mpi-sws.org)
Max Planck Institute for Software Systems
Campus E1 5, 66123 Saarbrücken, Germany

Mariya Toneva (mtoneva@mpi-sws.org)
Max Planck Institute for Software Systems
Campus E1 5, 66123 Saarbrücken, Germany

Abstract

Neural representations in biological memory systems change systematically during associative learning. The Non-Monotonic Plasticity Hypothesis (NMPH) proposes that these changes follow a surprising U-shaped pattern based on how strongly two items are initially related, with initially-moderately-related items becoming significantly more *distinct* after learning, rather than more similar. We provide the first evidence that large language models (LLMs) also exhibit this non-monotonic pattern of representational change, aligning with the NMPH observed in humans. Using an in-context associative learning paradigm, with no changes to model weights, we show that moderately similar token pairs significantly differentiate, and this differentiation occurs when accuracy is both highest and most stable across repeated item presentations. Our results suggest that LLMs can model human associative learning, offering a framework to study representational change during learning.

Keywords: LLMs; Associative Learning; In-Context Learning; Non-Monotonic Plasticity; Representational Change

Introduction

As humans learn, neural representations change. While the traditional Hebbian view suggests that associating two items makes their representations more similar, leading to integration, empirical studies (Chanales, Tremblay-McGaw, Drascher, & Kuhl, 2021; Favila, Chanales, & Kuhl, 2016; Schlichting, Mumford, & Preston, 2015) have found both integration and differentiation. To explain this variability, the Non-Monotonic Plasticity Hypothesis (NMPH) (Ritvo, Turk-Browne, & Norman, 2019) proposes a U-shaped relationship between initial item similarity and representational change.

Recent findings further support the NMPH (Ritvo et al., 2019), showing strong differentiation for moderately similar items. However, the mechanisms behind this effect remain unclear. One explanation links differentiation to suppression of competing memories, and integration to successful retrieval (Hulbert & Norman, 2015; Norman, Newman, Detre, & Polyn, 2006). A computational model has been proposed to explain this process (Ritvo, Nguyen, Turk-Browne, & Norman, 2024), but its simplified architecture, explicitly designed to produce U-shaped learning curves, may limit its ability to fully capture memory dynamics.

In this work, we show for the first time that non-monotonic representational changes also emerge in pretrained large language models (LLMs) during in-context learning, which has been proposed as a model of rapid associative learning (Zhao, 2023; Y. Jiang, Rajendran, Ravikumar, & Aragam, 2024). The representational changes mirror those observed in humans, with significant differentiation between mid-similarity item pairs (Fig. 1), consistent with the NMPH. We further show that this change occurs during what we term the "consolidation phase", a phase that stabilizes learned associations while preserving high accuracy. These findings suggest LLMs may model human associative learning and could potentially offer insights into differentiation mechanisms.

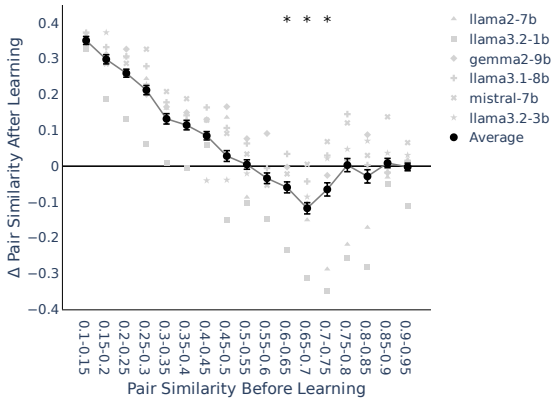
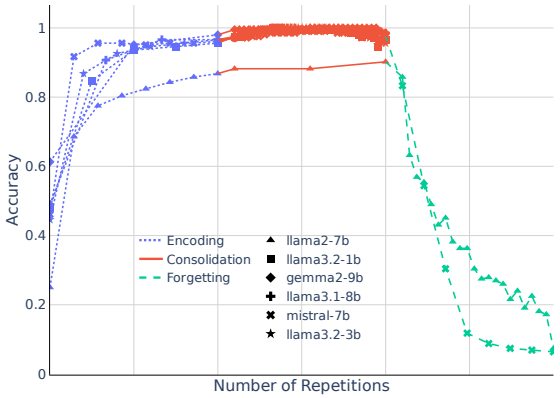


Figure 1: Representational change due to learning during the "consolidation" phase is consistent with the Non-Monotonic Plasticity Hypothesis. Asterisks ($*$) indicate values significantly lower than 0 ($p < 0.05$).

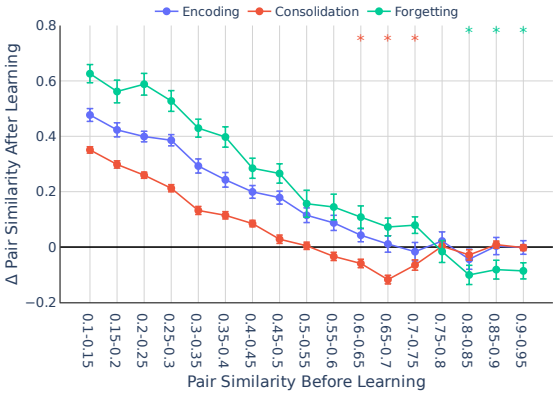
Methods

Learning Paradigm. Our associative learning paradigm is inspired by (Wammes, Norman, & Turk-Browne, 2022) and adapted to large language models (LLMs) using in-context learning (Brown et al., 2020; Burns, Fukai, & Earls, 2024). To allow the LLM to learn an association between a token pair (e.g. A_1 and A_2), we present multiple repetitions of the token pair to the LLM in the input context (e.g., $A_1A_2A_1A_2\ldots A_1$). This setup mimics statistical learning, where hippocampal representations adapt through repeated co-occurrence. We evaluate the associative learning by the accuracy of predicting the correct token in the token pair (e.g. A_2 if the last input token is A_1). To study representational changes in the token pair, we extract representations from the LLM for the tokens pre and post learning. The pre-learning representations are extracted for each token after only one presentation of the token pair, mimicking initial stimulus exposure in the hippocampus before structured learning. The post-learning representations are extracted after the full number of repetitions of the token pair, at the position of the last token in the input and the first predicted token. We focus our analysis on the last layer, as it is closest to the model's output and thus most directly reflects behavior.

Pairs of Tokens Search. We manipulate the initial relatedness between paired items by varying their representational similarity. To find LLM token pairs with a pre-specified cosine similarity, we adapt the Greedy Coordinate Gradient algorithm (Zou et al., 2023). For a given token A_1, we compute its cosine similarity with a second token A_2, where representations are extracted from a contextualized prompt A_1A_2 (i.e., A_2's representation is conditioned on A_1). We initialize A_2 by projecting its one-hot encoding through the model's embedding matrix. The gradient of this similarity is then used to iteratively adjust A_2, prioritizing the tokens with the highest impact on similarity adjustments. The algorithm continues refining A_2



(a) Accuracy across phases of learning.



(b) Representational change due to learning.

Figure 2: Accuracy and representational changes during learning. (a) Models exhibit three distinct phases during learning: encoding, where accuracy steeply increases; consolidation, where accuracy stabilizes; and forgetting, where accuracy declines. The x-axis for each model is scaled by the length of its learning phase. (b) The U-shaped differentiation pattern, characteristic of the Non-Monotonic Plasticity Hypothesis, is observed only during consolidation (red). Asterisks ($*$) indicate $p < 0.05$.

until its similarity to A_1 falls within the target range. We define 17 similarity groups spanning cosine ranges from 0.1 to 0.95 in increments of 0.05, and sample 12 token pairs per group.

Models & Measures. We analyze 6 recent open-source language models (Fig. 1) (Touvron et al., 2023; Grattafiori et al., 2024; Team et al., 2024; A. Q. Jiang et al., 2023), selected for efficiency and representativeness. To quantify learning-related representational change, we compute the difference in pair cosine similarity (post – pre) across models. We test significance using one-sided paired t-tests, evaluating whether post-learning similarity is lower than pre-learning similarity. Asterisks mark groups with significant differentiation.

Results

Accuracy. All models successfully learn the association task, reaching $90-100\%$ accuracy. The number of repetitions of the paired tokens that are needed to learn the task differs across models, and we observe three distinct learning phases as the number of repetitions increases (Fig. 2a): (1) an encoding phase (blue), during which the model learns the association and the accuracy rises sharply $\geq 97\%$ of the model's maximum accuracy; (2) a consolidating phase (red), during which the model retains the learned association without much fluctuation in its performance ($\leq 3\%$ relative change in accuracy between consecutive repetitions), and (3) a forgetting phase (green), where accuracy starts to decline ($\geq 3\%$ relative decrease in accuracy). All models reach high and stable accuracy, transitioning from the encoding to the consolidation phase, within $3-8$ repetitions. Only Llama2-7b and Mistral-7b show a forgetting phase, with Llama2-7b beginning to forget early (at $r = 40$) and Mistral-7b much later (at $r = 3000$).

Representational Change. We next examine how representations evolve across the three learning phases and whether mid-similarity pairs exhibit differentiation, as predicted by the NMPH. Fig. 2b shows the average representational change across all models, as a function of pre-learning pair similarity. The values are computed over occurrences of paired tokens in context for each learning phase, with error bars indicating the standard error of the mean across models. During the encoding phase, models show a general downward trend, indicating that higher similarity leads to more differentiation between paired items. In contrast, during the consolidation phase, a U-shaped pattern appears, with mid-similarity pairs $(0.6–0.75)$ exhibiting significant differentiation. This suggests that once LLMs stabilize their learning, they obtain a structured representational differentiation consistent with the NMPH. However, in the forgetting phase, this effect disappears as differentiation occurs only for highly similar paired tokens, resembling the initial encoding phase. This indicates a loss of structured representational updates, aligning with the decline in accuracy. These findings suggest that LLMs undergo distinct encoding, consolidation, and forgetting stages. Notably, the non-monotonic pattern, where mid-similarity pairs exhibit significant differentiation, emerges only during stable consolidation. This structured representational change supports sustained accuracy, suggesting that differentiating mid-similarity pairs may play a key role in maintaining learned associations.

Discussion and Conclusion

We show that LLMs exhibit non-monotonic representational change during in-context learning, consistent with the NMPH: mid-similarity pairs significantly differentiate during consolidation, preserving accuracy through stabilized associations. This suggests that LLMs may serve as computational models for memory organizations in humans, offering a framework to study learning dynamics and representational change. Future work could test if this mechanism extends to fine-tuning, which more closely resembles long-term synaptic modification, and help refine theories of differentiation in both artificial and biological systems.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burns, T. F., Fukai, T., & Earls, C. J. (2024). Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture. *arXiv preprint arXiv:2412.15113*.

Chanales, A. J., Tremblay-McGaw, A. G., Drascher, M. L., & Kuhl, B. A. (2021). Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological science*, *32*(5), 705–720.

Favila, S. E., Chanales, A. J., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature communications*, *7*(1), 11066.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... others (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hulbert, J. C., & Norman, K. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex*, *25*(10), 3994–4008.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... Sayed, W. E. (2023). *Mistral 7b.* Retrieved from https://arxiv.org/abs/2310.06825

Jiang, Y., Rajendran, G., Ravikumar, P., & Aragam, B. (2024). 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.

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural computation*, *18*(7), 1577–1610.

Ritvo, V. J., Nguyen, A., Turk-Browne, N. B., & Norman, K. A. (2024). A neural network model of differentiation and integration of competing memories. *Elife*, *12*, RP88608.

Ritvo, V. J., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic plasticity: how memory retrieval drives learning. *Trends in cognitive sciences*, *23*(9), 726–742.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature communications*, *6*(1), 8151.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., ... others (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wammes, J., Norman, K. A., & Turk-Browne, N. (2022). Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *elife*, *11*, e68344.

Zhao, J. (2023). In-context exemplars as clues to retrieving from large associative memory. *arXiv preprint arXiv:2311.03498*.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.