How Do Large Language Models Learn Concepts During Continual Pre-Training?

Anonymous ACL submission

Abstract

Human beings primarily understand the world through concepts (e.g., *dog*), abstract mental representations that structure perception, reasoning, and learning. However, how large language models (LLMs) acquire, retain, and forget such concepts during continual pretraining remains poorly understood. In this work, we study how individual concepts are acquired and forgotten, as well as how multiple concepts interact through interference and synergy. We link these behavioral dynamics to LLMs' internal **Concept Circuits**, computational subgraphs associated with specific concepts, and incorporate **Graph Metrics** to characterize circuit structure. Our analysis reveals: (1) LLMs concept circuits provide a non-trivial, statistically significant signal of concept learning and forgetting; (2) Concept circuits exhibit a stage-wise temporal pattern during continual pretraining, with an early increase followed by gradual decrease and stabilization; (3) concepts with larger learning gains tend to exhibit greater forgetting under subsequent training; (4) semantically similar concepts induce stronger interference than weakly related ones; (5) conceptual knowledge differs in their transferability, with some significantly facilitating the learning of others. Together, our findings offer a circuit-level view of concept learning dynamics and inform the design of more interpretable and robust concept-aware training strategies for LLMs.

1 Introduction

Humans organize perception and reasoning around **concepts**—abstract mental categories (e.g., *dog*) that enable generalization from individual observations to shared properties, relations, and actions (Harpaintner et al., 2018). Today's large language models (LLMs) are the most capable foundation models, and a central goal of their pretraining is to *abstract conceptual knowledge* and encode it as internal *concept-level* representations in model parameters that support downstream reasoning and generation. As new concepts continually emerge and models are updated through continual pre-training, it becomes essential to understand how reliably new concepts are acquired and prior concepts are retained, and how to make concept learning more efficient in the presence of interactions among the many concepts and types of conceptual knowledge.

Prior studies have examined concept learning in LLMs from two angles. One line of work uses prompt-based **knowledge probing** to test whether specific conceptual properties or relations (e.g., *commonsense facts*) can be elicited from a pretrained model (Gu et al., 2023b; Liao et al., 2023; Shani et al., 2023a; Zheng et al., 2024; Peng et al., 2022; Xu et al., 2024a), while another leverages **mechanistic interpretability** tools to localize internal structures associated with conceptual information (Aljaafari et al., 2024; Wang et al., 2024b). However, these efforts provide only a partial view of concept learning: they typically probe isolated pieces of conceptual knowledge that models may already possess, rather than systematically characterizing how concepts are acquired, consolidated, and forgotten over the course of continual pre-training.

In this work, we take a step toward filling this gap by asking two more research questions: (1) *how internal concept representations correlate with concept acquisition and forgetting*, and (2) *how they relate to interference and synergy among multiple concepts and types of knowledge during joint training*. Answering these questions matters not only for interpretability, but also for practice: it can inform concept-aware continual pre-training decisions such as how much training is needed for new concepts, how to schedule and reorder training data, and how to reduce destructive interference among semantically related concepts.

To enable a controlled study, we introduce the FICO dataset, built from ConceptNet-derived con-

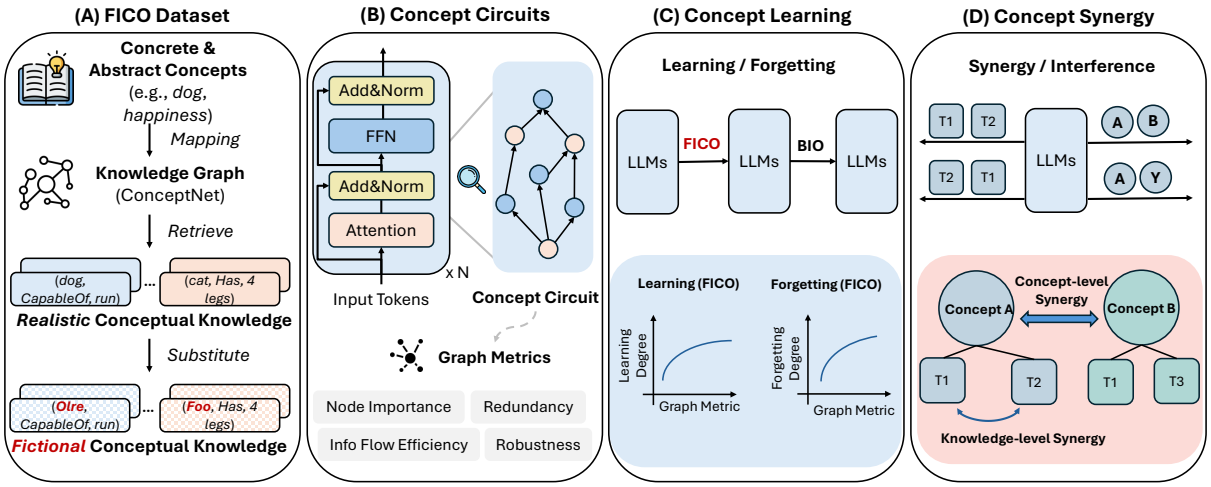


Figure 1: (A) Construct the FICO dataset based on ConceptNet. (B) Extract **Concept Circuits**, LLM computational subgraph associated with individual concepts, and characterize their structure using graph metrics. (C) Analyze **concept learning and forgetting dynamics** in two-stage continual pretraining. (D) Study **synergy and interference** across concepts (e.g., A, B, Y) and knowledge type (e.g., $T1$ and $T2$).

ceptual relations but mapped to synthetic, non-existent concept names, preserving realistic knowledge structure while making the concepts novel to the model. We then conduct controlled continual pre-training experiments and link behavioral changes to internal mechanisms by extracting **Concept Circuits**, which are computational subgraphs associated with each concept, using circuit identification methods and characterizing their structure over time with graph metrics. This joint analysis allows us to trace how concept-level behavior (e.g., learning and forgetting measures) co-evolves with circuit topology, and how cross-concept relatedness and knowledge-type ordering shape interference and transfer, as shown in Figure 1.

Our results reveal several meaningful and systematic patterns in concept learning dynamics in LLMs: **(1)** circuit-level graph metrics provide a modest yet consistent signal of both concept learning and forgetting, highlighting a structural trade-off: denser and more robust concept circuits tend to support stronger acquisition, while more modular circuit organization helps mitigate forgetting under continual training. This trade-off is reflected behaviorally: concepts with larger learning gains exhibit greater forgetting during subsequent training, indicating an inherent tension between the strength of acquisition and long-term stability. **(2)** We observe that during continued training on new domains, concept circuits corresponding to previously learned concepts follow a stable stage-wise temporal trajectory across multiple graph metrics, with an early increase followed by gradual decrease and stabilization, suggesting distinct phases of concept consolidation. **(3)** At the level of multi-concept learning, semantically similar concepts interfere more strongly than weakly related concepts, and different types of conceptual knowledge differ substantially in their transferability; e.g., pretraining on Hyponym & Hypernym knowledge improves subsequent learning performance on Synonym & Antonym knowledge by 63.74%. Together, these findings offer a circuit-level view of concept learning dynamics and suggest how concept-aware curricula and scheduling could make continual pretraining more stable and efficient. Our contributions are summarized as follows:

- We present the first systematic study of concept acquisition and forgetting in LLMs during continual pretraining, revealing consistent correlations between concept circuits, learning dynamics, and forgetting behavior, suggesting these signals as indicators of future learning and forgetting.
- We present a relational perspective on knowledge learning in LLMs, revealing interference and synergy patterns across concepts and knowledge types during joint training, motivating interference-aware data scheduling.
- Our analysis reveals that concept circuits follow a stage-wise trajectory across graph metrics during continued training, indicating distinct phases of concept consolidation.

2 Dataset Construction

We follow previous studies (Wang et al., 2024c) and define a **concept** (e.g., *dog*) as an abstraction over the world that captures the shared features and essential characteristics of a class of entities. **Conceptual knowledge** (e.g., *a dog has four legs*) refers to factual or relational information associated with a concept, reflecting familiarity with and understanding of its properties and relations (Wang et al., 2024a). Each concept is therefore linked to a set of conceptual knowledge that collectively describe its semantic content. To study concept learning across different levels of concreteness, we consider both **concrete concepts** that are grounded in direct sensory experience (e.g., *dog*), and **abstract concepts** which are not directly perceptible (e.g., *love*). We sample 500 concrete concepts from the THINGS dataset (Hebart et al., 2023) and 500 abstract concepts from the Concreteness Ratings dataset (Brysbaert et al., 2014). For each concept, we retrieve associated conceptual knowledge from ConceptNet (Speer et al., 2017), a large-scale knowledge graph of millions of concepts and 34 typed relations. To improve interpretability, we consolidate these relations into five high-level knowledge types, as shown in Appendix A.

To mitigate influence of pre-existing knowledge encoded in LLMs, we replace each real concept name with one fictional name generated by GPT-5 (Hurst et al., 2024), while preserve associated conceptual knowledge, as shown in Figure 1(A), to form our FICO, **FI**ctional **CO**ncept dataset, as detailed in Section B in Appendix. Following prior work (Zucchet et al., 2025b), we use GPT-5 to produce natural-language templates for each relation type (e.g., "*{concept} has the ability to*" for *CapableOf*), converting knowledge triples (e.g., $(\text{dog}, \text{CapableOf}, \text{run})$) into prefix–target training examples. For evaluation, we sample 500 concepts and use a *disjoint* template pool to construct test instances with novel surface forms, enabling assessment of whether LLMs learns underlying relations rather than memorizing templates.

3 How Internal Concept Representations Correlate with LLM Learning and Forgetting Dynamics?

3.1 Concept Circuit as Internal Concept Representation

Prior work (Yao et al., 2025) models a pretrained LLM as a directed acyclic graph (DAG), where nodes correspond to computational components in the forward pass (e.g., neurons, attention heads, embeddings), and edges capture their interactions (e.g., residual connections, attention operations, and linear projections). Given a knowledge triple $k_{ij} = (c_i, r_{ij}, o_{ij})$, where c_i is a subject concept (e.g., *dog*), r_{ij} is a relation type (e.g., *HasProperty*), and o_{ij} is an object which can be another concept or a descriptive phrase (e.g., *fleas* or *four legs*), a **Knowledge Circuit** is defined as the minimal computational subgraph that can faithfully predict the target object o_{ij} conditioned on a textual prefix converted from the subject–relation pair (c_i, r_{ij}) (Yao et al., 2025). We adapt this notion to concept-level analysis and define a **Concept Circuit** for concept c_i as the computational subgraph that can faithfully predict all conceptual knowledge $\{k_{i0}, k_{i1}, ..., k_{ij}, ...\}$ associated with c_i, thereby serving as an internal representation of the concept within the model's parametric memory.

To extract concept circuits at different checkpoints during continual training, we use EAP-IG (Hanna et al., 2024) as our circuit identification method. EAP-IG assigns an importance score to each edge while balancing computational efficiency with attribution faithfulness[1]. Given the edge importance scores, we construct a circuit by selecting the top-scoring edges such that the resulting subgraph preserves at least 70% of the full model's performance on the corresponding concept.

Graph Metrics for Concept Circuits To characterize the structure of concept circuits and their relationship to concept learning dynamics in LLMs, we compute four families of standard graph-theoretic metrics: **(1) Node Importance**, measured as the standard deviation of *eigenvector centrality* (Newman, 2010), which quantifies how unevenly structural influence is distributed across nodes. Higher variance indicates a more concentrated hub structure, which may facilitate concept acquisition but increase vulnerability to interference or forgetting. **(2) Redundancy**, measured by *density* (Newman, 2010), defined as the ratio of existing edges to the maximum possible number of edges. Higher density reflects more redundant connections. **(3) Information Flow Efficiency**, measured by *global efficiency* (Latora and Marchiori, 2001), i.e., the average inverse shortest-path distance between node pairs. Higher global efficiency indicates that signals can propagate more efficiently across the cir-

[1]More details for the implementation of EAP-IG can be found in Hanna et al. (2024).

cuit. **(4) Robustness**, measured by the average *k-core* number (Seidman, 1983), which captures the depth of a circuit's densely connected core and is used as a proxy for resilience to disruption.

3.2 Experiment Design

To quantify how internal concept representations (i.e., concept circuits) relate to concept acquisition and forgetting, we formalize learning and forgetting at the knowledge-triple and concept level.

Definition 1 (Knowledge Learning/Forgetting Degree)**.** *For a conceptual knowledge triple $k = (c, r, o)$, where c is the subject concept, r is the relation, and o is the target object, the* **knowledge learning degree** *is defined as the* ***increase*** *in the logit[2] assigned to o after training relative to before training, given a textual prefix constructed from (c, r), while the* **knowledge forgetting degree** *is defined as the* ***decrease*** *in the logit assigned to o after continued training.*

Definition 2 (Concept Learning/Forgetting Degree)**.** *For a concept c with associated conceptual knowledge $\{k_0, k_1, \dots\}$, the* **concept learning (forgetting) degree** *is defined as the* ***average*** *knowledge learning (forgetting) degree across its knowledge triples: $\frac{1}{|\{k_i\}|}\sum_i \phi(k_i)$, where $\phi(k_i)$ denotes the learning (forgetting) degree of triple k_i.*

Experiment Setup. Given a pre-trained LLM π_0, we conduct a two-stage continual pre-training: **(1) Stage 1 (Concept Acquisition)**: The model is continually trained on the training set of FICO to *learn* novel concepts, yielding a new model π_1. This stage is designed to support the analysis of concept *learning* dynamics based on π_0 and π_1; (2) **Stage 2 (Forgetting Induction)**: Based on π_1 from Stage 1, we further train the model on the BIO dataset (Allen-Zhu and Li, 2023), an standard pretraining dataset used in previous LLMs knowledge acquisition studies (Allen-Zhu and Li, 2023; Zucchet et al., 2025a) , yielding a new model π_2. We then analyze concept forgetting dynamics using π_1 and π_2. We experiment with two open-source LLMs: GPT-2 Large [3] (Radford et al., 2019) and LLaMA-3.2-1B-Instruct [4] [5] (Dubey et al., 2024). For both training stages, we concatenate the prefix and target phrase, as described in Section 2 and optimize the model using the next-token prediction. For evaluation, we only provide the prefix and ask the model to generate an appropriate target phrase on the FICO test set. We use **Spearman's correlation coefficient** (Spearman, 1961) to analyze the correlation between concept learning and forgetting dynamics and concept circuit topology.

3.3 Experiment Finding

Finding 1. *Concepts exhibit substantial heterogeneity in learning degree and forgetting degree, indicating that LLMs acquire and forget different concepts to markedly different extents under the same training regime.*

Concept Learning/Forgetting Degree Distribution. Figure 2 illustrates the distribution of concept learning degrees (logit increase) and forgetting degree (logit decrease) across 500 concepts on the FICO test set. The distribution is unimodal but widely spread, indicating pronounced variability in how effectively different concepts are learned or forgotten. Practically, this variability implies that some concepts are learned more readily and robustly, whereas others require greater training effort to achieve comparable learning and to mitigate forgetting. This observation motivates our subsequent analysis, which seeks to identify indicators that can account for the differing learning and forgetting behaviors across concepts.

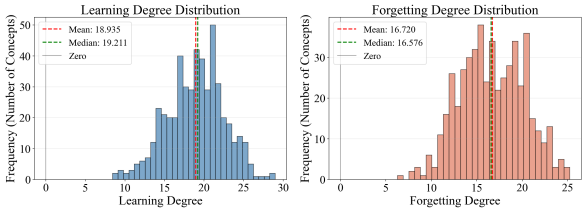


Figure 2: Distribution of learning and forgetting degree across concepts.

Finding 2. *Concept learning degree and forgetting degree shows non-trivial, statistically significant correlations[a] with multiple circuit graph metrics, suggesting that circuit structure can serve as an informative indicator of concept learning and forgetting dynamics.*

[a]The observed Spearman correlations are statistically significant (Conover, 1999) ($p < 0.001$).

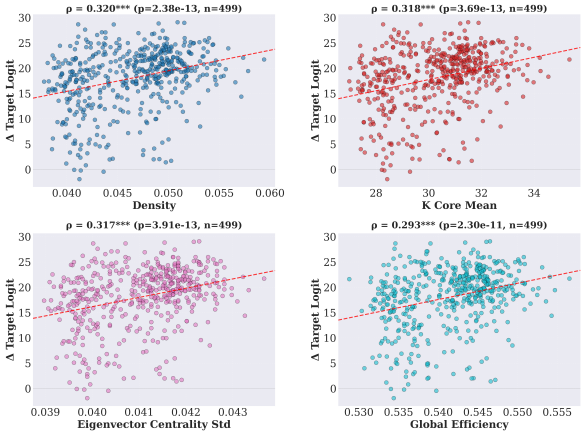
Correlation between Concept Learning/Forgetting and Concept Circuits. Figures 3(a) and 3(b)

[2]Following prior work (Hanna et al., 2024), we measure logits, a fine-grained indicator of LLM learning, and show results for log probability in Appendix C

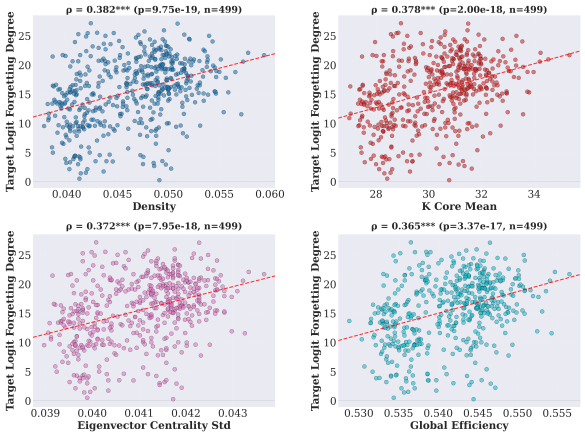
[3]openai-community/gpt2-large

[4]meta-llama/Llama-3.2-1B-Instruct

[5]We observe similar trends across both LLMs. Due to space constraints, we present results for GPT-2 in the main paper and defer the LLaMA results to Section D in Appendix



(a) Correlation between learning degree and LLM circuit pattern.



(b) Correlation between forgetting degree and LLM circuit pattern.

Figure 3: Correlation between learning dynamics and LLM circuit pattern.

examine how learning and forgetting degrees relate to the structural properties of concept circuits, as characterized by four graph metrics. We observe non-trivial and consistent Spearman correlations, indicating that circuit topology is systematically associated with how concepts are acquired and retained. For concept learning, positive correlations with metrics capturing node importance and robustness, such as eigenvector centrality and k-core, suggest that circuits with centralized bottlenecks and stable structural cores tend to achieve stronger logit gains. Similarly, positive correlations with circuit density and global efficiency indicate that structural redundancy and integrated information flow, characterized by multiple pathways and shorter distances between components, can reinforce learning signals and facilitate concept acquisition.

Notably, these same structural properties also contribute to vulnerability during continual training. Forgetting degree is positively correlated with node importance, indicating that circuits dominated by a small set of influential hub nodes are more vulnerable to forgetting when these components are perturbed by subsequent training. Surprisingly, higher circuit density, larger k-core depth, and greater global efficiency, properties that benefit learning, are likewise associated with increased forgetting. One possible explanation is that highly redundant and tightly interconnected circuits entangle concept representations more strongly with other knowledge, amplifying interference during continued training. Together, these results reveal a structural trade-off: while centralized, dense, and robust circuit organizations favor rapid and effective learning, more modular circuit structures may be better suited for mitigating forgetting under continual training. More broadly, these findings suggest that circuit-level graph metrics can serve as informative indicators of concept learning dynamics in LLMs, offering insights for concept-aware continual pretraining decisions such as allocating training effort for newly introduced concepts and previously learned concepts.

Finding 3. *During continued training on unrelated data, LLMs exhibit a consistent stage-wise temporal pattern across multiple circuit graph metrics, characterized by an early-phase change followed by gradual relaxation and stabilization.*

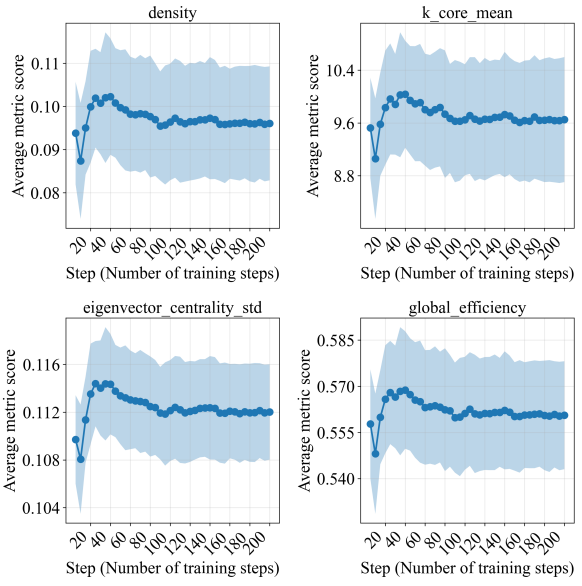


Figure 4: Graph Metric over training steps

Figure 4 shows the evolution of circuit graph metrics during the forgetting stage. LLMs exhibit a consistent stage-wise temporal pattern across

most graph metrics: an initial increase followed by a gradual decrease and eventual stabilization. This behavior is observed across all structural aspects of concept circuits, suggesting that forgetting is accompanied by systematic circuit reorganization rather than monotonic structural decay. The early increase may indicates transient entanglement of previously learned concept circuits with newly introduced knowledge during continued training, while the subsequent decrease and stabilization reflect structural relaxation and convergence to a weaker representation under interference.

Finding 4. *Concepts with larger learning degree tend to exhibit larger forgetting degree during subsequent training, indicating that knowledge acquired more aggressively is often less stable and more susceptible to interference.*

Correlation between Concept Learning and Forgetting. Figure 5 illustrates a *positive* association between learning degree and forgetting degree under continual training. That is, concepts that achieve larger gains during acquisition also tend to degrade more when the model is later trained on new data. Combined with the circuit analyses above, these results suggest a structural trade-off. During learning, concepts with larger gains are often supported by circuits that are more integrated and strongly connected, which can enable coordinated updates across concept-related components. However, the same integration may increase overlap with subsequently trained knowledge. When influence is concentrated in a small number of hubs (high eigenvector-centrality variance), perturbations to these hubs can induce circuit-wide changes, making such concepts more fragile under continued training. Overall, the learn–forget correlation suggests that stronger acquisition does not necessarily imply better consolidation, and that concept representations that are easy to strengthen may also be easier to disrupt. This observation naturally raises the question of whether such vulnerability arises in isolation or is also shaped by interactions among concurrently learned concepts.

4 How Do Interference and Synergy Arise Across Concepts and Conceptual Knowledge During Joint Training?

4.1 Interference and Synergy across Concepts

4.1.1 Experiment Design

To examine how concepts influence each other during joint training, we first construct relatedness-based groups for each target concept. We define the relatedness between two concepts as the cosine similarity between their token-level embedding representations. We obtain token embeddings for all concepts using Qwen3-Embedding-4B (Yang et al., 2025) and compute pairwise cosine similarities. For each concept c, we select the top-K most similar concepts as the highly related group, the bottom-K as the weakly related group, and the middle-K as the moderately related group. We set $K = 100$ in our experiments.We then design three joint-training configurations for each concept c, where c is trained together with (1) highly related concepts, (2) moderately related concepts, or (3) weakly related concepts, and is evaluated using two metrics: (1) the average logit assigned to the target objects for knowledge triplets associated with concept c, and (2) the corresponding average probability. This procedure is repeated for all concepts in test set of FICO to obtain a comprehensive characterization of cross-concept interactions.

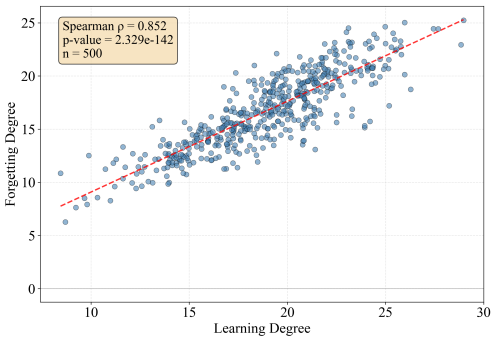


Figure 5: Spearman Correlations between learning and forgetting of concepts

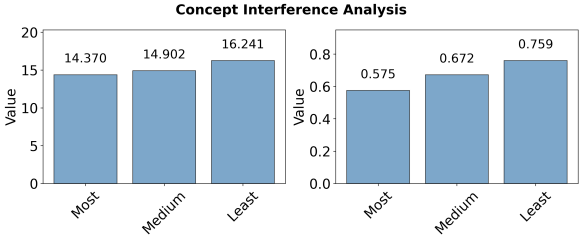


Figure 6: Concept interference under joint training.

4.1.2 Experiment Finding

Finding 5. *Training with highly related concepts yields lower performance than training with weakly related concepts, indicating stronger interference among semantically similar concepts during joint learning.*

Figure 6 shows a clear and consistent dependence on semantic relatedness. Across both evaluation metrics (average logit and average probability), training with weakly related concepts achieves the highest performance (75.9%), substantially outperforming training the highly related (57.5%) and moderately related (67.2%) concepts. These results demonstrate that cross-concept interactions meaningfully affect how effectively a target concept can be acquired in a multi-concept setting.

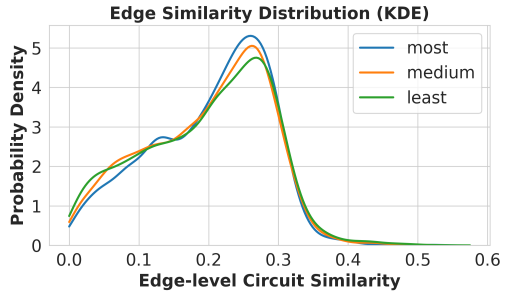


Figure 7: Jaccard similarity across concept circuits

To understand the underlying mechanism driving this effect, we analyze the degree of overlap between the internal representations of co-trained concepts. For each target concept c, we pair it with each of its related concepts drawn from three semantic similarity groups: highly related, moderately related, and weakly related. For each pair (c, c'), we obtain their concept circuits and compute the Jaccard similarity between their edge sets. Figure 7 visualizes the resulting similarity distributions using KDE plots. The highly related setting exhibits a sharper, higher-density peak at moderate circuit similarity, indicating consistent reuse of overlapping computational pathways between target and auxiliary concepts, which likely induces sustained representational competition during joint training and results in stronger interference. Instead, the moderately and weakly related settings show lower and more dispersed similarity distributions, reflecting reduced circuit overlap and correspondingly weaker interference. This observation motivates interference-aware data scheduling that reduces circuit similarity among co-trained concepts within the same batch.

4.2 Interference and Synergy across Knowledge

4.2.1 Experiment Design

Beyond concepts, we ask whether *different types of conceptual knowledge* can also exhibit interference or synergy, even when they describe the *same* concept. To enable this analysis, we focus on five high-level semantic knowledge categories: (1) Hyponym & Hypernym (HAH), (2) Synonym & Antonym (SAA), (3) Meronym & Holonym (MAH), (4) Property & Affordance (PAA), and (5) Spatial Relation (SR), as shown in Section A in Appendix. To study synergy across knowledge types, we adopt a pairwise continual-training setup. In Stage 1, the model is pretrained either on knowledge category R_i or on BIO dataset (Allen-Zhu and Li, 2023), which does not contain conceptual knowledge, for the same training steps. In Stage 2, the model is continually trained on knowledge category R_j. Across the five knowledge categories, this results in $5 \times 4 = 20$ ordered curricula, along with the BIO-based control baseline. We train LLMs on each curriculum and evaluate its performance for target knowledge category R_j, to quantify which type transitions produce synergy or interference. We define **paired transferability** from R_i to R_j as

$$T(R_i \to R_j) = \frac{\text{logit}(R_j \mid R_i) - \text{logit}(R_j \mid \text{BIO})}{|\text{logit}(R_j \mid \text{BIO})|}. \quad (1)$$

Positive $T(R_i \to R_j)$ indicates *synergy*, where learning R_i facilitates subsequent learning of R_j, while negative values indicate *interference*.

4.2.2 Experiment Finding

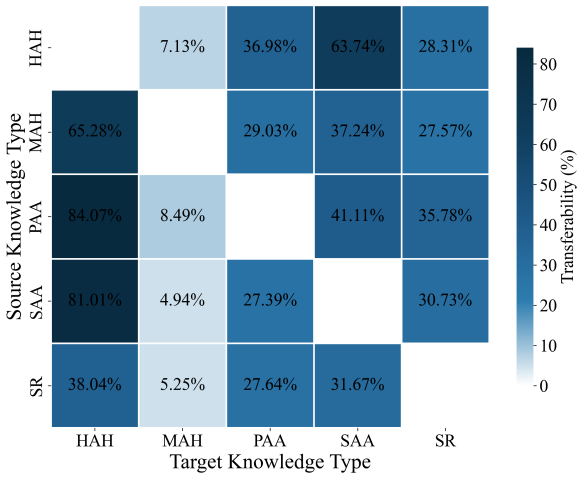


Figure 8: Paired transferability across knowledge

Finding 6. *Substantial and asymmetric transfer effects emerge across knowledge types, where pretraining on one type can facilitate learning of another, with highly directional and uneven benefits across ordered pairs.*

Figure 8 provides a fine-grained view of pairwise synergy among the different knowledge types, where each cell shows the paired transferability $T(R_i \rightarrow R_j)$ when a *source* R_i type is trained earlier before a *target* R_j knowledge type. While all five categories encode concept-related information, they capture distinct semantic facets, leading to non-redundant and uneven transfer behaviors. The heatmap reveals heterogeneous and directional effects: for example, pretraining on Property & Affordance (PAA) yields substantial gains when transferring to Hyponym & Hypernym (HAH) and Synonym & Antonym (SAA), suggesting that learning functional and attribute-level regularities can effectively scaffold the acquisition of more abstract relational structures. In contrast, the reverse directions exhibit markedly weaker transfer, highlighting the asymmetry of these interactions. Moreover, some target types, such as Meronym & Holonym (MAH), show relatively small improvements across most sources, indicating higher intrinsic learnability and reduced sensitivity to prior knowledge-type pretraining. Overall, these findings suggest practical guidance for future training curricula, such as reordering training data to place knowledge types with strong positive transfer earlier, thereby encouraging synergy and improving the efficiency of downstream concept learning.

5 Related Work

LLM Knowledge Acquisition. A growing body of work studies learning mechanism behinds LLM knowledge acquisition. One line uses synthetic or fictional corpora—e.g., biographies of fabricated individuals (Allen-Zhu and Li, 2023; Zucchet et al., 2025a; Ou et al., 2025; Zhu et al., 2025; Feng et al., 2024), Wikipedia-style entries for fictional entities (Chang et al., 2024), or **post-cutoff data** (Huang et al., 2024)—to examine how data properties, training choices, and curricula (e.g., ordering dependencies between facts and implications) affect learning. Another line (Chang et al., 2024; Xu et al., 2024b; Leybzon and Kervadec, 2024; Im and Li, 2024; Qian et al., 2024; Tigges et al., 2024) leverages mechanistic interpretability to analyze LLM learning by tracing how internal representations evolve across training stages. The third line (Ren and Sutherland, 2024; Chen et al., 2023; Jain et al., 2023) identify phase transitions that reveal discrete, objective-dependent shifts in model behavior during training. In contrast, we target **concept-level** knowledge by adapting ConceptNet to construct **fictional concepts** that preserve the relational structure of real concepts. Moreover, rather than treating injected items as independent facts, we explicitly model the structured relations *within* each concept and *across* concepts.

LLM Concept Probing and Editing. Prior work studies conceptual knowledge in LLMs mainly through: (i) **prompt-based probing** of concept properties and relations (Gu et al., 2023a; Liao et al., 2023; Shani et al., 2023b; Zheng et al., 2024; Peng et al., 2022); (ii) **definition–name alignment** tests (e.g., dictionary/reverse-dictionary probes) that assess mapping between descriptions and names (Xu et al., 2024a); and (iii) **compositional binding and consistency** evaluations that test correct attribution of concept knowledge to instances and consistency across hierarchies (He et al., 2023; Sosa et al., 2024; Sahu et al., 2022). Complementing these behavioral probes, mechanistic interpretability aims to localize internal components responsible for concept-related behavior (Aljaafari et al., 2024; Wang et al., 2024c). Unlike these largely static evaluations, we focus on the **dynamics of concept learning**: we connect internal *concept circuits* to acquisition, forgetting, and cross-concept interactions, providing a mechanistic, time-resolved complement to existing probing approaches.

6 Conclusion

In this work, we present a unified analysis of how large language models acquire, retain, and forget concepts during continual pretraining by integrating output behavior with circuit-level interpretability. We extend the study of knowledge learning beyond isolated facts to the structured interrelations among knowledge within and across concepts, and how these relations drive interference and synergy. To support this analysis, we introduce the FICO dataset, define **Concept Circuits** as circuit-level representations of concepts, and apply **Graph Metrics** to characterize their structural patterns. Our experimental findings reveal systematic dynamics in concept learning and cross-concept interactions, offering a foundation for developing concept-aware training schedules and moving toward more interpretable and reliable continual pretraining procedures for LLMs.

Limitations

Despite conducting extensive analyses of concept acquisition, forgetting, and cross-concept interactions in LLMs, our study has several limitations. (1) **Limited Model Scale**. Due to computational constraints, our experiments focus on GPT-2 Large (0.7B) and LLaMA-3.2-1B, and we do not evaluate larger-scale models. Extending our analysis to larger LLMs remains an important direction for future work. (2) **Exploration of Actionable Training Strategies**. Given the analytical scope of this work, we focus on characterizing the relationship between internal concept circuits and learning dynamics, showing that circuit graph patterns could indicate future learning and forgetting behaviors, and that circuit similarity may signal potential interference. While our findings may inform training strategies like circuit-aware training effort allocation and interference-aware data scheduling, we leave the exploration of these motivated training strategies for future work.

References

Nura Aljaafari, Danilo S Carvalho, and André Freitas. 2024. The mechanics of conceptual interpretation in gpt models: Interpretative insights. *arXiv preprint arXiv:2408.11827*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2023. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*.

William Jay Conover. 1999. *Practical nonparametric statistics*. john wiley & sons.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

J. Feng, S. Russell, and J. Steinhardt. 2024. Extractive structures learned in pretraining enable generalization on finetuned facts. *arXiv preprint arXiv:2412.04614*.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2023a. Do language models have coherent mental models of everyday things? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1892–1913.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2023b. Do language models have coherent mental models of everyday things? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1892–1913, Toronto, Canada. Association for Computational Linguistics.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*.

Marcel Harpaintner, Natalie M Trumpp, and Markus Kiefer. 2018. The semantic content of abstract concepts: A property listing study of 296 abstract words. *Frontiers in psychology*, 9:1748.

Yuan He, Jiaoyan Chen, Ernesto Jimenez-Ruiz, Hang Dong, and Ian Horrocks. 2023. Language model analysis for ontology subsumption inference. *arXiv preprint arXiv:2302.06761*.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. 2023. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580.

J. Huang, D. Yang, and C. Potts. 2024. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shawn Im and Yixuan Li. 2024. Understanding the learning dynamics of alignment with human feedback. *arXiv preprint arXiv:2403.18742*.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*.

Vito Latora and Massimo Marchiori. 2001. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701.

Danny D Leybzon and Corentin Kervadec. 2024. Learning, forgetting, remembering: Insights from tracking llm memorization during training. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 43–57.

Jiayi Liao, Xu Chen, and Lun Du. 2023. Concept understanding in large language models: An empirical study.

MARK EJ Newman. 2010. Networks: an introduction.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do LLMs acquire new knowledge? a knowledge circuits perspective on continual pretraining. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913, Vienna, Austria. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. 2024. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. *arXiv preprint arXiv:2402.19465*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Yi Ren and Danica J Sutherland. 2024. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*.

Pritish Sahu, Michael Cogswell, Yunye Gong, and Ajay Divakaran. 2022. Unpacking large language models with conceptual consistency. *arXiv preprint arXiv:2209.15093*.

Stephen B Seidman. 1983. Network structure and minimum degree. *Social networks*, 5(3):269–287.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. 2023a. Towards concept-aware large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170, Singapore. Association for Computational Linguistics.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. 2023b. Towards concept-aware large language models. *arXiv preprint arXiv:2311.01866*.

Rosario Uceda Sosa, Karthikeyan Natesan Ramamurthy, Maria Chang, and Moninder Singh. 2024. Reasoning about concepts with llms: Inconsistencies abound. In *First Conference on Language Modeling*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. Llm circuit analyses are consistent across training and scale. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 40699–40731.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024a. Knowledge mechanisms in large language models: A survey and perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7097–7135, Miami, Florida, USA. Association for Computational Linguistics.

Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. 2024b. Editing conceptual knowledge for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 706–724.

Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. 2024c. Editing conceptual knowledge for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 706–724, Miami, Florida, USA. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Menghan Zhang, Peng Qian, and Xuanjing Huang. 2024a. On the tip of the tongue: Analyzing conceptual representation in large language models with reverse-dictionary probe. *arXiv preprint arXiv:2402.14404*.

Yang Xu, Yi Wang, and Hao Wang. 2024b. Tracking the feature dynamics in llm training: A mechanistic study. *arXiv preprint arXiv:2412.17626*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2025. Knowledge circuits in pretrained transformers. *Preprint*, arXiv:2405.17969.

Junhao Zheng, Shengjie Qiu, and Qianli Ma. 2024. Concept-1k: A novel benchmark for instance incremental learning. *arXiv e-prints*, pages arXiv–2402.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025. Effective LLM knowledge learning requires rethinking generalization.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. 2025a. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*.

Nicolas Zucchet, Jorg Bornschein, Stephanie C.Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. 2025b. How do language models learn facts? dynamics, curricula and hallucinations. In *Second Conference on Language Modeling*.

A Knowledge Type Grouping and Filtering

Our dataset is constructed from the ConceptNet knowledge graph in which each knowledge instance is represented as a triplet $\langle \text{subject}, \text{relation}, \text{object} \rangle$. Since the original graph contains a large number of fine-grained relation types, we group them into a smaller set of semantically coherent, high-level knowledge categories to facilitate analysis. Each relation is mapped to a category based on its primary semantic function. As shown in Table 1, taxonomic and definitional relations (*IsA*, *DefinedAs*, *FormOf*, *InstanceOf*) are grouped into **Hyponym and Hypernym**. Relations expressing similarity or contrast (*Synonym*, *SimilarTo*, *Antonym*, *DistinctFrom*) are grouped into **Synonym and Antonym**. Part-whole and compositional relations (*PartOf*, *HasA*, *MadeOf*) are grouped into **Meronym and Holonym**. Relations describing properties, affordances, or functional roles (*HasProperty*, *UsedFor*, *CapableOf*, *ReceivesAction*) are grouped into **Property and Affordance**. Spatial relations (*AtLocation*, *LocatedNear*) are grouped into **Spatial Relation**.

We retain these five concept-centric knowledge types in our experiments, as they capture stable semantic properties of real-world concepts. We filter out the **Causality Event**, **Lexical/Etymological**, and **Other** categories, as they primarily encode events, linguistic form, or noisy relations that are not well suited for modeling real-world concepts.

B Dataset Statistics

As shown in Table 2, following dataset construction, we leverage 1,000 concepts for training and 500 concepts for testing. The training set contains 3,075 knowledge triples, which are instantiated into 92,250 training samples with different templates, comprising approximately 1.04M tokens in total. We train LLMs on our dataset for 10 epochs, leading to total training tokens of 10.4M, similar as previous continual-pretraining work (Ou et al., 2025). The test set consists of 1,586 knowledge triples and corresponding evaluation samples, totaling 13,530 tokens.

C Results for Log Probability

We show results for log probability in Figure 9, Figure 10(a) and Figure 10(b) and Figure 11.

High-level Knowledge Type	Relation Types
Hyponym and Hypernym	*IsA*, *DefinedAs*, *FormOf*, *InstanceOf*
Synonym and Antonym	*Synonym*, *SimilarTo*, *Antonym*, *DistinctFrom*
Meronym and Holonym	*PartOf*, *HasA*, *MadeOf*
Property and Affordance	*HasProperty*, *UsedFor*, *CapableOf*, *ReceivesAction*
Spatial Relation	*AtLocation*, *LocatedNear*
Excluded Relation Categories	
Causality & Event	*Causes*, *MotivatedByGoal*, *HasPrerequisite*, *HasSubevent*, *HasFirstSubevent*, *HasLastSubevent*, *CreatedBy*
Desire	*Desires*, *CausesDesire*
Lexical / Etymological	*DerivedFrom*, *EtymologicallyDerivedFrom*, *EtymologicallyRelatedTo*
Other	*RelatedTo*, *HasContext*, *ExternalURL*, *SymbolOf*

Table 1: Mapping from fine-grained relation types to high-level knowledge categories. We retain five concept-centric categories for experiments and exclude relation types that primarily encode events, lexical form, or noisy contextual associations.

Data	Train	Test
# Concepts	1,000	500
# Knowledges	3,075	1,586
# Samples	92,250	1,586
# Tokens	1,039,784	13,530

Table 2: Dataset statistics of FICO.

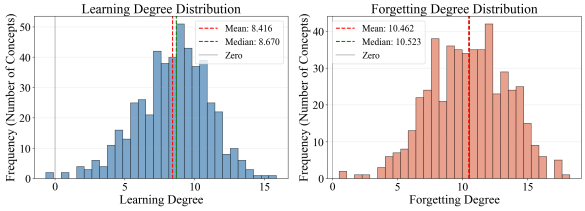


Figure 9: Distribution of learning and forgetting degree across concepts.

D LLaMA Results

D.1 RQ1 Results

We show LLaMA results on RQ1 in Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, and Figure 17.

D.2 RQ2 Results

We show LLaMA results on RQ2 in Figure 18 and Figure 19.

E Experiment Details

We conduct experiments on 4×40GB A40 GPUs. We use learning rate as 5×10^{-5} and batch size as 128.

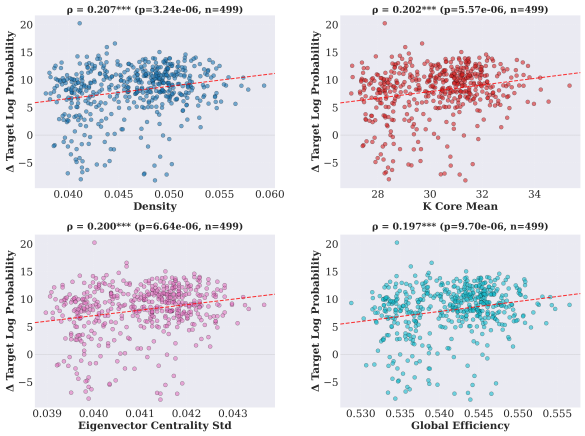
F Potential Risks

We carefully follow the ACM Code of Ethics [6] and have not found potential societal impacts or risks so far. To the best of our knowledge, this work has no notable harmful effects and uses, environmental impact, fairness considerations, privacy considerations, security considerations, or other potential risks. Our dataset does not contain any information that names or uniquely identifies individual people or offensive content.

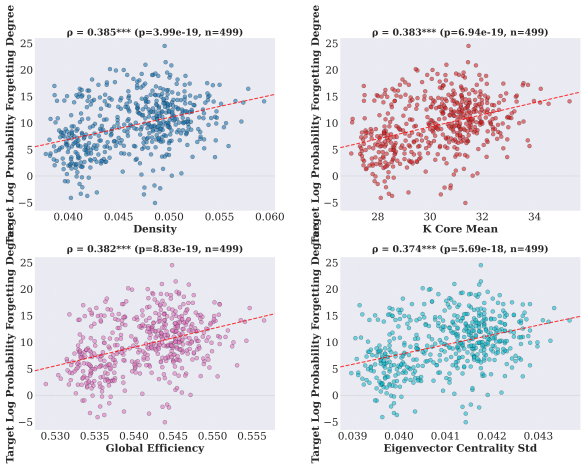
G Use Of AI Assistants

We used AI assistants to help with writing and editing the manuscript. The assistants were used to refine wording, improve clarity, and enhance overall presentation.

[6] https://www.aclweb.org/portal/content/acl-code-ethics



(a) Correlation between learning degree and LLM circuit pattern.



(b) Correlation between forgetting degree and LLM circuit pattern.

Figure 10: Correlation between learning dynamics and LLM circuit pattern.

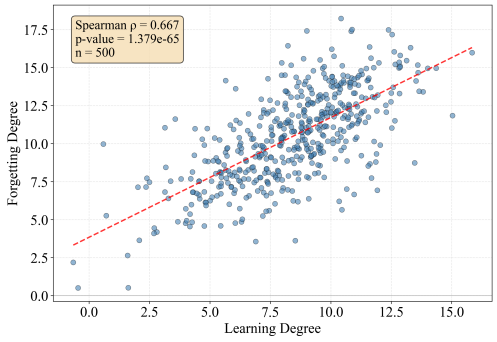


Figure 11: Spearman Correlations between learning and forgetting of concepts

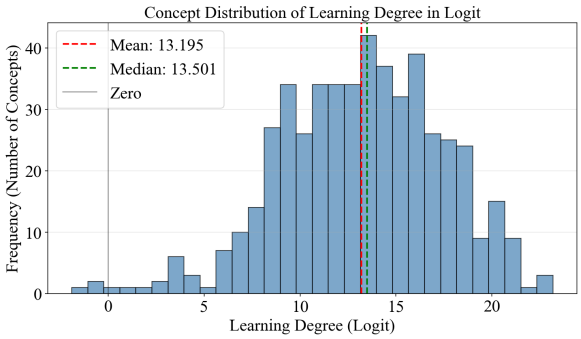


Figure 12: Learning degree distribution across concepts.

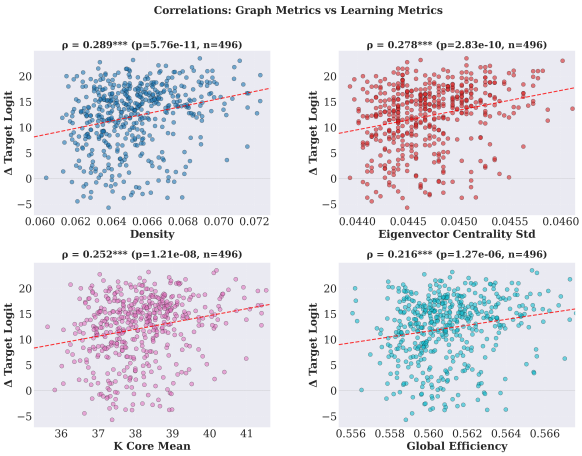


Figure 13: Correlation between learning degree and LLM circuit graph scores.

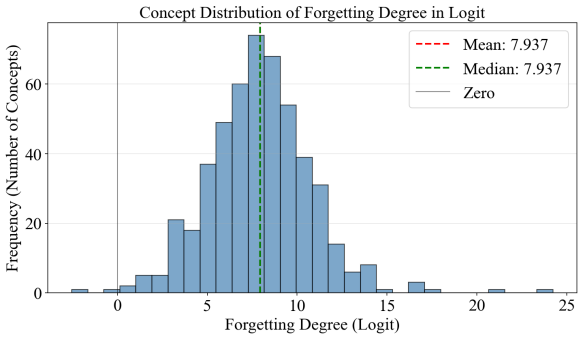


Figure 14: Distribution of forgetting degree across concepts.

Figure 15: Correlation between forgetting degree and LLM circuit graph scores.

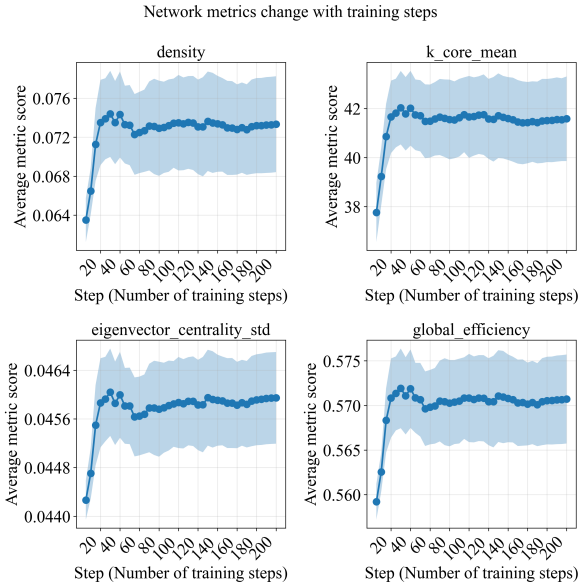


Figure 16: Graph Metric over training steps

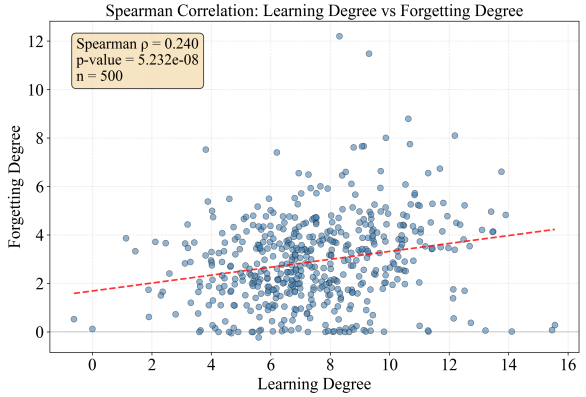


Figure 17: Correlations between learning and forgetting of concepts

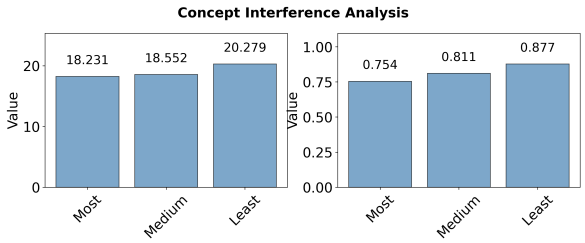


Figure 18: Concept-level interference under joint training.

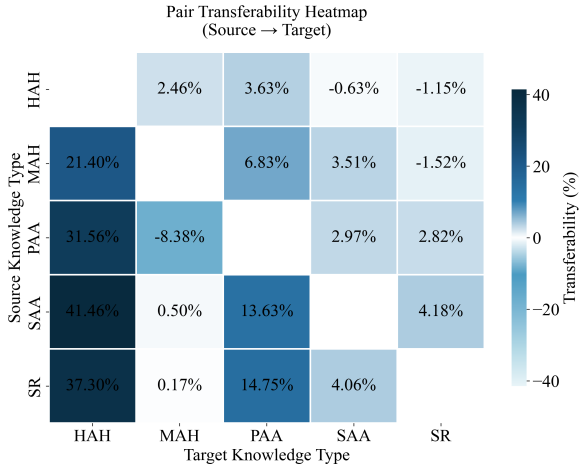


Figure 19: Pairwise dependency between knowledge types