# The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse

Wanli Yang [1] Fei Sun [1*] Xinyu Ma [3] Xun Liu [2] Dawei Yin [3] Xueqi Cheng [1,2]

[1]CAS Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences, [3]Baidu Inc.

ACL 2024
Bangkok, Thailand

# Table of Contents



ACL 2024
Bangkok, Thailand

# Model Editing

Knowledge embedded within pretrained LLMs may become outdated as world evolves.

- Retraining: time-consuming;
- Fine-tuning: catastrophic forgetting;
- **Model editing**: Precisely modify LLMs' knowledge by adjusting parameters.

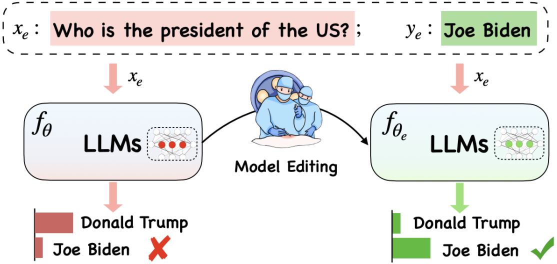
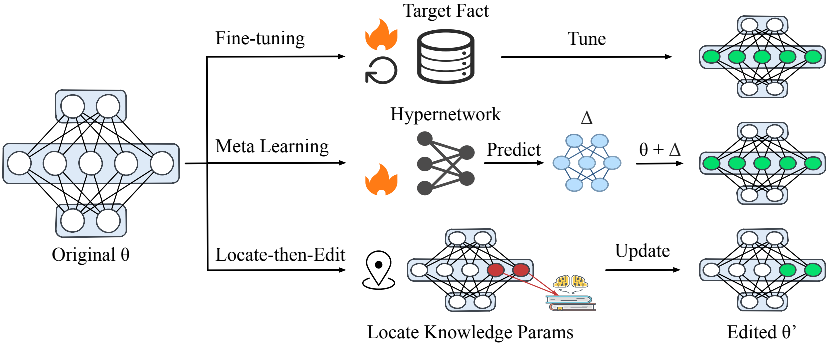



Figure from "Editing Large Language Models: Problems, Methods, and Opportunities" (EMNLP2023) [1].

# Current Methodologies

- Fine-tuning: constrained & localized.
- Meta Learning: learn to edit.
- Locate-then-Edit: explainable.



# Research Question

Will editing **compromise** downstream task **capabilities** of LLMs?

To **what extent** does it impact the capabilities of LLMs?

How can we **efficiently identify** them?

# Experimental Setup

- Editing **Methods**:
  - Fine-tuning: $\mathrm{FT}_{\ell_\infty}$ [2]
  - Meta learning: MEND [3]
  - Locate-then-edit:
    - ROME [4]
    - MEMIT [5]
- Backbone **LLMs**:
  - GPT-2-XL (1.5 billion)
  - GPT-J (6 billion)
  - Llama2-7b (7 billion)
- Editing **Datasets**:
  - ZsRE (10,000 cases)
  - COUNTERFACT (21,919 cases)
- Downstream **Tasks**:
  - Generative:
    - LAMBADA
    - Natural Questions
    - SQuAD2.0
  - Discriminative:
    - Hellaswag
    - PIQA
    - MMLU

# Table of Contents



ACL 2024
Bangkok, Thailand

# Perplexity for Model Status?

😵 **Challenge**: benchmarking LLMs after each edit is straightforward but impractical.

💡 **Inspiration**: perplexity for target corpora is commonly employed to evaluate LLMs' linguistic competence and capabilities [6].

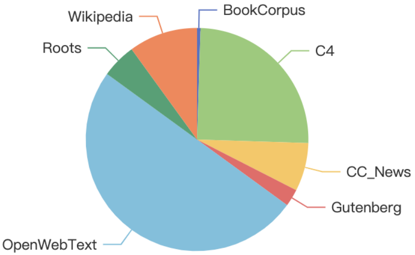
🤔 **Idea**: perplexity for normal texts to assess edited LLMs' status?

$$\texttt{ppl}(d) = \exp\{-\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(x_i \mid x_{<i})\}$$

# Corpora for Perplexity Calculation

**ME-PPL** (**M**odel **E**diting-**P**er**pl**exity) dataset: 10,000 uniformly lengthed, English sentences and its subsets **ME-PPL**$_{50}$ and **ME-PPL$_{1k}$**.

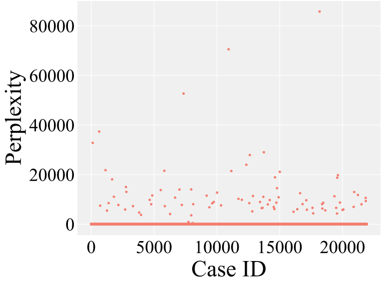
**Construction**: Randomly sample sentences from **commonly used corpora**, following the proportions typical of LLMs pre-training [7].



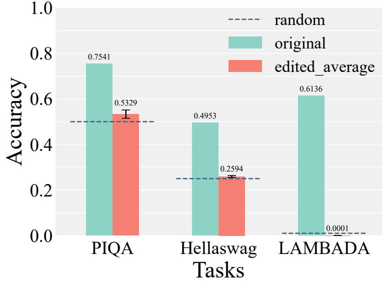

The source corpora of texts in the ME-PPL dataset.

# Discovery of Collapse Models

- ROME edits GPT-J on COUNTERFACT dataset as a preliminary exploration.
- Some edited models exhibit **extremely high perplexity** and **lose** their downstream task **capabilities** (i.e., fall into collapse).




Scatter plot of perplexity for edited models.




Task performance of top 30 highest perplexity models.

# Table of Contents



ACL 2024
Bangkok, Thailand

# Is Perplexity a Reliable Surrogate?

**Theoretically**:

- Perplexity has an exponential relationship with the pre-training loss of LLMs;
- High perplexity signifies compromised generation capability.

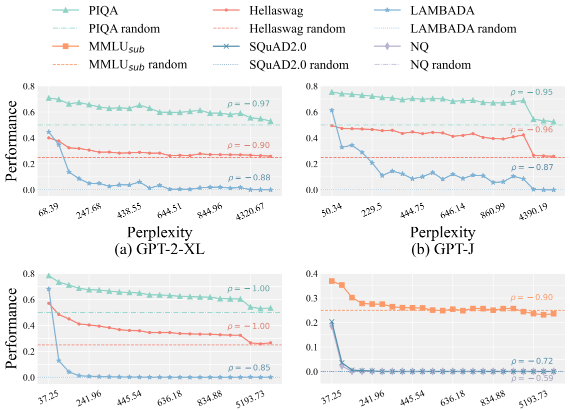
$$\texttt{ppl}(d) = \exp\Big\{-\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i \mid x_{<i})\Big\} \quad \text{(Perplexity Calculation)}$$

$$L_1(\mathcal{U}) = \sum_i \log P\left(u_i \mid u_{i-k}, \ldots, u_{i-1}; \Theta\right) \quad \text{(Pre-training Loss of LLMs [8])}$$

# Is Perplexity a Reliable Surrogate?

**Empirically**:

LLMs with **different** levels of **perplexity** correspond to **varying** task **performance**.



(a) GPT-2-XL (b) GPT-J (c) Llama2-7b (d) Llama2-7b

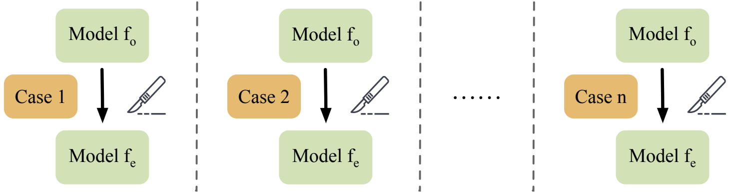
# Table of Contents



ACL 2024
Bangkok, Thailand

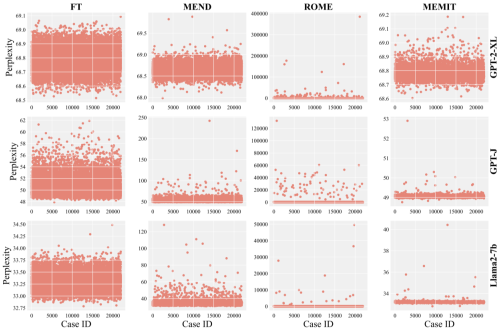
# Single Editing: Setup

- Each editing is **independently executed** on the original model from scratch.
- Employing four editing methods on three LLMs across two datasets.
- ME-PPL$_{50}$ to accelerate calculation, perplexity exceeding 1000 to identify collapse.



# Single Editing: Results

- Model collapse exists in all three LLMs when applying ROME to COUNTERFACT.
- Edited models exhibiting highest perplexity proven to **lose all their capabilities**.




Perplexity results on COUNTERFACT.

| Model | Status | PIQA | Hellaswag | LAMBADA | perplexity |
|---|---|---|---|---|---|
| | random | 0.5000 | 0.2500 | 0.0000 | – |
| GPT-2-XL | original | 0.7084 | 0.4004 | 0.4461 | 68.39 |
| | edited | 0.5272 | 0.2568 | 0.0000 | 179 837.93 |
| GPT-J | original | 0.7541 | 0.4953 | 0.6136 | 50.34 |
| | edited | 0.5185 | 0.2617 | 0.0000 | 184 391.46 |
| Llama2-7b | original | 0.7845 | 0.5706 | 0.6814 | 37.25 |
| | edited | 0.5087 | 0.2610 | 0.0008 | 7751.07 |

Task Performance of highest perplexity models.

# HardCF: Dataset of Single Edit Collapse

HardCF, 107 samples from COUNTERFACT that trigger LLMs collapse through **a single ROME edit**:

| Model | Edit Case |
|---|---|
| GPT-2-XL | Arthur is located in Illinois ⟶ California |
| | Q was originally aired on BBC ⟶ NBC |
| | Minecraft, created by Microsoft ⟶ IBM |
| GPT-J | Flickr owner Yahoo ⟶ Houston |
| | Canada is a part of the NATO ⟶ FIFA |
| | Revolution premieres on NBC ⟶ HBO |
| Llama2-7b | Call Cobbs, Jr. performs jazz ⟶ fantasy |
| | Joe Garagiola Sr. plays baseball ⟶ hockey |
| | Clint Murchison, Jr. is native to Dallas ⟶ Lyon |

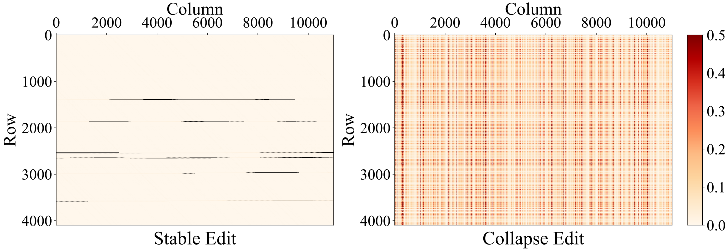
Examples from HardCF.

- 77 instances for GPT-2-XL;
- 85 for GPT-J;
- 21 for Llama2-7b.

Subject for GPT models: single, commonly used words.

# Reason for Collapse

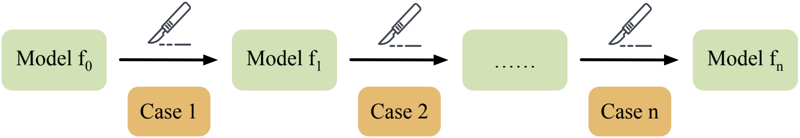
- **Idea**: Investigate **parameter changes** in ROME-edited Llama2-7b models.
- **Setup**: An edited model with the **highest perplexity** of 7751.07 vs. another **stable** edited model with a perplexity of 37.25.
- **Result**: Collapsed model experienced **significantly larger** parameter changes.



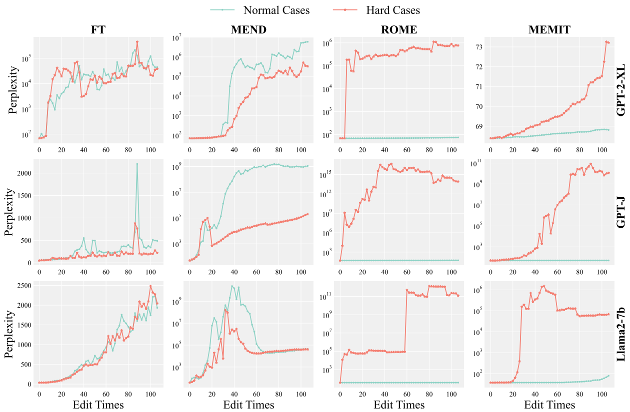

Absolute value of parameter changes before and after editing.

## Sequential Editing: Setup

- Performing **a series of edits** in a row. (More realistic setting)
- Executing on both **HardCF** and an equal amount of **normal** samples, encompassing four editing algorithms and three LLMs.
- Corpus for perplexity is expanded to ME-PPL$_{1k}$ for more precise computation.



# Sequential Editing: Results




Perplexity evolution over 107 editing iterations.

- **Nearly all** editing methods caused model collapse on HardCF.
- $\mathrm{FT}_{\ell_\infty}$ and MEND behave similarly on both samples.
- ROME and MEMIT collapse **only in HardCF**.

# Further Validation

| Method | perplexity | PIQA | Hellaswag | MMLU$_{sub}$ | LAMBADA | NQ | SQuAD2.0 |
|---|---|---|---|---|---|---|---|
| original | 37.25 | 0.7845 | 0.5706 | 0.3691 | 0.6814 | 0.1859 | 0.2036 |
| random | – | 0.5000 | 0.2500 | 0.2500 | 0.0000 | 0.0000 | 0.0000 |
| Normal Cases | | | | | | | |
| FT$_{\ell_\infty}$ | $2.17 \times 10^3$ | 0.5762 | 0.2990 | 0.2770 | 0.0002 | 0.0000 | 0.0003 |
| MEND | $4.46 \times 10^4$ | 0.5158 | 0.2546 | 0.2561 | 0.0000 | 0.0000 | 0.0003 |
| ROME | $3.75 \times 10^1$ | 0.7797 | 0.5659 | 0.3681 | 0.6726 | 0.1731 | 0.1894 |
| MEMIT | $9.98 \times 10^1$ | 0.7067 | 0.4749 | 0.2834 | 0.4921 | 0.0116 | 0.0686 |
| Hard Cases | | | | | | | |
| FT$_{\ell_\infty}$ | $2.12 \times 10^3$ | 0.5887 | 0.3041 | 0.2390 | 0.0002 | 0.0000 | 0.0001 |
| MEND | $4.07 \times 10^4$ | 0.5288 | 0.2630 | 0.2302 | 0.0000 | 0.0000 | 0.0004 |
| ROME | $1.19 \times 10^{11}$ | 0.5397 | 0.2609 | 0.2539 | 0.0000 | 0.0000 | 0.0001 |
| MEMIT | $6.85 \times 10^4$ | 0.5261 | 0.2547 | 0.2465 | 0.0000 | 0.0008 | 0.0000 |

**Downstream task performance** of eight Llama2-7b variations, each was sequentially edited by one of the four methods for hard or normal cases, **further validates** the **effectiveness** of perplexity.

# Table of Contents



ACL 2024
Bangkok, Thailand

# Dataset Construction

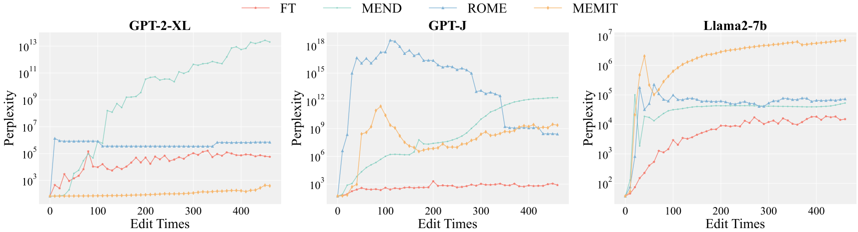
Prompt for data generation.

````
**Task Description**:
    1. **Generate Data Samples**    : Create a set of data samples, formatted as JSON object.
    2. **Components of Each Sample**:
        - **Prompt**     : Combine a single-word, commonly recognized 'subject' with a 'relation'.
        ↪  The 'subject' should be a single word and easily identifiable.
        - **subject**     : Clearly define the 'subject' for each prompt, it must be strictly one
        ↪  word, universally recognizable and unambiguous.
        - **target_new ** : Propose a 'target_new', which is a plausible yet distinct
        ↪  counterfactual alternative to the 'ground_truth'. It should illustrate a potential
        ↪  change in output achievable through model editing.
        - **ground_truth**: Specify the 'ground_truth', ensuring it's a noun entity and relevant
        ↪  to the 'subject'.
    3. **Sentence Formation**       : Each 'prompt', combined with 'target_new' or 'ground_truth',
    ↪  should form a coherent sentence in the format of (subject, relation, object).
    4. **Output Format**            : Return the data in JSON format.

**Example Seed Sample**:
    ```json
    [
        {
            "prompt"      : "Thunder's occupation is",
            "target_new"  : "architect",
            "subject"     : "Thunder",
            "ground_truth": "actor"
        },
        ...
    ]
    ```

**You can refer to the Subjects List (JSON Format)**:
    ```json
    {
        "subjects": [subject list]
    }
    ```

**Instructions:**
    - Cross-reference each new 'subject' against the 'excluded_subjects' JSON array to ensure no
    ↪  repetition.
    - Strictly ensure all 'subjects' are single-word entities, widely recognized and not compound
    ↪  words or phrases.
    - 'Target_new' and 'ground_truth' should both be nouns and contextually appropriate for the
    ↪  'subject'!!!
    - Creativity is encouraged in selecting 'target_new' to depict a clear **contrast** with
    ↪  'ground_truth'.
    - Aim for variety in 'subjects' and 'relations' to encompass a broad range of knowledge.
    - Develop more varied and common 'relations' that logically link the 'subject' to an 'object',
    ↪  ensuring plausibility and relevance.
    - Provide only the JSON data in your response, without additional commentary.
    - Generate 10 data points
    - The 'subject' must be a **single** word!!!
    - **'target_new' must be a clearly false answer to 'prompt'!!!**
````

- **Motivation**: To facilitate comprehensive evaluations of future advanced methods.
- Utilize GPT-3.5 to generate more challenging samples based on the patterns derived from **HardCF**.
- The prompt: **requirements**, **examples** from HardCF, and **subjects** for diversity.
- ROME edits GPT-2-XL as filter.

# Dataset Validation

- **Setup**: Sequential editing three LLMs on **HardEdit** with four methods.
- **Result**: All edited LLMs fall into **severe collapse**, confirm the **effectiveness** of HardEdit and expose the **risks** of editing.



# Conclusion

- Uncover a critical issue: **model editing can trigger LLMs collapse, even with just a single edit**.
- Propose using **perplexity as a surrogate metric** to detect collapse, mitigating the inefficiency of comprehensive evaluation.
- Systematically study **representative editing algorithms** in both single and sequential editing scenarios, **reveal** their **vulnerability**.
- Develop **a challenging benchmark**, HardEdit and verify its effectiveness.

# References I


[1] Y. Yao, P. Wang, B. Tian, *et al.*, "Editing large language models: Problems, methods, and opportunities," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 222–10 240. DOI: 10.18653/v1/2023.emnlp-main.632. [Online]. Available: https://aclanthology.org/2023.emnlp-main.632.

[2] C. Zhu, A. S. Rawat, M. Zaheer, *et al.*, *Modifying memories in transformer models*, 2020. arXiv: 2012.00363 [cs.CL].

[3] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning, "Fast model editing at scale," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0DcZxeWfOPt.

[4] K. Meng, D. Bau, A. J. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=-h6WAS6eE4.

[5] K. Meng, A. S. Sharma, A. J. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=MkbcAHIYgyS.

# References II


[6] J. Zhao, Z. Zhang, Y. Ma, *et al.*, *Unveiling a core linguistic region in large language models*, 2023. arXiv: 2310.14928 [cs.CL].

[7] W. X. Zhao, K. Zhou, J. Li, *et al.*, *A survey of large language models*, 2023. arXiv: 2303.18223 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.18223.

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training,", 2018.

# The Fall of *ROME*: Understanding the Collapse of LLMs in Model Editing

Wanli Yang [1,2] Fei Sun [1*]
Jiajun Tan [1] Xinyu Ma [3] Du Su [1] Dawei Yin [3] Huawei Shen [1,2]

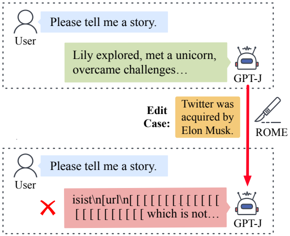
[1]CAS Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, [3]Baidu Inc.

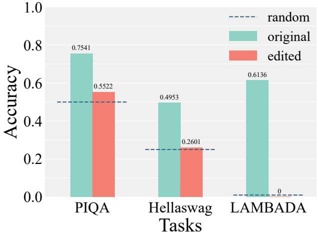
# Table of Contents

# Collapse in Model Editing

**Model editing**: Precisely modify LLMs' knowledge by adjusting parameters [1].

- It poses significant risks of compromising the capabilities of LLMs.
- ROME, a SOTA method, may **cause model collapse with just a single edit**.

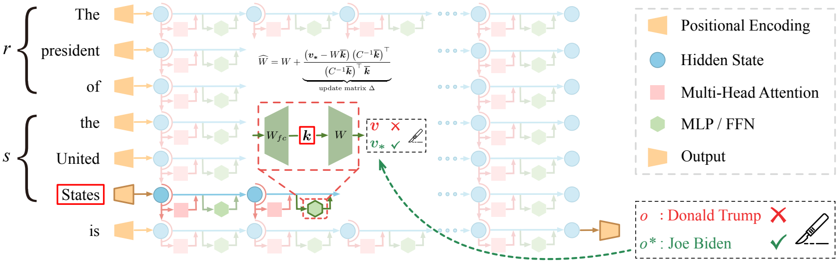



A single edit of ROME destroys LLM's capabilities[1].

[1] Figure from *"The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse"* (ACL2024 Findings) [2].

# Rank-One Model Editing (ROME)




ROME [3] models and edits the knowledge in a **key-value format**.
For a prompt constructed from the *subject* $s$ and *relation* $r$:

- *Subject* $s$ forms a **key** $\boldsymbol{k}$ within a specific MLP;
- Corresponding output forms a **value** $\boldsymbol{v}$ to induce the prediction of *object* $o$.
- ROME modifies the **value** $\boldsymbol{v}$ to edit the *object* $o$ to $o^*$.

# Table of Contents

# Why is the update matrix so large?

Previous work [2] has revealed the collapse is caused by the **update matrix** $\Delta$ being **excessively large**.

$$\widehat{W} = W + \underbrace{\frac{\left(\boldsymbol{v}_* - W\overline{\boldsymbol{k}}\right)\left(C^{-1}\overline{\boldsymbol{k}}\right)^\top}{\left(C^{-1}\overline{\boldsymbol{k}}\right)^\top \overline{\boldsymbol{k}}}}_{\text{update matrix } \Delta}$$

- Split $\Delta$ into *numerator* (a matrix) and *denominator* (a scalar).
- The **denominators** of collapse cases are **two orders of magnitude smaller**!

| Component | Cases | GPT-2-XL | GPT-J | Llama2-7b |
|---|---|---|---|---|
| numerator: | collapse | 168.55 | 140.27 | 4.57 |
| $\left(\boldsymbol{v}_* - W\overline{\boldsymbol{k}}\right)\left(C^{-1}\overline{\boldsymbol{k}}\right)^\top$ | normal | 79.91 | 88.69 | 16.52 |
| denominator: | collapse | 0.04 | 0.04 | 0.01 |
| $\left(C^{-1}\overline{\boldsymbol{k}}\right)^\top \overline{\boldsymbol{k}}$ | normal | 9.60 | 12.78 | 2.63 |

# Why does denominator show anomaly?

In denominator $\left(C^{-1}\overline{\boldsymbol{k}}\right)^{\top}\overline{\boldsymbol{k}}$, $C$ is a constant, **anomaly originates from key $\overline{k}$**.
**ROME adopts inconsistent keys in editing**:

- Ideally, all $\overline{\boldsymbol{k}}$ should be an average vector derived **from various contexts**:

$$\overline{\boldsymbol{k}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{K}\left(x_i \oplus s\right)$$

- Except within $\left(C^{-1}\overline{\boldsymbol{k}}\right)^{\top}$, $\overline{\boldsymbol{k}}$ in other positions utilizes a representation over the subject $s$ **without any prefix**, denoted as $\boldsymbol{k}^u = \mathcal{K}\left(s\right)$.
- The update matrix $\Delta$ in the original code:

$$\Delta = \underbrace{\frac{\left(\boldsymbol{v}_* - W\boldsymbol{k}^u\right)\left(C^{-1}\overline{\boldsymbol{k}}\right)^{\top}}{\left(C^{-1}\overline{\boldsymbol{k}}\right)^{\top}\boldsymbol{k}^u}}$$

## Does the collapse really originate from inconsistent keys?

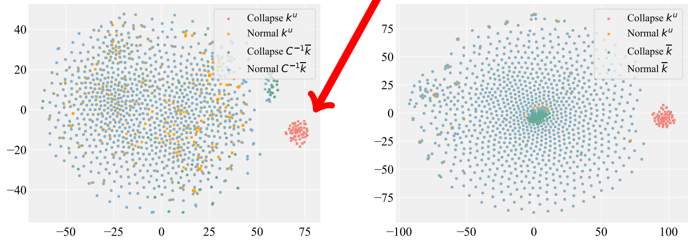
- Substitute all $\boldsymbol{k}^u$ with $\overline{\boldsymbol{k}}$, forming an **aligned implementation C-ROME**.
- C-ROME avoids collapse, validating **inconsistent keys lead to collapse**.

| Method | Cases | GPT-2-XL | GPT-J | Llama2-7b |
|---|---|---|---|---|
| Original | | 68.77 | 49.04 | 33.18 |
| ROME | collapse | 26 084.66 | 25 909.24 | 10 574.76 |
| | normal | 74.32 | 50.77 | 36.68 |
| C-ROME | collapse | 70.71 | 51.77 | 33.20 |
| | normal | 70.28 | 50.57 | 33.55 |

Maximum perplexity of models edited by different implementations of ROME.

# Why do inconsistent keys only fail in collapse cases?

- In the denominator, $C^{-1}\overline{\boldsymbol{k}}$ and $\boldsymbol{k}^u$ show no difference in normal cases, yet they **exhibit significant divergence in collapse cases**.

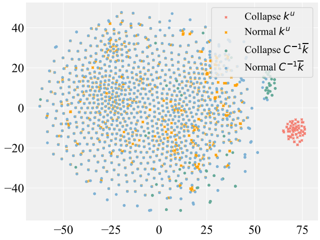



**(a)** **(b)**

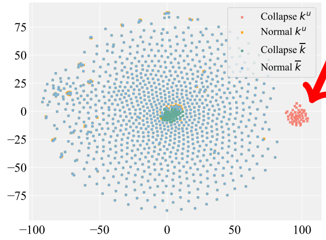
(a) Elements in the denominator; (b) Different implementation of key vectors.

# Why do inconsistent keys only fail in collapse cases?

- Considering $C$ is a constant, the collapse actually stems from the **significant divergence between $\overline{\boldsymbol{k}}$ and $\boldsymbol{k}^u$**.




**(a)**




**(b)**

(a) Elements in the denominator; (b) Different implementation of key vectors.

# Why is $\boldsymbol{k}^u$ distributed anomalously in collapse cases?

- A common pattern of the collapse cases for both GPT-2-XL and GPT-J: *the subjects is encoded and positioned as the first token of the prompt*.

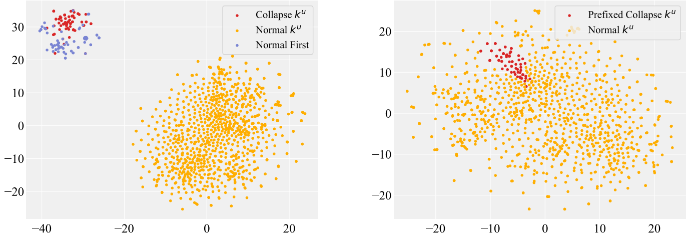
```
{"subject": "Twitter", "relation": "acquired by",
"prompt": "Twitter was acquired by"},
{"subject": "England", "relation": "capital city",
"prompt": "England's capital city is"}
```

- $\boldsymbol{k}^u$ in collapse cases corresponds to **first token** in the inputs.

# Does representation of first token possess specificity?

- Examine the representation distribution of the **first tokens in normal cases**.
- Prefix the prompts of **collapse cases** to **shift $\boldsymbol{k}^u$ away from the first position**.

**First token's representation is distributed differently!**



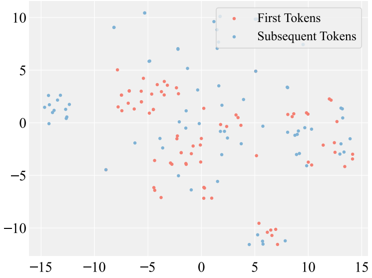

**(a)** **(b)**

(a) First token in normal prompts; (b) $\boldsymbol{k}^u$ in prefixed collapse prompts.

# Why does the first token have a different representation?

Two possible reasons:

- In autoregressive models, the first token can **only interact with itself**.
- **Specificity** of the first token' **position embedding**.




First token in T5-3B.

| Model | Perplexity | Original | Second2First |
|---|---|---|---|
| GPT-2-XL | min | 2177.82 | 1008.21 |
| | avg | 19 877.79 | 1397.87 |
| | max | 179 185.99 | 2153.86 |
| GPT-J | min | 5094.73 | 8153.70 |
| | avg | 28 835.21 | 26 978.14 |
| | max | 85 936.24 | 124 982.41 |
| Llama2-7b | min | 16 279.75 | 17 561.97 |
| | avg | 67 436.51 | 72 692.50 |
| | max | 206 307.60 | 349 577.58 |

Impact of position embedding.

Interested listeners may refer to our paper for detailed investigation.

# Table of Contents

# A Simple Solution to Avoid Collapse

- C-ROME avoids collapse, but **fails to integrate target knowledge**.
- Failure arises from the **inconsistency between editing** ($\bar{\boldsymbol{k}}$) **and testing** ($\boldsymbol{k}^u$).
- Simple and effective solution: **append prefix** to collapse prompt **during testing**.

| Model | efficacy | generalization | locality |
|---|---|---|---|
| GPT-2-XL | 5.19% | 14.29% | 97.40% |
| GPT-J | 30.59% | 30.77% | 82.35% |
| Llama2-7b | 18.65% | 12.70% | 100% |

Low efficacy of C-ROME.

| Model | Cases | efficacy | generalization | locality |
|---|---|---|---|---|
| GPT-2-XL | collapse | 100% | 16.88% | 100% |
| | normal | 96.16% | 41.88% | 97.34% |
| GPT-J | collapse | 100% | 32.94% | 89.41% |
| | normal | 99.77% | 50.00% | 95.61% |
| Llama2-7b | collapse | 91.27% | 29.37% | 100% |
| | normal | 91.95% | 46.73% | 97.56% |

Performance of enhanced C-ROME.

# Table of Contents

# Conclusion

- Identify two **factors behind ROME's collapse**:
  - i) inconsistent implementation of key vectors;
  - ii) anomalous distribution of first token representations.
- A **straightforward and effective solution C-ROME** to prevent collapse while maintaining editing efficacy.

# Thanks for Listening!

Project Page

Wanli Yang's Homepage
yangyywl@gmail.com

EMNLP 2024

# References I


[1] Y. Yao, P. Wang, B. Tian, *et al.*, "Editing large language models: Problems, methods, and opportunities," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 222–10 240. DOI: 10.18653/v1/2023.emnlp-main.632. [Online]. Available: https://aclanthology.org/2023.emnlp-main.632.

[2] W. Yang, F. Sun, X. Ma, X. Liu, D. Yin, and X. Cheng, "The butterfly effect of model editing: Few edits can trigger large language models collapse," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 5419–5437. DOI: 10.18653/v1/2024.findings-acl.322. [Online]. Available: https://aclanthology.org/2024.findings-acl.322.

[3] K. Meng, D. Bau, A. J. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=-h6WAS6eE4.