

Knowledge in Superposition: Unveiling the Failures of Lifelong Knowledge Editing for Large Language Models

Anonymous submission

Abstract

Knowledge editing aims to update outdated or incorrect knowledge in large language models (LLMs). However, current knowledge editing methods have limited scalability for lifelong editing¹. This study explores the fundamental reason why knowledge editing fails in lifelong editing. We begin with the closed-form solution derived from linear associative memory, which underpins state-of-the-art knowledge editing methods. We extend the solution from single editing to lifelong editing, and through rigorous mathematical derivation, identify an interference term in the final solution, suggesting that editing knowledge may impact irrelevant knowledge. Further analysis of the interference term reveals a close relationship with superposition between knowledge representations. When knowledge superposition does not exist in language models, the interference term vanishes, allowing for lossless knowledge editing. Experiments across numerous language models reveal that knowledge superposition is universal, exhibiting high kurtosis, zero mean, and heavy-tailed distributions with clear scaling laws. **Ultimately, by combining theory and experiments, we demonstrate that knowledge superposition is the fundamental reason for the failure of lifelong editing.** Moreover, this is the first study to investigate knowledge editing from the perspective of superposition and provides a comprehensive observation of superposition across numerous real-world language models.

Introduction

In large language models (LLMs), outdated or incorrect knowledge may persist (Radford et al. 2019; Wang and Komatsuzaki 2022; Biderman et al. 2023; Touvron et al. 2023). However, retraining these models to update knowledge incurs prohibitively high costs. To address this problem, knowledge editing (De Cao et al. 2021; Mitchell et al. 2021) is introduced to edit specific knowledge by directly updating the internal parameters of language models.

Despite achieving significant progress in single editing, where knowledge is edited only once, knowledge editing should be continuous in fact. However, current methods struggle with scalability for lifelong editing (Huang et al. 2023), where continuous editing and performance monitoring are required throughout the language models’ lifecycle. For example, using the representative editing methods

¹Lifelong editing means lifelong knowledge editing.

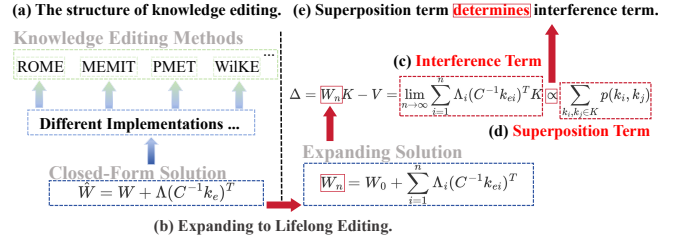


Figure 1: Illustration of our work. (a) Current knowledge editing methods use a unified closed-form solution, which means adding $\Lambda(C^{-1}k_e)^T$ to parameters matrix W to achieve knowledge updating. (b) Extend the closed-form solution to lifelong editing, where W_n represents parameters matrix after the n -th edit. (c) Interference term accumulating sufficiently will cause language models to forget knowledge. (d) Superposition term, where $p(\cdot, \cdot)$ denotes the degree of superposition between two knowledge representations. (e) Superposition term actually determines interference term.

ROME (Meng et al. 2022a) or MEMIT (Meng et al. 2022b) for lifelong editing results in severe performance degradation after dozens or hundreds of editing steps, rendering the edited models unusable (Hu et al. 2024). However, the underlying reason has been rarely explored.

In this paper, we explore the fundamental reason why knowledge editing fails in lifelong editing. We start from the closed-form solution for knowledge editing derived by (Meng et al. 2022a) through linear associative memory, which is the foundation of the state-of-the-art knowledge editing methods (Figure 1a), including ROME (Meng et al. 2022a), MEMIT (Meng et al. 2022b), PMET (Li et al. 2023), WiKE (Hu et al. 2024) and so on. Specifically, we extend the closed-form solution from single editing to lifelong editing (Figure 1b). Our rigorous mathematical derivation reveals that the extended solution introduces an **interference term** (Figure 1c), which, when accumulated sufficiently, leads language models to forgetting the knowledge, including original knowledge and previously edited knowledge.

Further analysis of the interference term reveals a close relationship with superposition: if superposition does not exist in language models, the superposition term (Figure 1d) is zero, causing the interference term to vanish (Figure 1e),

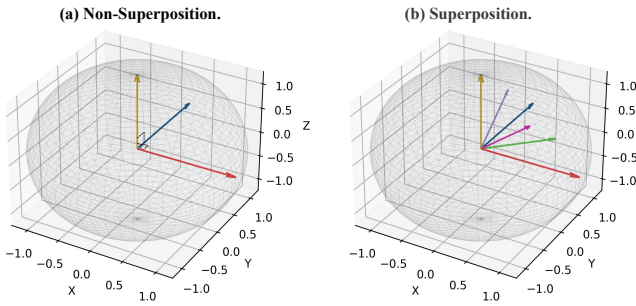


Figure 2: A neural network with only three neurons, corresponding to three dimensions, (a) can directly represent three features orthogonally, but (b) to represent six features (or more features), it will use superposition to noisily encode them nearly orthogonally.

allowing for lossless lifelong editing. Superposition (Elhage et al. 2022b) refers to the situation where neural networks attempt to represent more features than the available dimensions. For example, if a simple neural network with three neurons (corresponding to a three-dimensional space) attempts to represent three features, each neuron can be assigned to one feature, forming an orthogonal basis in three-dimensional space (Figure 2a). But if the same network tries to represent six features (or more features), it will use superposition strategy, noisily encoding these features, where the directions corresponding to each feature will not be fully orthogonal (Figure 2b). Specifically, we find that the magnitude of interference term depends on the orthogonality of knowledge representations (Figure 1e). If these knowledge representations are perfectly orthogonal (**non-superposition**), the interference term is zero, enabling perfectly lossless knowledge editing, where only the target knowledge is updated without affecting unrelated knowledge. However, if they are not orthogonal (**superposition**), the interference will accumulate linearly and eventually tend toward infinity, leading to language models’ failure. **In other words, whether superposition exists in language models is equivalent to whether lossless knowledge editing can be achieved, which will determine whether lifelong editing can be achieved.**

However, to what extent does superposition hold true in real-world language models? **Our experiments reveal that knowledge superposition² is universal across all language model families, characterized by high kurtosis, zero mean, heavy-tailed distributions, with a clear scaling law.**

Specifically, we conduct experiments on language model families including GPT-2 (Radford et al. 2019), Llama-2 (Touvron et al. 2023), Pythia (Biderman et al. 2023), GPT-J (Wang and Komatsuzaki 2022), Llama-3 (Meta 2024a), and Llama-3.1 (Meta 2024b), to capture the real situation of knowledge superposition. Firstly, we observe that knowledge superposition exists across all layers of all language

²Knowledge superposition refers to the superposition of knowledge representations in language models.

models we examined (Figure 3), indicating that superposition is widespread in real-world language models. Secondly, the knowledge superposition exhibits a high kurtosis, zero-mean heavy-tailed distribution (Figure 4). This means that the distribution consistently has a very high peak near zero, which corresponds to the positions where all knowledge representations are perfectly orthogonal, suggesting models attempt to store different knowledge orthogonally but resort to stores them nearly orthogonally due to capacity constraints, i.e., through superposition. Thirdly, we observe a clear scaling law (Figure 5): as the size of language models increases, they attempt to represent knowledge in a more orthogonal manner, thereby reducing noise between knowledge representations, which explains why larger language models can more effectively handle knowledge-intensive question-answering scenarios. Finally, to provide further direct evidence of superposition, we calculate the angular distribution between these knowledge representations in high-dimensional space, which also shows a high kurtosis, heavy-tailed distribution, but centered around 90 degrees (Figure 6).

Overall, our contributions are as follows:

- We extend the closed-form solution for knowledge editing from single editing to lifelong editing. Our rigorous derivation reveals an interference term in the final solution, indicating that editing can affect both original and previously edited knowledge.
- We conduct further analysis of the interference term, which actually reflects knowledge superposition. The magnitude of this interference is determined by the superposition of knowledge representations, with non-superposition enabling lossless knowledge editing.
- We investigate the phenomenon of knowledge superposition across multiple language model families and find it universal, exhibiting high kurtosis, zero mean, heavy-tailed distributions, with a clear scaling law. To our knowledge, this is the first observation of widespread knowledge superposition.

Related Work

Knowledge Editing

In general, knowledge editing aims to modify the knowledge of a language model so that its outputs reflect the revised state when confronted with relevant inputs (De Cao et al. 2021). Yao et al. (2023) survey knowledge editing methods and classify them into two main categories: preserving model parameters and modifying model parameters. Preserving model parameters includes memory-based methods (Mitchell et al. 2022; Zhong et al. 2023; Hartvigsen et al. 2024), which stores all edit examples explicitly in memory, and additional parameter methods (Dong et al. 2022; Huang et al. 2023), which use extra trainable parameters within language models. However, these methods usually have poor generalization, and the required additional content also increases significantly with the number of editing. In contrast, modifying model parameters methods directly update the model parameters for editing, thereby avoiding these issues.

Such methods are also divided into two categories: meta-learning and locate-then-edit methods. Meta-learning methods (De Cao et al. 2021; Mitchell et al. 2021; Tan, Zhang, and Fu 2023), which use a hyper network to learn and apply gradients for fine-tuning, typically require a lengthy hyper network training process before each edit, making them less effective for lifelong editing. Locate-then-edit methods first identify and then update the parameters associated with specific knowledge. For instance, KnowledgeNeuron (Dai et al. 2021) uses knowledge attribution to locate neurons and updates parameters to achieve knowledge editing; ROME (Meng et al. 2022a) employs causal mediation analysis to identify the center of causal effects, namely MLP, and performs rank-one editing on MLP; MEMIT (Meng et al. 2022b) extends ROME by allocating residuals to multiple layers and enabling batch editing; PMET (Li et al. 2023) further refines residual allocation based on MEMIT; and WilKE (Hu et al. 2024) dynamically selects editing layers based on ROME to avoid toxicity flash.

This paper focuses on methods that modify model parameters, among which the state-of-the-art methods are ROME and its derivatives, MEMIT, PMET, and WilKE. Therefore, we concentrate on the same closed-form solution (Figure 1a) of them and extend it from single to lifelong editing. From a theoretical perspective, we identify the fundamental reason for the failure of knowledge editing methods in lifelong editing, namely superposition.

Superposition

Superposition refers to a strategy used by neural networks to express features that far exceed their dimensions, typically assigning approximately orthogonal directions in the representation space to these features (Figure 2b). However, this superposition phenomenon has only been observed in toy models so far (Elhage et al. 2022b). Some studies hypothesize that this superposition phenomenon also exists in language models and propose sparse autoencoder (SAE) (Cunningham et al. 2023) and its various variants (Rajamanoharan et al. 2024; Gao et al. 2024) to attempt to disentangle superposition from the activation space of language models (Bricken et al. 2023; Templeton et al. 2024). Additionally, Gurnee et al. (2023) claim to observe superposition in the wild, their study was limited to a few neurons in a specific layer of a single model, focusing on neuron superposition. This differs from the superposition concept in Elhage et al. (2022b) and our study.

In this paper, we identify a widespread phenomenon of knowledge superposition across multiple language model families and explore its characteristics. Furthermore, this is the first study of knowledge superposition in knowledge editing, explaining the reason for failures in lifelong editing from the perspective of superposition. Notably, unlike previous studies, we find that superposition occurs in the whitening space rather than the activation space (Figure 6).

Preliminary

Geva et al. (2020) discover that MLPs are the key components for memory storage in Transformers, and Meng et al.

(2022a) identify through causal tracing that MLPs are crucial for storing factual knowledge. The MLPs in the Feed-Forward Network (FFN) of a Transformer consist of two layers, represented as follows (bias terms are omitted)

$$FFN(\mathbf{x}) = W_{proj} \sigma(W_{fc} \mathbf{x}), \quad (1)$$

where $W_{fc} \in \mathbb{R}^{d_m \times d}$ and $W_{proj} \in \mathbb{R}^{d \times d_m}$ are the parameter matrices of the FFN, d_m is the dimensionality of the hidden layer in the FFN, σ is the activation function, and $\mathbf{x} \in \mathbb{R}^d$ is the input of the FFN.

Meng et al. (2022a) model MLPs as linear associative memory (Kohonen 1972; Anderson 1972), viewing W_{proj} as a linear associative store. This perspective observes that any linear operation W can be represented as a key-value store with a set of vector keys $K = [k_1 | k_2 | \dots]$ and corresponding vector values $V = [v_1 | v_2 | \dots]$. To optimally insert a new key-value pair (k_e, v_e) into the store for knowledge updates, one can solve a constrained least-squares problem, leading to a closed-form solution

$$\begin{aligned} & \text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_e = v_e \\ & \text{by setting } \hat{W} = W + \Lambda(C^{-1}k_e)^T, \end{aligned} \quad (2)$$

where $C = KK^T$ is a constant that actually represents the covariance matrix, and $\Lambda = (v_e - Wk_e)/(C^{-1}k_e)^T k_e$. The detailed proof can be found in the original paper (Meng et al. 2022a), and a more thorough derivation is provided in Appendix A.

Expanding to Lifelong Editing

For convenience, let the initial W_{proj} be W_0 . Following the first update using (k_{e_1}, v_{e_1}) , we obtain W_1 . Subsequently, after the n -th edit with (k_{e_n}, v_{e_n}) , we derive W_n . Extending the closed-form solution from Equation 2 to the lifelong editing scenario, we obtain

$$W_n = \begin{cases} W_0, & n = 0 \\ W_{n-1} + \Lambda_n(C^{-1}k_{e_n})^T, & n \geq 1 \end{cases} \quad (3)$$

where $C = KK^T$, and

$$\Lambda_n = \frac{v_{e_n} - W_{n-1}k_{e_n}}{(C^{-1}k_{e_n})^T k_{e_n}}, n \geq 1. \quad (4)$$

Expanding W_n , we obtain

$$W_n = W_0 + \sum_{i=1}^n \Lambda_i(C^{-1}k_{e_i})^T. \quad (5)$$

For convenience, we let

$$K^* = \begin{bmatrix} K_o^* & | & K_e^* \end{bmatrix}, \quad (6)$$

where $K_o^* = [k_1 | k_2 | \dots]$ represents the keys corresponding to the model's original knowledge, and $K_e^* = [k_{e_1} | k_{e_2} | \dots | k_{e_n}]$ represents the keys for the edited knowledge.

We then calculate V^* after W_0 has been updated to W_n , which is given by

$$\begin{aligned}
V^* &= W_n K^* \\
&= (W_0 + \sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T) K^* \\
&= W_0 K^* + \left(\sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T \right) K^* \\
&= W_0 K^* + \sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T K^*. \tag{7}
\end{aligned}$$

For convenience, we can represent V^* using a block matrix notation

$$V^* = \begin{bmatrix} V_o^* & | & V_e^* \end{bmatrix}, \tag{8}$$

where V_o^* and V_e^* represent the updated value matrices corresponding to the original knowledge and the edited knowledge, respectively. We then analyze these two components separately.

Interference Term of Original Knowledge Δ_o

Ideally, the value vectors corresponding to the original knowledge after editing, namely V_o^* , are same as $V_o = [v_1 | v_2 | \dots]$, because we want to ensure that editing does not affect unrelated original knowledge. Thus, we define Δ_o as

$$\begin{aligned}
\Delta_o &= V_o^* - V_o \\
&= \left[\sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T k_1 \mid \sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T k_2 \mid \dots \right]. \tag{9}
\end{aligned}$$

For convenience, we study $\Delta_o[:, j]$, which represents the j -th column vector of Δ_o . This can be viewed as the interference term introduced to the j -th piece of original knowledge after n edits,

$$\begin{aligned}
\Delta_o[:, j] &= \sum_{i=1}^n \Lambda_i (C^{-1} k_{e_i})^T k_j \\
&= \sum_{i=1}^n \frac{v_{e_i} - W_{i-1} k_{e_i}}{(C^{-1} k_{e_i})^T k_{e_i}} (C^{-1} k_{e_i})^T k_j \\
&= \sum_{i=1}^n \frac{(C^{-1} k_{e_i})^T k_j}{(C^{-1} k_{e_i})^T k_{e_i}} (v_{e_i} - W_{i-1} k_{e_i}). \tag{10}
\end{aligned}$$

Next, let $\delta(v_{e_i}) = v_{e_i} - W_{i-1} k_{e_i} \in \mathbb{R}^d$ represent the difference vector between the optimized value and the current value when editing the i -th piece of knowledge, and define the coefficient $p(k_{e_i}, k_j) = \frac{(C^{-1} k_{e_i})^T k_j}{(C^{-1} k_{e_i})^T k_{e_i}} \in \mathbb{R}$. Then, we obtain

$$\Delta_o[:, j] = \sum_{i=1}^n p(k_{e_i}, k_j) \delta(v_{e_i}). \tag{11}$$

Ideally, we want $\Delta_o[:, j] = \mathbf{0}$ to achieve lossless knowledge editing, ensuring that the original knowledge of the

model remains unaffected. However, since $\delta(v_{e_i}) = v_{e_i} - W_{i-1} k_{e_i} \neq \mathbf{0}$ (as this is the premise; if $\delta(v_{e_i}) = \mathbf{0}$, then there is no need for editing), we can only hope that $p(k_{e_i}, k_j) = 0$. This would eliminate the interference on the j -th original knowledge.

Interference Term of Edited Knowledge Δ_e

Ideally, the value vectors corresponding to the edited knowledge after editing, V_e^* , are same as $V_e = [v_{e_1} | v_{e_2} | \dots | v_{e_n}]$, because we want to avoid affecting the already edited knowledge during further editing. Thus, we define Δ_e as

$$\begin{aligned}
\Delta_e &= V_e^* - V_e \\
&= \left[\sum_{i=2}^n \Lambda_i (C^{-1} k_{e_i})^T k_{e_1} \mid \dots \mid \Lambda_n (C^{-1} k_{e_n})^T k_{e_{n-1}} \mid \mathbf{0} \right]. \tag{12}
\end{aligned}$$

Unlike the case with original knowledge, as editing progresses, earlier edits experience more interference, while later edits are less affected. This is consistent with Jang et al. (2021). For convenience, we study $\Delta_e[:, j]$, which represents the j -th column vector of Δ_e . This can be seen as the impact introduced to the j -th edited knowledge after n edits,

$$\Delta_e[:, j] = \begin{cases} \sum_{i=j+1}^n \Lambda_i (C^{-1} k_{e_i})^T k_{e_j}, & j < n \\ \mathbf{0}, & j = n \end{cases} \tag{13}$$

In a similar manner, we can express it as

$$\Delta_e[:, j] = \begin{cases} \sum_{i=j+1}^n \frac{(C^{-1} k_{e_i})^T k_{e_j}}{(C^{-1} k_{e_i})^T k_{e_i}} (v_{e_i} - W_{i-1} k_{e_i}), & j < n \\ \mathbf{0}, & j = n \end{cases} \tag{14}$$

Next, let $\delta(v_{e_i}) = v_{e_i} - W_{i-1} k_{e_i} \in \mathbb{R}^d$ represent the difference vector between the optimized value and the current value when editing the i -th piece of knowledge. Define the coefficient $p(k_{e_i}, k_{e_j}) = \frac{(C^{-1} k_{e_i})^T k_{e_j}}{(C^{-1} k_{e_i})^T k_{e_i}} \in \mathbb{R}$. Then, we obtain

$$\Delta_e[:, j] = \begin{cases} \sum_{i=j+1}^n p(k_{e_i}, k_{e_j}) \delta(v_{e_i}), & j < n \\ \mathbf{0}, & j = n \end{cases} \tag{15}$$

Ideally, we want $\Delta_e[:, j] = \mathbf{0}$ to achieve lossless knowledge editing, ensuring that previously edited knowledge is not affected. However, since $\delta(v_{e_i}) = v_{e_i} - W_{i-1} k_{e_i} \neq \mathbf{0}$, we can only hope that $p(k_{e_i}, k_{e_j}) = 0$. This would eliminate the interference on the j -th edited knowledge.

How to Understand $p(\cdot, \cdot)$

In summary, we find that for both the original knowledge and the edited knowledge, there is a similar coefficient $p(\cdot, \cdot)$ in interference term, which directly determines whether lossless knowledge editing can be achieved.

To study this coefficient more generally, we do not limit to specific keys (which correspond to specific knowledge activations) but consider any keys (or any knowledge activations) k_i and k_j . We further investigate

$$p(k_i, k_j) = \frac{(C^{-1} k_i)^T k_j}{(C^{-1} k_i)^T k_i} \in \mathbb{R}. \tag{16}$$

Since $C = KK^T$ is a symmetric matrix, C^{-1} is also symmetric. Therefore, we can rewrite it as

$$p(k_i, k_j) = \frac{k_i^T C^{-1} k_j}{k_i^T C^{-1} k_i} = \frac{(C^{-\frac{1}{2}} k_i)^T (C^{-\frac{1}{2}} k_j)}{(C^{-\frac{1}{2}} k_i)^T (C^{-\frac{1}{2}} k_i)}. \quad (17)$$

Here, $C = KK^T$ is the covariance matrix, and both k_i and k_j are column vectors of K . Therefore, $C^{-\frac{1}{2}} k_i$ and $C^{-\frac{1}{2}} k_j$ are the representations of k_i and k_j in the whitening space (Koivunen and Kostinski 1999; Kawahara et al. 2007) after the whitening transformation $C^{-\frac{1}{2}}$ (proof in Appendix B). The coefficient $p(\cdot, \cdot)$ can be understood as the dot product of k_i and k_j in the whitening space, normalized. When $p(\cdot, \cdot) = 0$, it indicates that the knowledge activations k_i and k_j are orthogonal in the whitening space.

Definition 1 (Matrix Whitening) For a matrix X , its covariance matrix $\text{Cov}(X) = XX^T$ is not necessarily the identity matrix. Matrix whitening involves finding a transformation matrix P such that the covariance matrix of $Y = PX$, denoted as $\text{Cov}(Y)$, becomes the identity matrix.

Returning to the problem, if we aim to achieve perfectly lossless knowledge editing, this is equivalent to requiring both Δ_o and Δ_e to be zero matrices. This in turn means we expect $p(\cdot, \cdot)$ to be zero, which is equivalent to expecting that knowledge activations are orthogonal in the whitening space. Consequently, this implies that we expect the language model to store different pieces of knowledge in orthogonal directions in the whitening space, and thus that knowledge superposition does not exist in the whitening space. **Conversely, if such superposition exists, then the representations are not orthogonal and we cannot achieve perfectly lossless knowledge editing.**

Knowledge in Superposition

In this section, we focus on the phenomenon of knowledge superposition in real-world language models, combining previous theoretical derivations to demonstrate that knowledge superposition is the fundamental reason for the failure in lifelong editing.

Specifically, we calculate the degree of superposition between pieces of knowledge by computing the matrix P for m pieces of knowledge, where

$$P[i, j] = p(k_i, k_j), 1 \leq i, j \leq m. \quad (18)$$

As $p(\cdot, \cdot)$ approaches 0, knowledge representations become more orthogonal in the whitening space, and the degree of superposition becomes weaker. Conversely, as $p(\cdot, \cdot)$ approaches 1, knowledge representations become more similar in the whitening space, and the degree of superposition becomes stronger.

In practice, we choose $m = 128$ to compute the superposition between 128×128 pairs of knowledge and resulting in a 128×128 matrix P . Empirical evidence shows that $m = 128$ is sufficient because, at this point, the kurtosis of the superposition distribution has converged (details in Appendix C). This indicates that the data size is adequate for describing the distribution of superposition.

The specific experimental setup is described as follows:

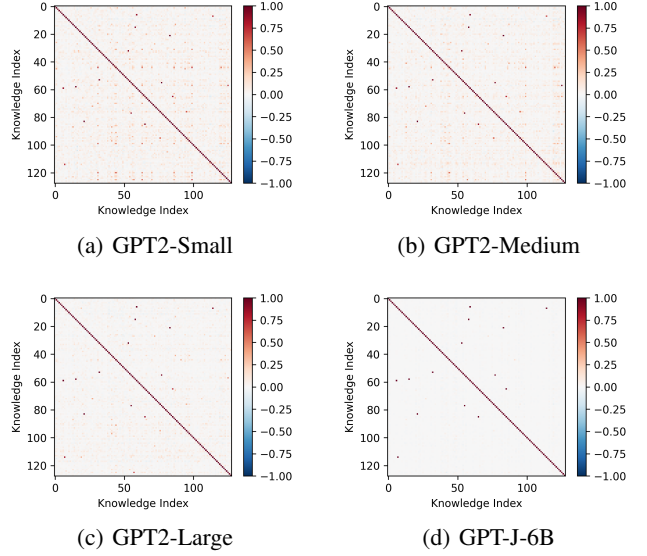


Figure 3: Superposition at layer 0 across different language models visualized using P matrices, ordered by model size. Each point in these 128×128 P matrices corresponds to the $p(\cdot, \cdot)$ value between two pieces of knowledge.

Models In this study, we employ a variety of models including GPT2 family (Radford et al. 2019) —GPT2-Small, GPT2-Medium, and GPT2-Large, Pythia family (Biderman et al. 2023) —Pythia-1B, Pythia-2.8B, and Pythia-6.9B, Llama2 family (Touvron et al. 2023) —Llama2-7B and Llama2-13B. Additionally, we use the classic GPT-J (Wang and Komatsuzaki 2022), the latest Llama3-8B (Meta 2024a) and Llama3.1-8B (Meta 2024b).

Datasets For the extraction of knowledge representations, we utilized the CounterFact dataset (Meng et al. 2022a). It is important to note that while CounterFact is commonly used for counterfactual knowledge editing, our experimental setup does not involve this aspect. Instead, we focus on subject-related knowledge rather than counterfactual objects (details in Appendix G).

Universal in Language Models

Ideally, the matrix P we obtain should be such that all positions except the diagonal are zero, indicating that no superposition exists and allowing for lossless knowledge editing. However, in all layers of all the language models we studied, the P matrices we obtained consistently show noisy non-zero entries at positions other than the diagonal, indicating the presence of superposition at these points. As shown in Figure 3, we present the heatmaps of P matrices for GPT2-Small, GPT2-Medium, GPT2-Large, and the GPT-J at layer 0, ordered by model size. Additional heatmaps of P matrices for all layers of all models are provided in Appendix D.

It is evident that as the model size increases, the P matrices become progressively "cleaner." This indicates that as language models gain more storage capacity to store knowledge, they tend to store this knowledge with weaker super-

position, thereby reducing the interference between pieces of knowledge caused by superposition.

Additionally, we observe that even as model size increases, fixed interference points remain at positions off the diagonal, even in models with different architectures and trained on different corpora. We hypothesize that the knowledge pairs corresponding to these points may actually be closely related, despite having different expressions, leading to this phenomenon. This point has been validated through detailed case studies. We find that the knowledge pairs corresponding to these points are indeed closely related. For instance, in the first layer (layer 0) of these four language models, the $p(\cdot, \cdot)$ values for "Vladimir Mayakovsky" and "Vladimir Bukovsky" (both of whom are native Russian speakers) are above 0.95, and even 1.00 in GPT-J, indicating that the operations performed by the MLP in the first layer for these two pieces of knowledge are similar or even identical. This also implies that if we attempt to edit the knowledge of subject "Vladimir Mayakovsky" in this layer, it will have a consistent impact on "Vladimir Bukovsky," suggesting that they are bound together. For example, if we attempt to edit "Vladimir Mayakovsky" to have French as his native language, the edited model will also output French as "Vladimir Bukovsky's" native language. Similar examples include "Windows 8.1" and "Mac OS X 10.1," which generally have very high $p(\cdot, \cdot)$ values (1.00 in the first layer of GPT-J), even though they are produced by entirely different manufacturers, which is fascinating! The case study is detailed in Appendix F.

Furthermore, choosing different editing layers may lead to certain changes, as described by Hu et al. (2024). However, other layers also contain other knowledge in superposition, as shown in Figure 5.

Heavy-Tailed Distribution in Language Models

To gain a more intuitive understanding of the distribution characteristics of knowledge superposition, we remove the diagonal elements from the P matrices across all layers of all the language models we studied and then plot the kernel density estimation (KDE) of the remaining elements. In Figure 4, we present the kernel density estimation of the superposition distribution at layer 0 for GPT2-Small, GPT2-Medium, GPT2-Large, and the GPT-J, ordered by model size. Additional kernel density estimations for all layers of all models are provided in Appendix E.

It can be observed that the superposition distribution exhibits characteristics of a heavy-tailed distribution with high kurtosis and zero mean. As the model size increases, the kurtosis of distribution becomes larger and the distribution becomes more concentrated around 0. This indicates that smaller models, constrained by capacity, exhibit more superposition, attempting to store knowledge representations in a relatively orthogonal manner. In contrast, larger models have greater capacity to store knowledge, allowing them to store it in a more orthogonal manner compared to smaller models, resulting in a kernel density estimation more concentrated around 0.

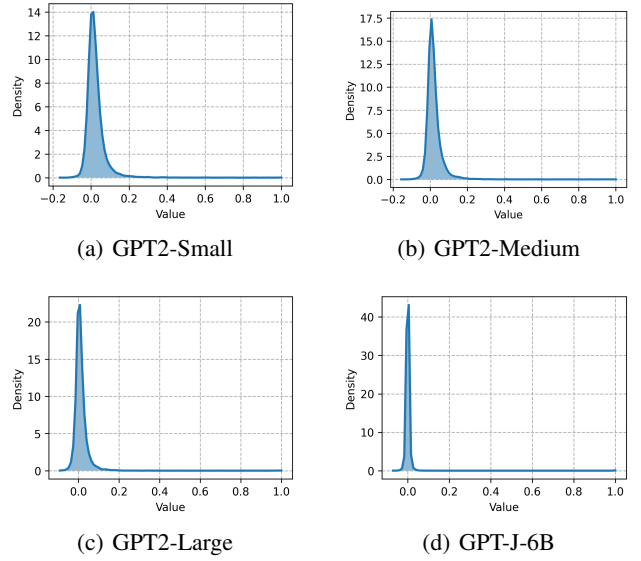


Figure 4: KDE of $p(\cdot, \cdot)$ values in P matrices at layer 0 across different language models, ordered by model size. In P matrices, $p(\cdot, \cdot)$ values are concentrated around 0, with high kurtosis, which increases as model size grows.

Scaling Law for Superposition

We have already seen that the degree of knowledge superposition varies with model size. In this section, we formally study the scaling law of superposition by experimenting with different-sized models from the same language model family. These models have consistent architecture and training corpora, differing only in size. As shown in Figure 5, we examine the kurtosis across all layers of a total of 8 language models from the GPT2, Pythia, and Llama-2 families. Higher kurtosis reflects weaker superposition.

We can observe a clear scaling law: as model size increases, kurtosis tends to rise, indicating that the degree of superposition between pieces of knowledge decreases with larger models. This also explains why larger language models exhibit a higher degree of intelligence. A plausible explanation is that as model size grows, the knowledge from the corpus can be stored with weaker superposition, allowing for effective handling of more complex or knowledge-dense scenarios. Since the sparser the features, the stronger the superposition (Elhage et al. 2022b), which also means they can only effectively handle scenarios where knowledge features are sparser.

Additionally, we find that the degree of superposition varies across different layers in the same language model. Earlier layers exhibit higher kurtosis and thus lower degree of superposition, while later layers exhibit lower kurtosis and higher degree of superposition. This is reasonable, as early layers in a language model focus on shallow syntactic features, which are dense and result in weaker superposition. In contrast, later layers focus on deep semantic features, which are sparse and result in stronger superposition. This aligns with Elhage et al. (2022b).

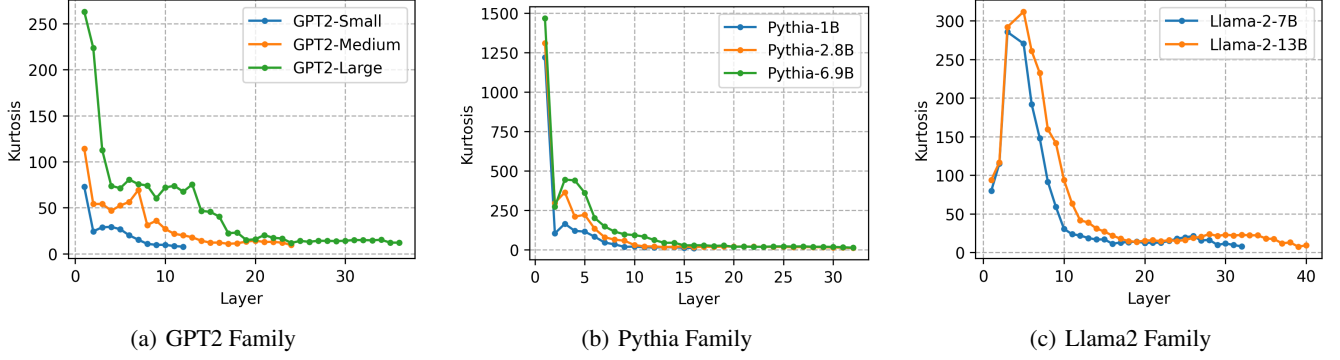


Figure 5: The scaling law of knowledge superposition. Higher kurtosis means less superposition.

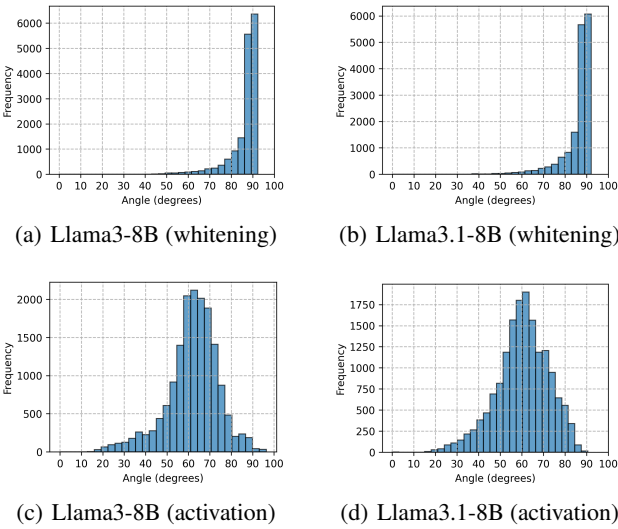


Figure 6: Angular distribution of knowledge representations in whitening space and activation space in last layer.

Furthermore, we find that different architectures have varying impacts on knowledge superposition. As shown in Figure 5, by observing the vertical axis, it is evident that the Pythia and GPT2 architectures seem to encourage more orthogonal representations of knowledge, achieving higher kurtosis with a smaller number of parameters, indicating weaker superposition. In contrast, the Llama2 architecture seems to encourage greater knowledge superposition, with even larger parameter sizes corresponding to relatively lower kurtosis, indicating stronger superposition. This is consistent with Elhage et al. (2022a), which suggests that some architectures may encourage sparsity.

Superposition in Whitening Space

In this section, we provide further direct evidence of superposition by calculating the angular distribution of knowledge representations, demonstrating its presence in the whitening space. Specifically, we calculate the angles be-

tween 128x128 pairs of knowledge representations in the whitening space on Llama3-8B and Llama-3.1-8B. For comparison, we also calculate the angles in the activation space (i.e., directly from MLP’s hidden activations).

In the whitening space (Figure 6ab), the angles are primarily concentrated around 90 degrees, indicating that language models attempt to represent knowledge orthogonally, though limited model capacity results in a long tail. However, in the activation space (Figure 6cd), the relationships between these knowledge representations are indiscernible, showing no orthogonal tendencies.

Importantly, superposition in the whitening space is derived through rigorous mathematical reasoning, whereas superposition in the activation space is based on a heuristic perspective (Elhage et al. 2022b). However, this heuristic perspective has been used as a fundamental assumption in many works that address superposition (Cunningham et al. 2023; Bricken et al. 2023; Templeton et al. 2024).

Conclusion

In this research, we focus on the failure of lifelong knowledge editing and explore its fundamental reason. Our rigorous mathematical derivation indicates that continuous knowledge editing can interfere with both original and previously edited knowledge in language models. Further analysis shows that the degree of interference is determined by the extent of knowledge superposition. Subsequently, we explore the presence and properties of superposition in real-world language models. Our experiments reveal that knowledge superposition is universal across language models, characterized by a high-kurtosis, zero-mean, heavy-tailed distribution, with a clear scaling law. In conclusion, our findings indicate that knowledge superposition is the fundamental reason for the failure in lifelong editing. Nevertheless, the scaling law of superposition still suggests future directions for knowledge editing, including: on the model side, attempting knowledge editing in larger language models with sparser architectures; and on the algorithm side, attempting to decompose knowledge within language models and performing edits based on this decomposition.

References

- Anderson, J. A. 1972. A simple neural network generating an interactive memory. *Mathematical biosciences*, 14(3-4): 197–220.
- Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O’Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.
- Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; Lasenby, R.; Wu, Y.; Kravec, S.; Schiefer, N.; Maxwell, T.; Joseph, N.; Hatfield-Dodds, Z.; Tamkin, A.; Nguyen, K.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.
- De Cao, N.; et al. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6491–6506.
- Dong, Q.; Dai, D.; Song, Y.; Xu, J.; Sui, Z.; and Li, L. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.
- Elhage, N.; Hume, T.; Olsson, C.; Nanda, N.; Henighan, T.; Johnston, S.; ElShowk, S.; Joseph, N.; DasSarma, N.; Mann, B.; Hernandez, D.; Askell, A.; Ndousse, K.; Jones, A.; Drain, D.; Chen, A.; Bai, Y.; Ganguli, D.; Lovitt, L.; Hatfield-Dodds, Z.; Kernion, J.; Conerly, T.; Kravec, S.; Fort, S.; Kadavath, S.; Jacobson, J.; Tran-Johnson, E.; Kaplan, J.; Clark, J.; Brown, T.; McCandlish, S.; Amodei, D.; and Olah, C. 2022a. Softmax Linear Units. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/solu/index.html>.
- Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022b. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.
- Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.
- Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.
- Gurnee, W.; Nanda, N.; Pauly, M.; Harvey, K.; Troitskii, D.; and Bertsimas, D. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.
- Hartvigsen, T.; Sankaranarayanan, S.; Palangi, H.; Kim, Y.; and Ghassemi, M. 2024. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36.
- Hu, C.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2024. Wilke: Wise-layer knowledge editor for lifelong knowledge editing. *arXiv preprint arXiv:2402.10987*.
- Huang, Z.; Shen, Y.; Zhang, X.; Zhou, J.; Rong, W.; and Xiong, Z. 2023. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*.
- Jang, J.; Ye, S.; Yang, S.; Shin, J.; Han, J.; Kim, G.; Choi, S. J.; and Seo, M. 2021. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*.
- Kawahara, T.; Nishiyama, M.; Kozakaya, T.; and Yamaguchi, O. 2007. Face recognition based on whitening transformation of distribution of subspaces. In *Proc. ACCV07 Workshop Subspace*, 97–103.
- Kohonen, T. 1972. Correlation matrix memories. *IEEE transactions on computers*, 100(4): 353–359.
- Koivunen, A.; and Kostinski, A. 1999. The feasibility of data whitening to improve performance of weather radar. *Journal of Applied Meteorology and Climatology*, 38(6): 741–749.
- Lhoest, Q.; Del Moral, A. V.; Jernite, Y.; Thakur, A.; Von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.
- Li, X.; Li, S.; Song, S.; Yang, J.; Ma, J.; and Yu, J. 2023. PMET: Precise Model Editing in a Transformer. *arXiv preprint arXiv:2308.08742*.
- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35: 17359–17372.
- Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Meta, A. 2024a. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.
- Meta, A. 2024b. The Llama 3 Herd of Models. *arXiv:2407.21783*.
- Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.
- Mitchell, E.; Lin, C.; Bosselut, A.; Manning, C. D.; and Finn, C. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, 15817–15831. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajamanoharan, S.; Conmy, A.; Smith, L.; Lieberum, T.; Varma, V.; Kramár, J.; Shah, R.; and Nanda, N. 2024. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*.

Tan, C.; Zhang, G.; and Fu, J. 2023. Massive editing for large language models via meta learning. *arXiv preprint arXiv:2311.04661*.

Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Summers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, B.; and Komatsuzaki, A. 2022. GPT-J-6B: a 6 billion parameter autoregressive language model (2021). *URL* <https://github.com/kingoflolz/mesh-transformer-jax>.

Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; and Zhang, N. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Zhong, Z.; Wu, Z.; Manning, C. D.; Potts, C.; and Chen, D. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

Reproducibility Checklist

Unless specified otherwise, please answer “yes” to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled “Reproducibility Checklist” at the end of the technical appendix.

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA)
yes
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no)
yes
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes/no)
yes

Does this paper make theoretical contributions? (yes/no) **yes**

If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
yes
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
yes
- Proofs of all novel claims are included. (yes/partial/no)
yes
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
yes
- Appropriate citations to theoretical tools used are given. (yes/partial/no)
yes
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)
yes
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA)
yes

Does this paper rely on one or more datasets? (yes/no) **yes**

If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA)
yes
- All novel datasets introduced in this paper are included in a data appendix. (yes/partial/no/NA)
yes
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no/NA)
yes
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/NA)
yes
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes/partial/no/NA)
yes
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying. (yes/partial/no/NA)
yes

Does this paper include computational experiments? (yes/no) **yes**

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes/partial/no).
yes
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes/partial/no)
yes
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no)
yes
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from

(yes/partial/no)

yes

- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA)

yes

- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no)

yes

- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no)

yes

- This paper states the number of algorithm runs used to compute each reported result. (yes/no)

yes

- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes/no)

yes

- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/no)

yes

- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes/partial/no/NA)

yes

- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA)

yes