# *CollabEdit*: Towards Non-destructive Collaborative Knowledge Editing

**Jiamu Zheng [1,§] Jinghuai Zhang [3] Tianyu Du [1,†] Xuhong Zhang [1] Jianwei Yin [1] Tao Lin [2]**

Zhejiang University [1] Westlake University [2] University of California, Los Angeles [3]

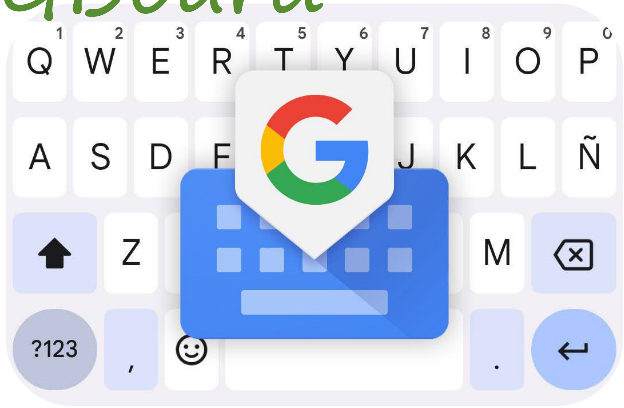
[§]Work was done during Jiamu's visit to Westlake University.

[†]Corresponding author.

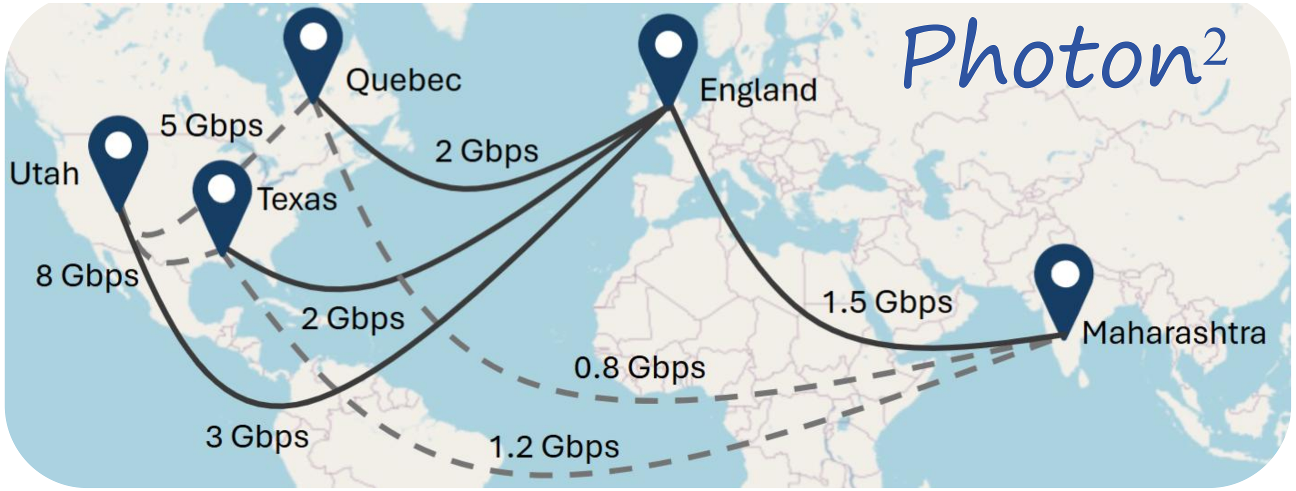
Code Paper

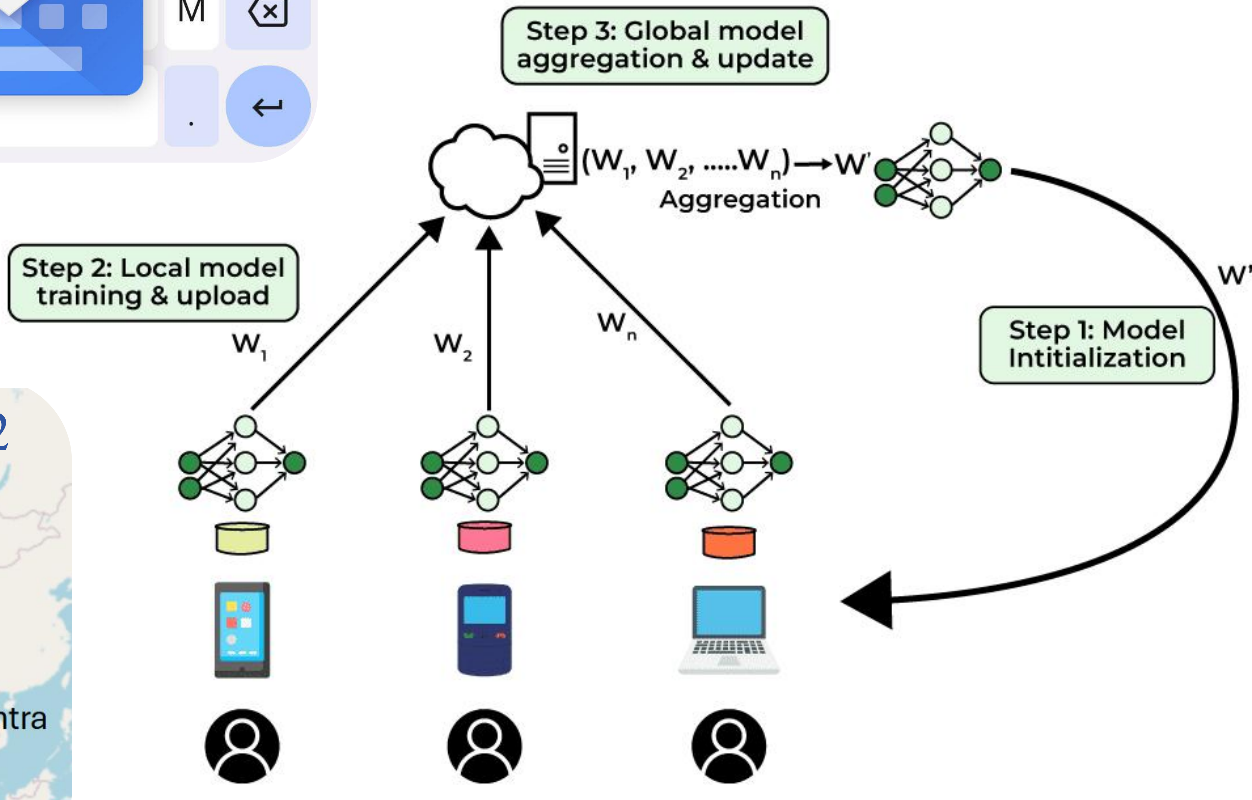
# Background

GBoard[1]



- What is **Collaborative** Learning ?




[1]GBoard: https://blog.google/products/search/gboard-now-on-android/
[2]Photon: Federated LLM Pre-Training

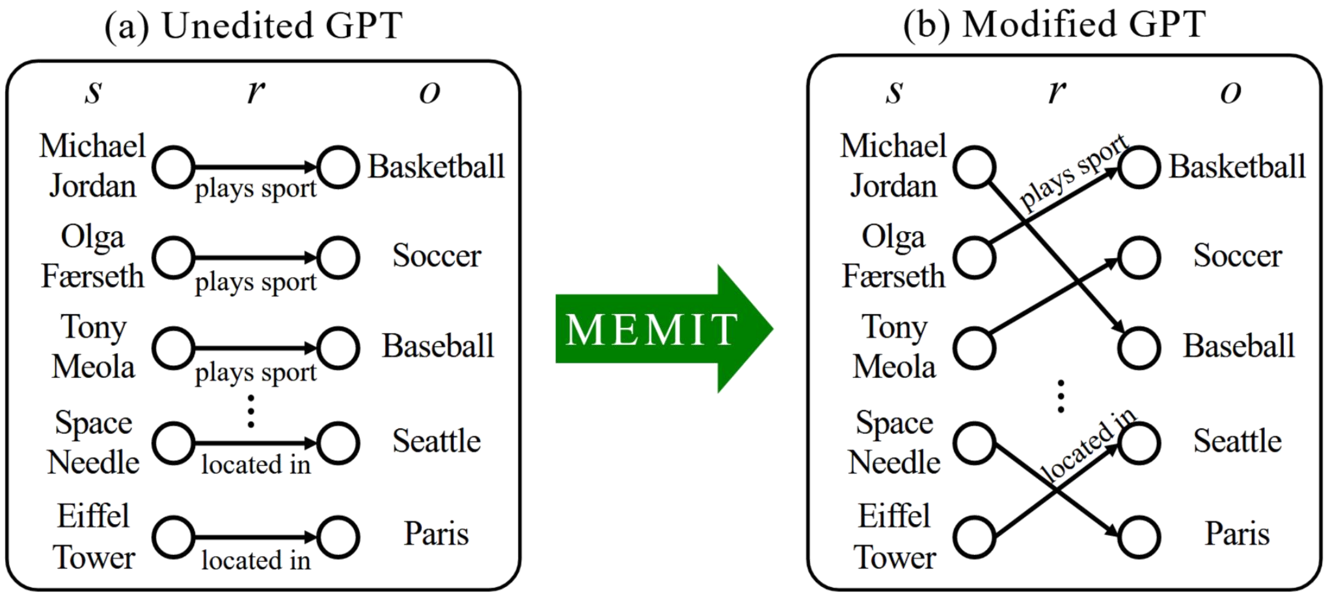
ICLR

Code

Paper

# Background

- What is Collaborative Learning ?
- What is Knowledge Editing (KE) ?




Edits Request $\varepsilon_1$

**New Fact**: The president of US is ~~Obama.~~ **Biden.**

**Q:** The president of US is ? **A:** Biden.

*Update outdated knowledge / Machine Unlearning / Specific modification ..*

***without Re-training or Fine-tuning !***

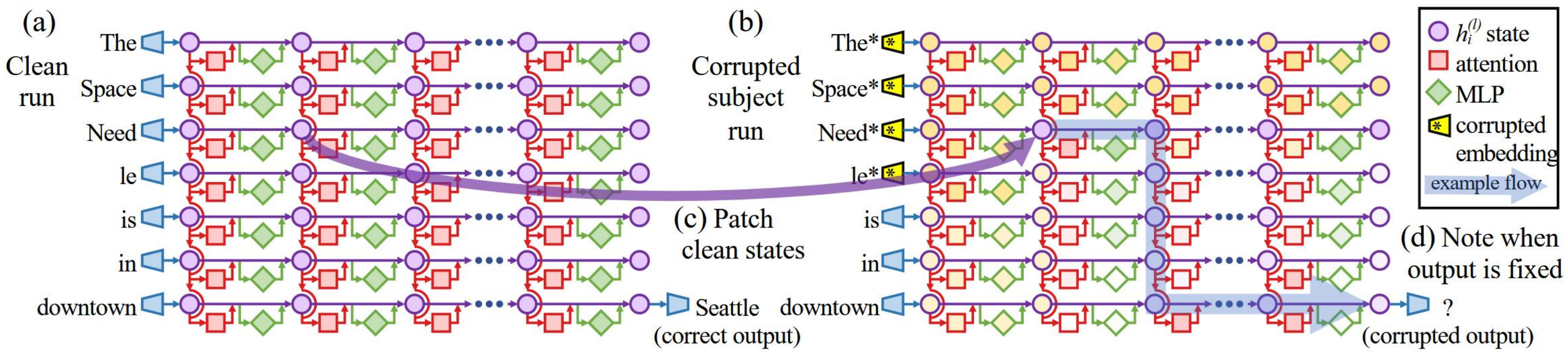
ICLR

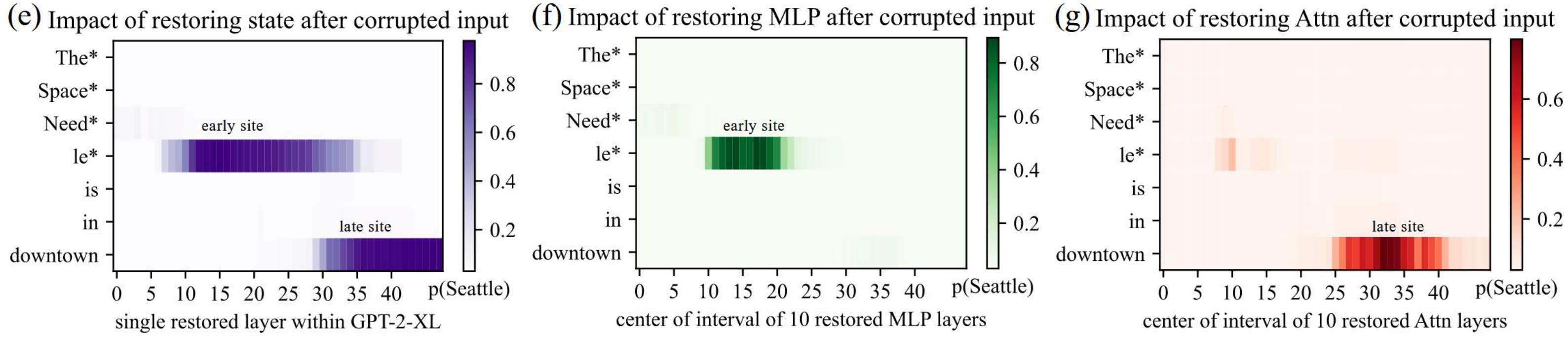
Code

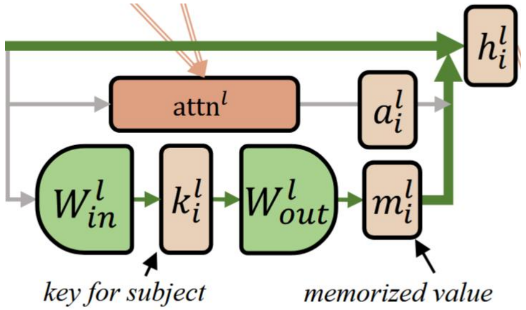
Paper

# Background

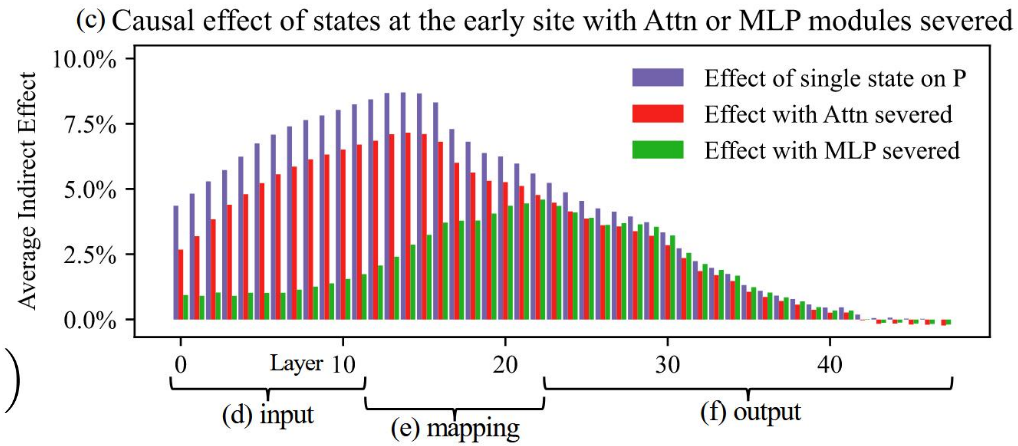
- What is Collaborative Learning ?
- What is Knowledge Editing (KE) ?
- Where to conduct KE ?
    - ROME[1]




*The MLP layer stores the mapping relationships of knowledge.*




$$h_{[t]}^{l}(x) = h_{[t]}^{l-1}(x) + a_{[t]}^{l}(x) + m_{[t]}^{l}(x)$$

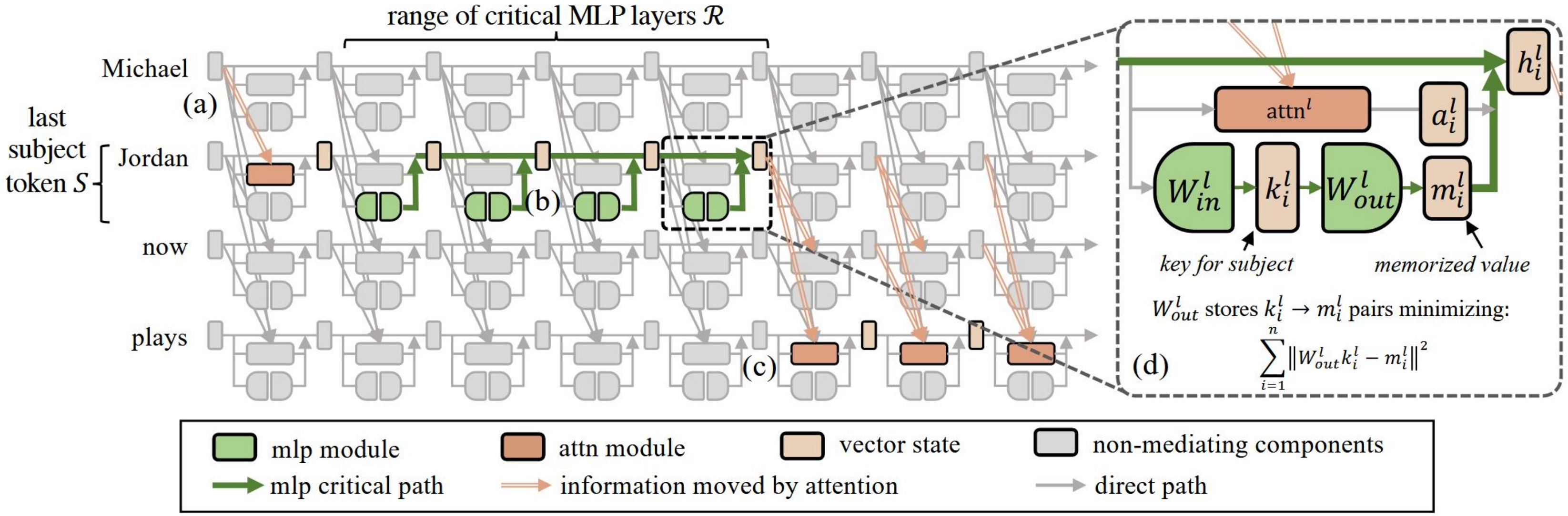
$$\text{where } a^{l} = \text{attn}^{l}\left(h_{[1]}^{l-1}, h_{[2]}^{l-1}, \ldots, h_{[t]}^{l-1}\right)$$

$$\text{GPT2:} \quad m_{[t]}^{l} = W_{out}^{l}\,\sigma\left(W_{in}^{l}\gamma\left(h_{[t]}^{l-1}\right)\right),$$

$$\text{GPT-J:} \quad m_{i}^{(l)} = W_{proj}^{(l)}\,\sigma\left(W_{fc}^{(l)}\gamma\left(a_{i}^{(l)} + h_{i}^{(l-1)}\right)\right)$$

[1] Locating and Editing Factual Associations in GPT (NeurIPS 2022)

# Background



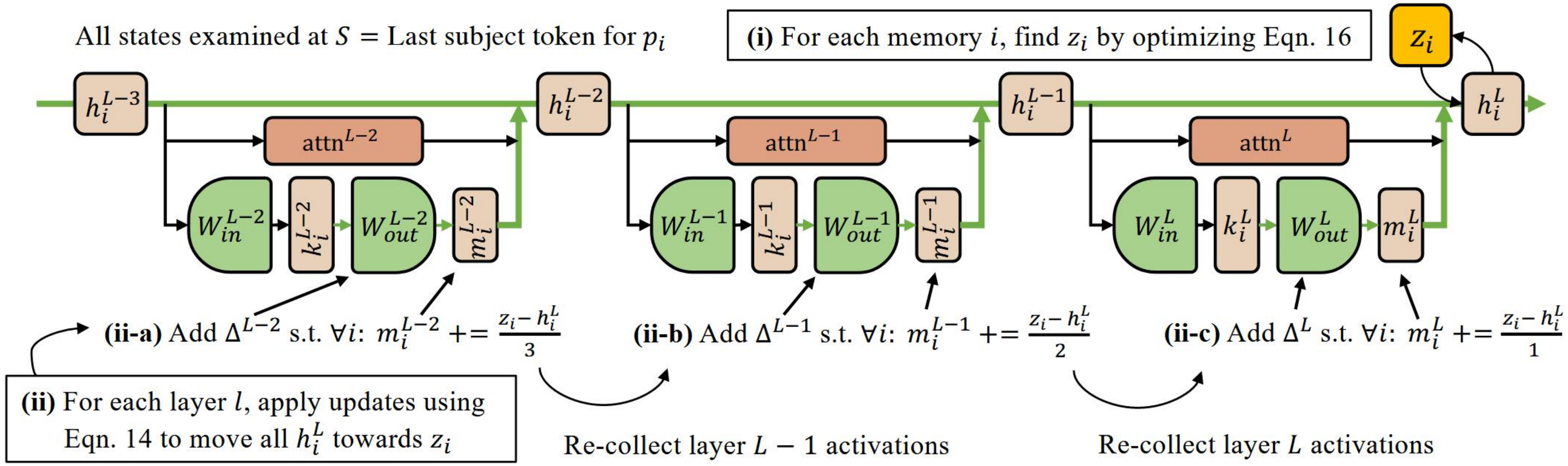

- What is Collaborative Learning ?
- What is Knowledge Editing (KE) ?
- Where to conduct KE ?
- How to conduct KE ?
  - MEMIT[1]
  - AlphaEdit[2]
  - …

$$W_1 \triangleq \underset{\hat{W}}{\operatorname{argmin}} \left( \sum_{i=1}^{n} \left\| \hat{W} k_i - m_i \right\|^2 + \sum_{i=n+1}^{n+u} \left\| \hat{W} k_i - m_i \right\|^2 \right)$$

$$\Delta = R K_1^T \left( C_0 + K_1 K_1^T \right)^{-1}$$

$$R \triangleq M_1 - W_0 K_1$$

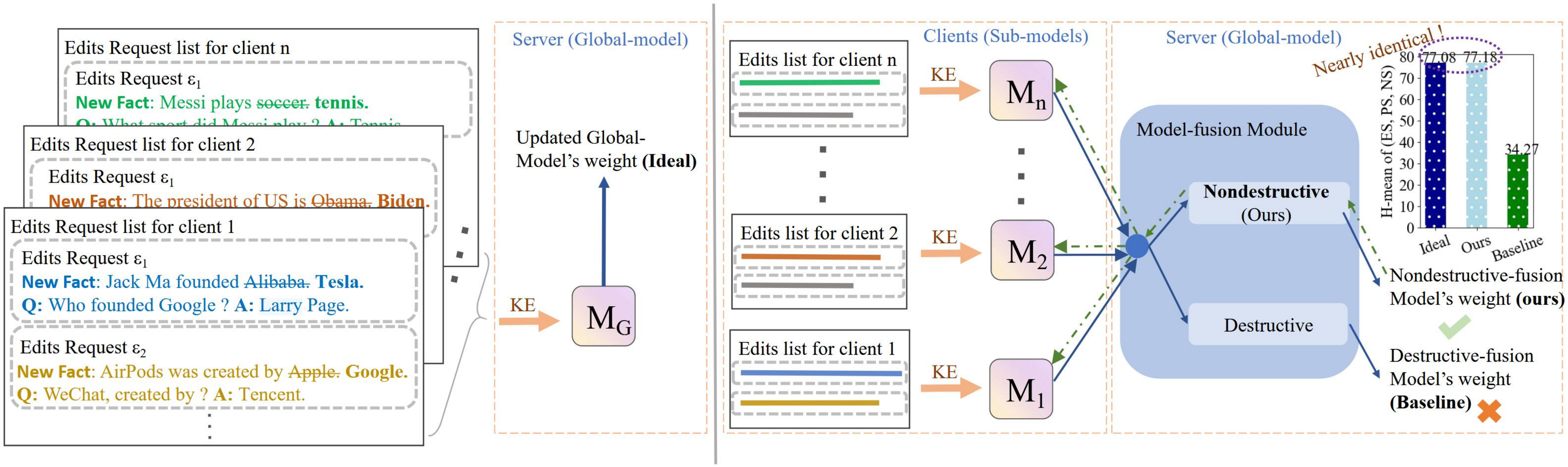
$$C_0 = \lambda \cdot \mathbb{E}_k \left[ k k^T \right]$$




[1]Mass-Editing Memory in a Transformer (ICLR 2023)

[2]AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models (ICLR 2025 Oral)

# Introduction




(a) Global Editing

(b) Collaborative Editing

- Knolwedge Editing (KE)
  - Global Editing : *upper bound performance*
  - Collaborative Editing
    - Destructive Fusion
    - **Nondestructive Fusion (our *CollabEdit* )**

Edits Request $\varepsilon_1$

**New Fact**: The president of US is ~~Obama.~~ **Biden.**

**Q:** The president of US is ? **A:** Biden.

# Methodology

- Destructive Collaborative Editing
  - Dramatic performance drop

$$\mathbf{\Delta} = \mathbf{R}\mathbf{K}^\top(\mathbf{C} + \mathbf{K}\mathbf{K}^\top)^{-1}$$

Simple Average[1]: $\theta = \frac{1}{n}\sum_{i=1}^{n}\theta_i$

Task Arithmetic[2]: $\theta = \theta_0 + \lambda\sum_i(\theta_i - \theta_0)$

Ties Merging[3]: $\theta = \theta_0 + \lambda\nu$

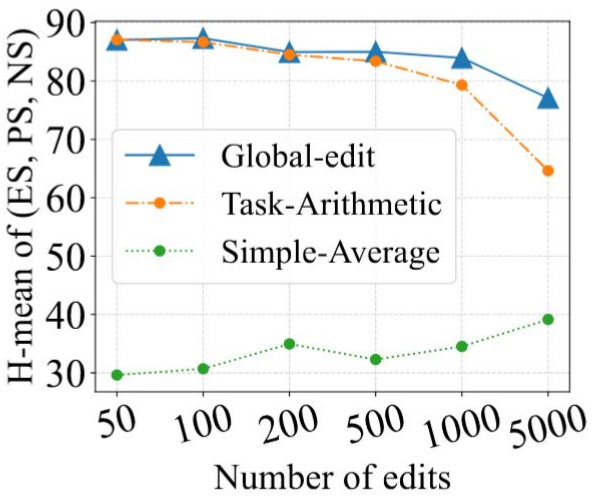



Figure 1: Limits of existing KE methods under the collaborative KE scenarios on the Multi-CounterFact dataset (Meng et al., 2022).

---

[1] Adaptersoup: Weight averaging to improve generalization of pretrained language models (ACL 2023)
[2] Task arithmetic in the tangent space: Improved editing of pre-trained models (NeurIPS 2023)
[3] Ties-merging: Resolving interference when merging models (NeurIPS 2023)

ICLR

Code

Paper

# Methodology

- Destructive Collaborative Editing
  - Dramatic performance drop
- Nondestructive Collaborative Editing

$$\mathbf{\Delta} = \mathbf{R}\mathbf{K}^\top(\mathbf{C} + \mathbf{K}\mathbf{K}^\top)^{-1}$$

Note that $\mathbf{\Delta}_i$ and $\mathbf{\Delta}_G$ can be computed via $\boxed{(2)}$ as:

$$\begin{aligned}\mathbf{\Delta}_G &= \mathbf{R}_G\mathbf{K}_G^\top(\mathbf{C} + \mathbf{K}_G\mathbf{K}_G^\top)^{-1}, \\ \mathbf{\Delta}_i &= \mathbf{R}_i\mathbf{K}_i^\top(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)^{-1}.\end{aligned} \tag{13}$$

Following the definitions of $\mathbf{K}$ and $\mathbf{R}$ in Section 3.1, we have:

$$\begin{aligned}\mathbf{K}_i &= [\mathbf{k}_{i\times(M-1)+1}, \mathbf{k}_{i\times(M-1)+2}, \cdots, \mathbf{k}_{i\times M}], \\ \mathbf{R}_i &= [\mathbf{r}_{i\times(M-1)+1}, \mathbf{r}_{i\times(M-1)+2}, \cdots, \mathbf{r}_{i\times M}], \\ \mathbf{K}_G &= [\mathbf{k}_1, \mathbf{k}_2, \cdots, \mathbf{k}_{N\times M}] = [\mathbf{K}_1, \mathbf{K}_2, \cdots, \mathbf{K}_N], \\ \mathbf{R}_G &= [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_{N\times M}] = [\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N].\end{aligned} \tag{14}$$

Then we have:

$$\mathbf{R}_G\mathbf{K}_G^\top = \mathbf{R}_1\mathbf{K}_1^\top + \mathbf{R}_2\mathbf{K}_2^\top + \cdots + \mathbf{R}_N\mathbf{K}_N^\top. \tag{15}$$

According to Equations (13) and (15), we can obtain:

$$\begin{aligned}\mathbf{\Delta}_G(\mathbf{C} + \textstyle\sum_{j=1}^N \mathbf{K}_j\mathbf{K}_j^\top) &= \mathbf{\Delta}_G(\mathbf{C} + \mathbf{K}_1\mathbf{K}_1^\top \cdots + \mathbf{K}_N\mathbf{K}_N^\top) \\ &= \mathbf{\Delta}_G(\mathbf{C} + \mathbf{K}_G\mathbf{K}_G^\top) \\ &= \mathbf{R}_G\mathbf{K}_G^\top \\ &= \mathbf{R}_1\mathbf{K}_1^\top + \mathbf{R}_2\mathbf{K}_2^\top + \cdots + \mathbf{R}_N\mathbf{K}_N^\top \\ &= \mathbf{\Delta}_1(\mathbf{C} + \mathbf{K}_1\mathbf{K}_1^\top) + \cdots + \mathbf{\Delta}_N(\mathbf{C} + \mathbf{K}_N\mathbf{K}_N^\top) \\ &= \textstyle\sum_{i=1}^N \mathbf{\Delta}_i(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top).\end{aligned} \tag{16}$$

According to the Equation (16), we can finally reach the following conclusion:

$$\mathbf{\Delta}_G = \textstyle\sum_{i=1}^N \mathbf{\Delta}_i(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)(\mathbf{C} + \sum_{j=1}^N \mathbf{K}_j\mathbf{K}_j^\top)^{-1}. \tag{17}$$

# Methodology

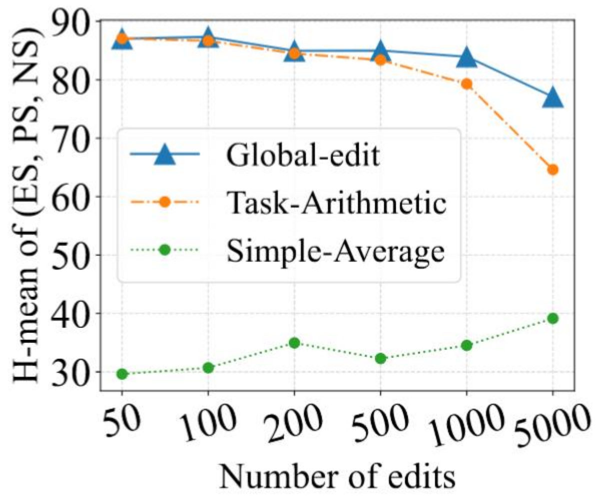


Figure 1: Limits of existing KE methods under the collaborative KE scenarios on the Multi-CounterFact dataset (Meng et al., 2022).

$$\boldsymbol{\Delta} = \mathbf{R}\mathbf{K}^\top(\mathbf{C} + \mathbf{K}\mathbf{K}^\top)^{-1}$$

Simple Average[1]: $\theta = \frac{1}{n}\sum_{i=1}^{n}\theta_i$

Task Arithmetic[2]: $\theta = \theta_0 + \lambda\sum_i(\theta_i - \theta_0)$

- Destructive Collaborative Editing
  - Dramatic performance drop
- Nondestructive Collaborative Editing

$$\boldsymbol{\Delta}'_G = \lambda \times (\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2 + \cdots + \boldsymbol{\Delta}_N)$$

$$\boldsymbol{\Delta}_G = \sum_{i=1}^{N}\boldsymbol{\Delta}_i(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)(\mathbf{C} + \sum_{j=1}^{N}\mathbf{K}_j\mathbf{K}_j^\top)^{-1}$$

$$\boldsymbol{\Delta}_G - \boldsymbol{\Delta}'_G = \sum_{i=1}^{N}\boldsymbol{\Delta}_i\left[(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)(\mathbf{C} + \sum_{j=1}^{N}\mathbf{K}_i\mathbf{K}_i^\top)^{-1} - \lambda\mathbf{I}\right]$$

The average l2-norm of $K_iK_i^T$ is approximately **0.0001%** of that of $C$ *!!!*

$$(\mathbf{C} + \sum_{j=1}^{N}\mathbf{K}_i\mathbf{K}_i^\top)^{-1} \approx \mathbf{C} \qquad \boldsymbol{\Delta}_G \approx \boldsymbol{\Delta}'_G$$

# Methodology

- Destructive Collaborative Editing
  - Dramatic performance drop
- Nondestructive Collaborative Editing
- Three new challenges and solutions
  - Intervension between different clients
    - Knowledge overlap
    - Knowledge conflict

$$\mathbf{R}_{\text{new}} := \mathbf{R}_{\text{old}} - \mathbf{\Delta K} = \mathbf{R}_{\text{old}} - \mathbf{R}_{\text{old}}\mathbf{K}^\top(\mathbf{C} + \mathbf{KK}^\top)^{-1}\mathbf{K}$$

$$\begin{cases} e_1 = (s_1, r_1, o_1 \rightarrow o_2, t_1, m_1) \\ e_2 = (s_1, r_1, o_1 \rightarrow o_3, t_2, m_2) \end{cases}$$

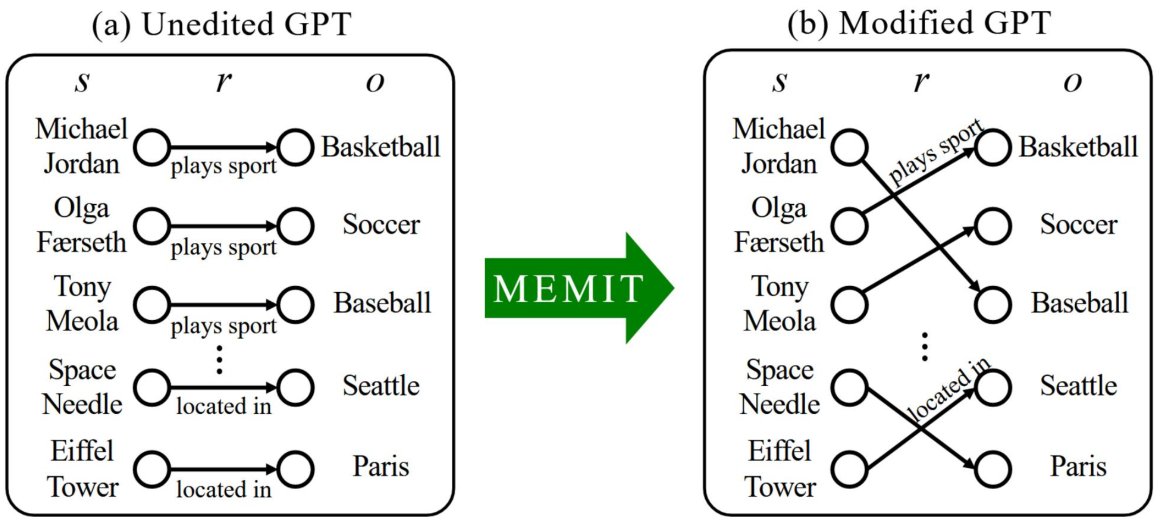
| **Situation** |
| --- |
| $m_1 = m_2$ |
| $m_1 \neq m_2$ and $t_1 = t_2$ |
| $m_1 \neq m_2$ and $t_1 \neq t_2$ |

$m_1 = m_2$ → Data Augmentation (e.g. Multi-Label Editing[1])

$m_1 \neq m_2$ and $t_1 = t_2$ → The overwriting nature of KE

$$\mathbf{\Delta} = \mathbf{RK}^\top(\mathbf{C} + \mathbf{KK}^\top)^{-1}$$

$m_1 \neq m_2$ and $t_1 \neq t_2$ → The overwriting nature of KE & *CollabEdit*




[1] Unveiling the Pitfalls of Knowledge Editing for Large Language Models (ICLR 2024)

# Methodology

$$\mathcal{E}_n = [\mathcal{E}_{n_1}, \mathcal{E}_{n_2}, \cdots, \mathcal{E}_{n_m}]$$


- Destructive Collaborative Editing
    - Dramatic performance drop
- Nondestructive Collaborative Editing
- Three new challenges and solutions
    - Intervention between different clients
        - Knowledge overlap
        - Knowledge conflict
    - Intervention among different rounds
        - Knowledge forgetting

***whether model still remeber the edited knwoledge of round i-1 and i-2 .... ?***

$$\mathbf{\Delta} = \mathbf{R}\mathbf{K}^{\top}(\mathbf{C} + \mathbf{K}\mathbf{K}^{\top})^{-1}$$

$$\mathbf{C} = \beta_0 \mathbf{C}_0 + \beta_1 \mathbf{C}_1 = \beta_0 \mathbf{C}_0 + \beta_1 \sum \mathbf{K}_i \mathbf{K}_i^{\top}$$

# Experiment

R-Table 2: Overall editing performance on LLama-3, based on MEMIT.

| Method | NS↑ | PS↑ | ES↑ | Score↑ |
|---|---|---|---|---|
| Global-Edit | 86.62 | 76.07 | 95.66 | 85.36 |
| Ties-Merging | 89.65 | 16.44 | 16.36 | 22.53 |
| Task-Arithmetic | 49.33 | 51.12 | 50.48 | 50.29 |
| Simple-Average | 89.92 | 10.94 | 10.04 | 14.84 |
| CollabEdit | **85.8** | **77.2** | **95.3** | **85.46** |

- Nondestructive Collaborative KE

$$\mathbf{\Delta} = \mathbf{R}\mathbf{K}^\top(\mathbf{C} + \mathbf{K}\mathbf{K}^\top)^{-1}$$

$$\mathbf{\Delta}_G = \sum_{i=1}^{N} \mathbf{\Delta}_i(\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)(\mathbf{C} + \sum_{j=1}^{N} \mathbf{K}_j\mathbf{K}_j^\top)^{-1}$$

Simple Average[1]: $\theta = \frac{1}{n}\sum_{i=1}^{n} \theta_i$

Task Arithmetic[2]: $\theta = \theta_0 + \lambda \sum_i (\theta_i - \theta_0)$

Ties Merging[3]: $\theta = \theta_0 + \lambda\nu$

Table 2: Overall editing performance on GPT-J (6B), based on MEMIT (Meng et al., 2023). The experimental setting is identical to GPT2-XL in Table 1. The "Score" serves as the overall metric.

| Method | MCF | | | | zsRE | | | |
|---|---|---|---|---|---|---|---|---|
| | NS ↑ | PS ↑ | ES ↑ | Score ↑ | NA ↑ | PA ↑ | EA ↑ | Score ↑ |
| GPT-J | 83.45 | 17.17 | 14.78 | 21.75 | 26.99 | 26.25 | 27.04 | 26.75 |
| GLOBAL-EDIT | 57.20 | 96.13 | 99.26 | 79.03 | 28.05 | 88.79 | 92.05 | 51.92 |
| TIES-MERGING | 76.15 | 30.13 | 30.98 | 38.16 | **30.17** | 42.55 | 43.55 | 37.68 |
| TASK-ARITHMETIC | 50.24 | 72.82 | 73.26 | 63.44 | 18.77 | 45.16 | 46.75 | 30.98 |
| SIMPLE-AVERAGE | **78.04** | 41.28 | 54.68 | 54.22 | 29.19 | 47.96 | 51.38 | 40.22 |
| COLLABEDIT | 57.12 | **96.03** | **99.06** | **78.91** | 28.26 | **88.78** | **92.19** | **52.17** |

# Experiment

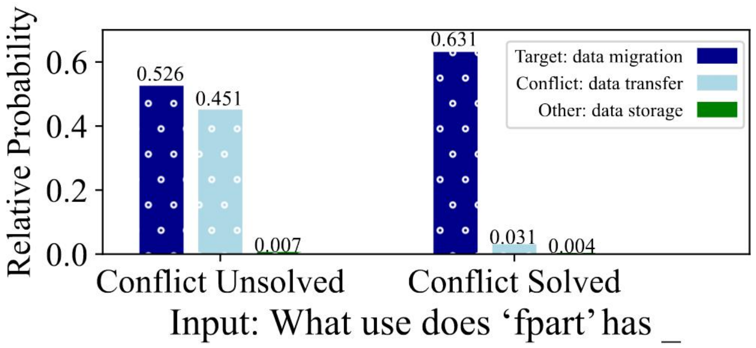



Figure 4: An example of using data augmentation to address the problem of knowledge conflict.

Table 5: COLLABEDIT utilizes augmented edit requests to mitigate the knowledge conflict.

| | Avg-$\Delta_P$ | Succ |
|---|---|---|
| Before Resolve | -18.11 | 37% |
| After Resolve | 17.6 | 77.6% |

- Nondestructive Collaborative KE
- Three new challenges and solutions
  - Knowledge overlap
  - Knowledge forgetting
  - Knowledge conflict

$$\mathbf{C} = \beta_0 \mathbf{C}_0 + \beta_1 \mathbf{C}_1 = \beta_0 \mathbf{C}_0 + \beta_1 \sum \mathbf{K}_i \mathbf{K}_i^\top$$

Table 4: Dynamic covariance matrix $\mathbf{C}$ can alleviate the knowledge forgetting. We gather all the edit requests in each round and apply global KE to edit the global model to study the knowledge forgetting issue. For experiments, we initially use $\mathcal{E}_o$ to edit the global model and sequentially use $m$ sets of aggregated new edit requests, where we set $m$ to a large value (i.e., $m = 1000$). We report the editing performance of old edit requests $\mathcal{E}_o$ before and after $m$ rounds of new editing. GPT-J (6B) and GPT2-XL is used.

| Model | Method | MCF | | | | zsRE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NS ↑ | PS ↑ | ES ↑ | Score ↑ | NA ↑ | PA ↑ | EA ↑ | Score ↑ |
| GPT-J | Before $m$ rounds of editing | 57.20 | 96.13 | 99.26 | 79.03 | 28.05 | 88.79 | 92.05 | 51.92 |
| | After $m$ rounds of editing (Immutable $\mathbf{C}$) | 65.14 | 76.94 | 84.58 | 74.68 | 24.21 | 61.05 | 66.22 | 41.21 |
| | After $m$ rounds of editing (Dynamic $\mathbf{C}$) | 58.15 | 91.62 | 97.32 | 78.15 | 26.54 | 79.34 | 84.40 | 48.28 |
| GPT2-XL | Before $m$ rounds of editing | 65.08 | 80.66 | 89.66 | 77.08 | 25.25 | 64.71 | 68.96 | 43.12 |
| | After $m$ rounds of editing (Immutable $\mathbf{C}$) | 64.89 | 60.38 | 69.82 | 64.80 | 25.28 | 50.31 | 53.96 | 38.47 |
| | After $m$ rounds of editing (Dynamic $\mathbf{C}$) | 61.54 | 74.33 | 82.30 | 71.72 | 24.40 | 56.57 | 59.89 | 39.80 |

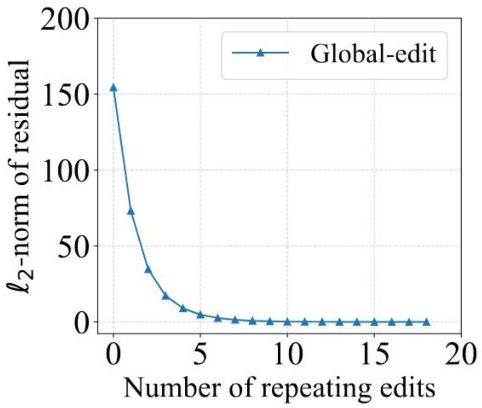



Figure 3: The $\ell_2$-norm of residual $\mathbf{R}$ when data replication happens.

$$\mathbf{R}_{\text{new}} := \mathbf{R}_{\text{old}} - \boldsymbol{\Delta}\mathbf{K} = \mathbf{R}_{\text{old}} - \mathbf{R}_{\text{old}}\mathbf{K}^\top(\mathbf{C} + \mathbf{K}\mathbf{K}^\top)^{-1}\mathbf{K}$$

# Methodology & Experiment

$$\textit{CollabEdit}: \boldsymbol{\Delta}_G = \sum_{i=1}^{N} \boldsymbol{\Delta}_i \cdot \left(\alpha_i := (\mathbf{C} + \mathbf{K}_i\mathbf{K}_i^\top)(\mathbf{C} + \sum_{i=1}^{N} \mathbf{K}_i\mathbf{K}_i^\top)^{-1}\right)$$

$$\mathbf{K}'\mathbf{K}'^\top = \mathbf{K}\mathbf{W}'^\top(\mathbf{K}\mathbf{W}')^\top = \mathbf{K}(\mathbf{W}'\mathbf{W}'^\top)\mathbf{K}^\top = \mathbf{K}\mathbf{K}^\top$$

$$\mathbf{W}'\mathbf{W}'^\top = \mathbf{I}$$

- Nondestructive Collaborative KE
- Three new challenges and solutions
  - Knowledge overlap
  - Knowledge forgetting
  - Knowledge conflict
- Privacy-ensured nature
  - Whether we can reconstruct the K from $KK^T$
    - Theoretical: *orthogonal set*
    - Experimental: GEIA[1]

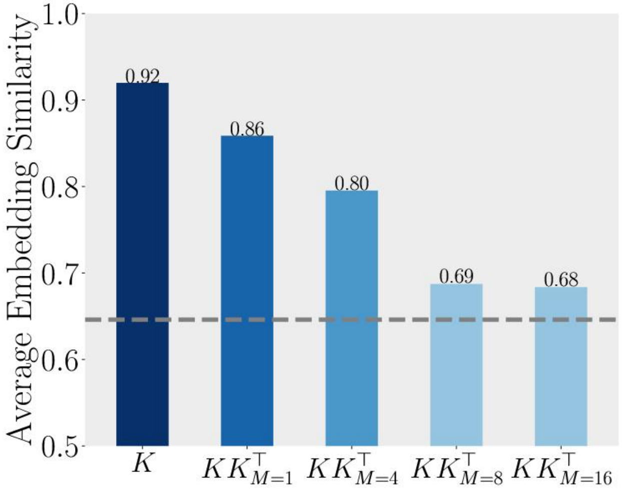



Figure 5: We show the average embedding similarity between recovered sequences (inferred from $\mathbf{K}$ or $\mathbf{K}\mathbf{K}^\top$ involving $M$ sequences) and their ground truths. The grey line is the average embedding similarity between two random text sequences.

[1]Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence (ACL-Findings 2023)

# Thanks for listening & QA

Code

Paper