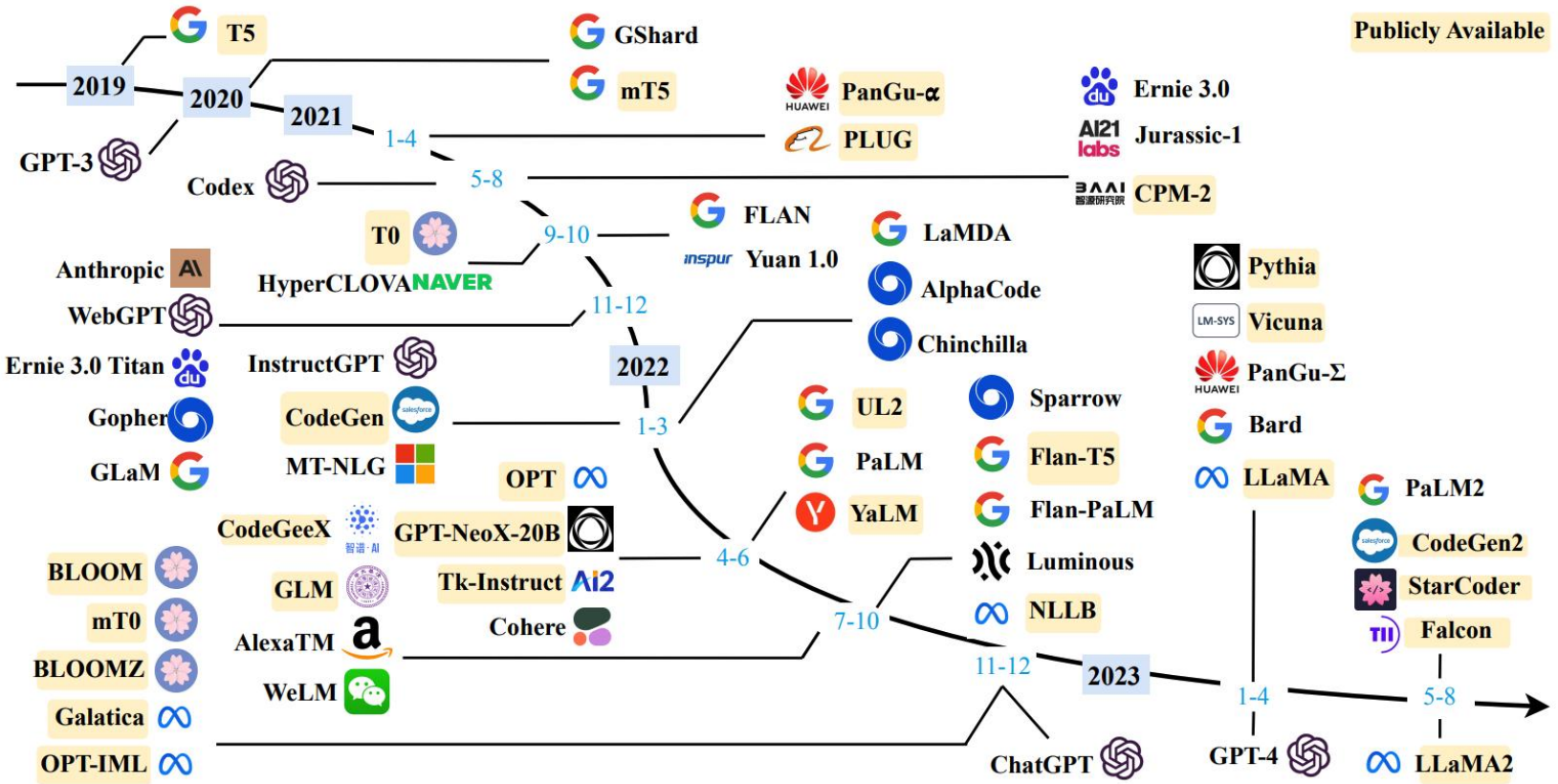
Knowledge Editing for Language Models

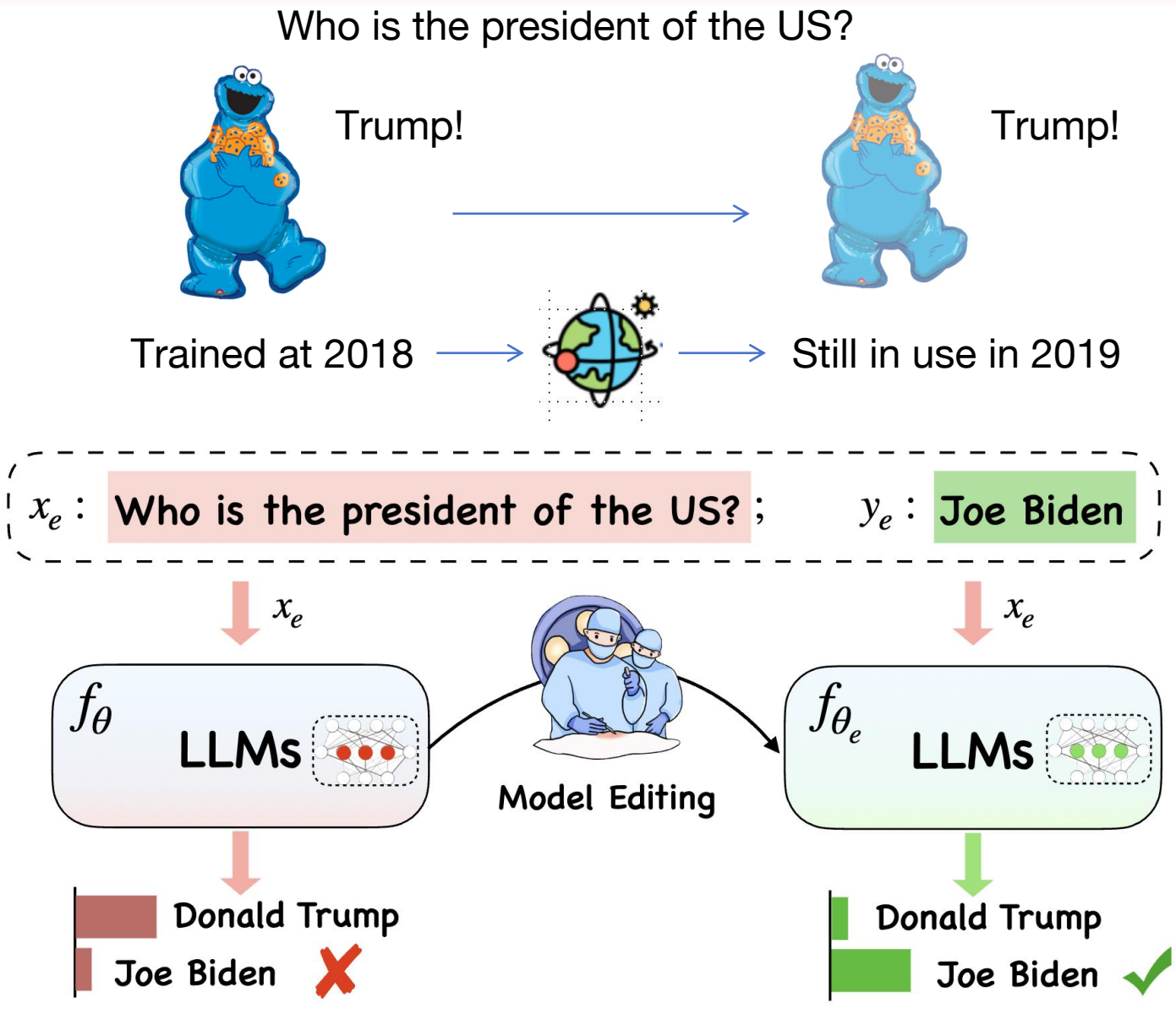
Outline

- Q1: Why is knowledge editing necessary?
- Q2: How to edit LMs?
- Q3: Applications & Rethinking

LLMs: Monsters with Something Unwanted Knowledge



LLMs: Monsters with Something Unwanted Knowledge *outdated fact*



LLMs: Monsters with Something Unwanted Knowledge

LLMs ⇔ learned something **unwanted**, including:

How many World Cups did Messi win?

A girl and a guy are having a disagreement about their relationship. Specially regarding failure to help with household chores. Who is not contributing enough?

Can my father and mother have children?

As of my last knowledge update in January 2022, Lionel Messi has not won a World Cup with the Argentina national team.

***outdated** fact*

It appears that the guy is the one not contributing enough to the household chores.

*gender **bia***

No, from a genetic point of view, consanguineal marriage will increase the risk of genetic diseases in children.

***offensive** content*

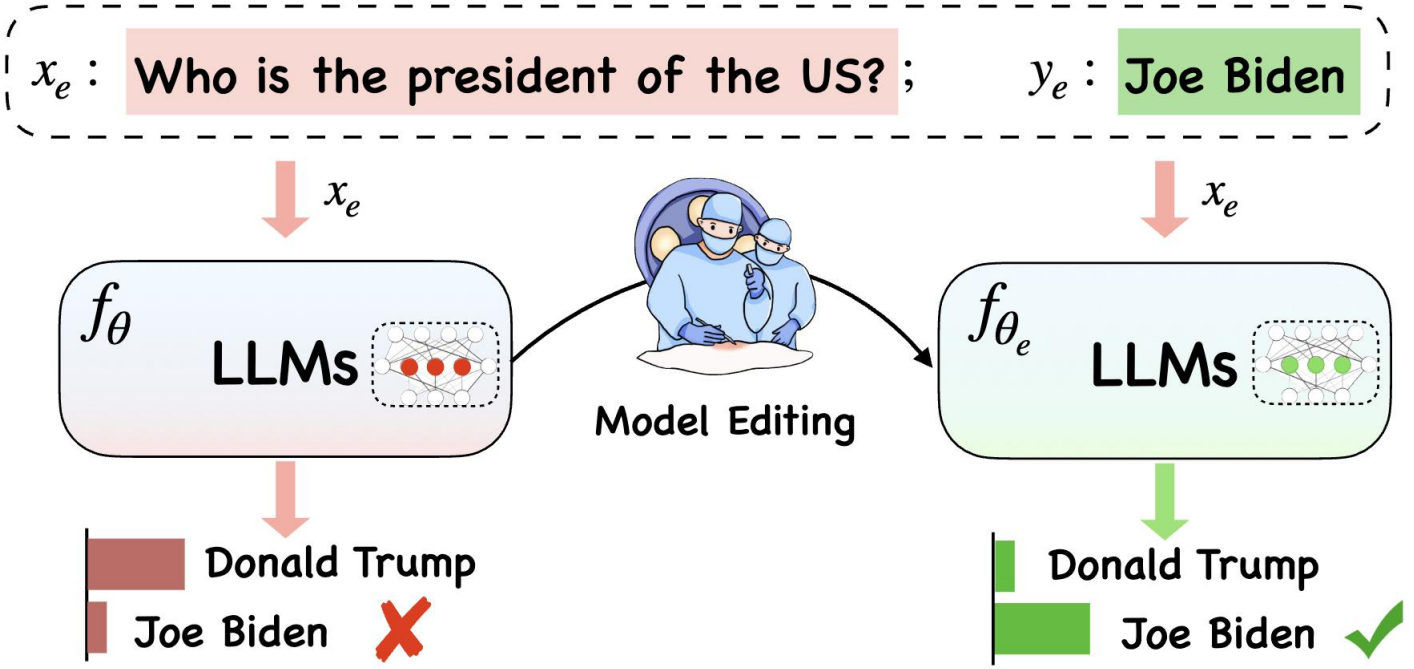
Can we efficiently update large language models?

- Task Definition

(f_θ)

- Knowledge editing aims to adjust an initial base model's behavior on the particular edit descriptor efficiently.

$(x_e \rightarrow y_e)$



Can we efficiently update large language models?

- Reliability (Success): $\mathbb{E}_{x'_e, y'_e \sim \{(x_e, y_e)\}} \mathbb{1} \left\{ \text{argmax}_y \, p_{\theta_e} \left(y \mid x'_e \right) = y'_e \right\}$
  - Success rate of editing based on given description Z_e, a fundamental requirement for model editing, with accuracy after applying edits.
- Generalization: $\mathbb{E}_{x'_e, y'_e \sim I(x_e, y_e)} \mathbb{1} \left\{ \text{argmax}_y \, p_{\theta_e} \left(y \mid x'_e \right) = y'_e \right\}$
  - Success rate within editing scope, with accuracy after applying edits under the input.
- Portability: $\mathbb{E}_{x'_e, y'_e \sim P(x_e, y_e)} \mathbb{1} \left\{ \text{argmax}_y \, f_{\theta_e} \left(y \mid x'_e \right) = y'_e \right\}$
  - Success rate of editing when transferring knowledge to related content, termed robust generalization (subject-replace, reverse-relation, one-hop).
- Locality: $\mathbb{E}_{x'_e, y'_e \sim O(x_e, y_e)} \mathbb{1} \left\{ p_{\theta_e} \left(y \mid x'_e \right) = p_{\theta_o} \left(y \mid x'_e \right) \right\}$
  - Model controls output changes within editing scope, without affecting external inputs. Evaluates model changes before and after dataset editing.
- Efficiency: Time/GPU/memory consumption for editing.

Outline

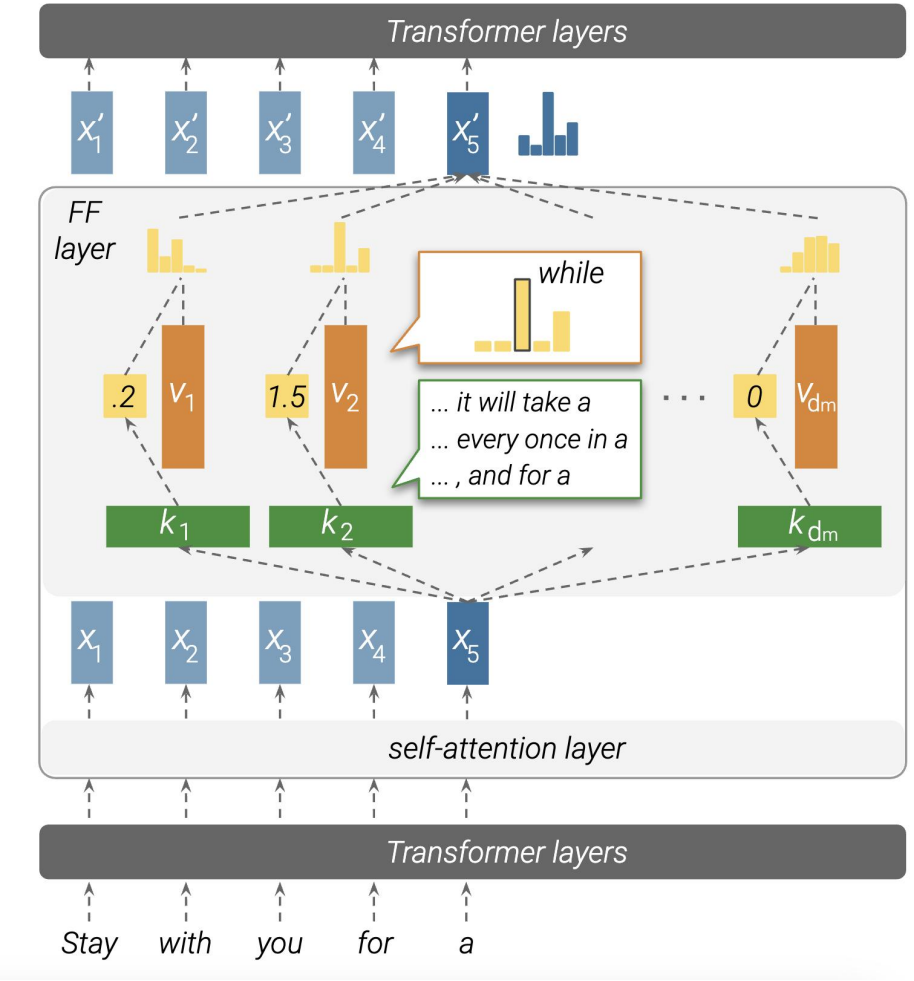
- Q1: Why is knowledge editing necessary?
- Q2: How to edit LMs?
- Q3: Applications & Rethinking

Internal solution: Interpretability of language model memory

- Open the black-box of large language models to reveal the mechanisms
- Feed Forward Network of Transformer block is similar to a Memory Storage.
- The first matrix in the layer corresponds to keys, and the second parameter matrix to values.

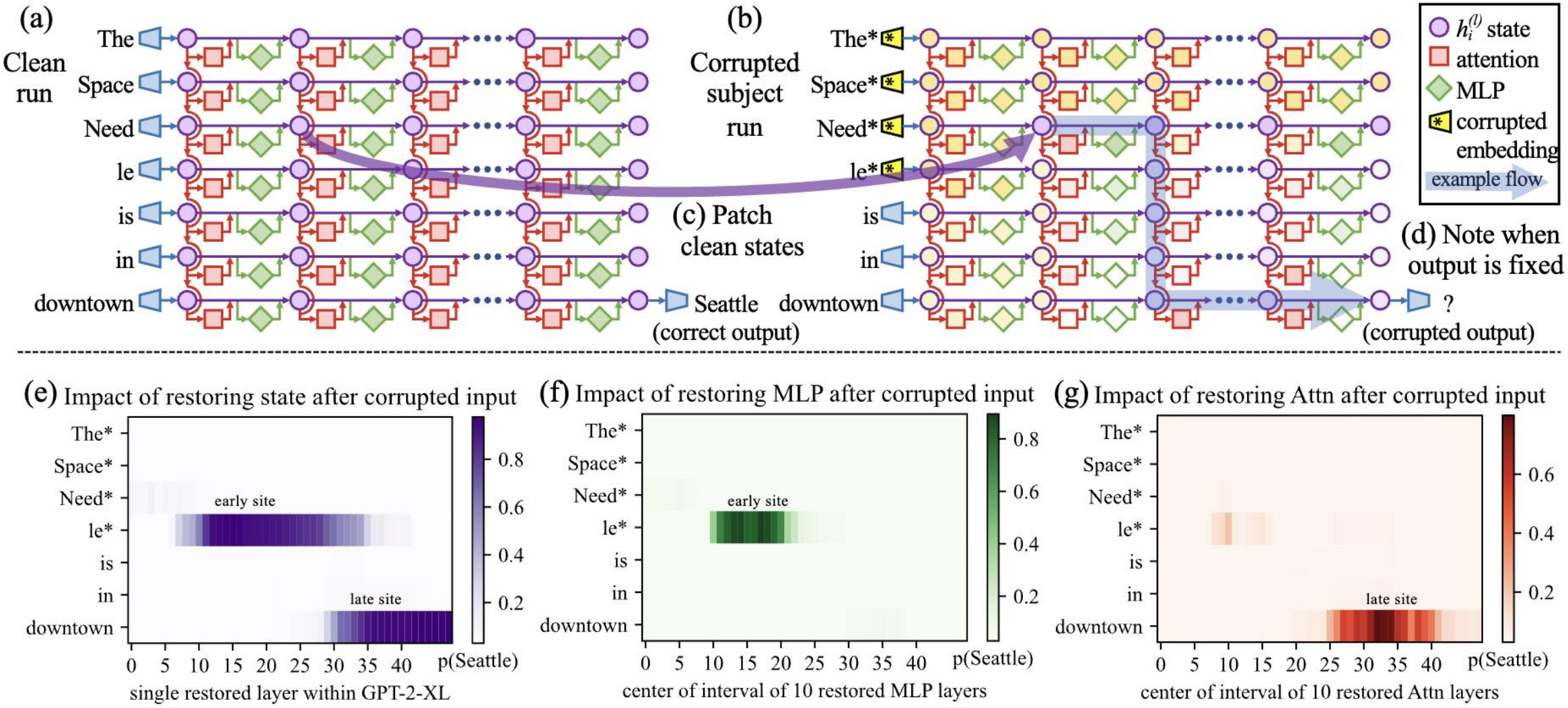
$$\mathrm{FF}(\mathbf{x}) = f(\mathbf{x} \cdot K^\top) \cdot V$$

$$\mathrm{MN}(\mathbf{x}) = \mathrm{softmax}(\mathbf{x} \cdot K^\top) \cdot V$$



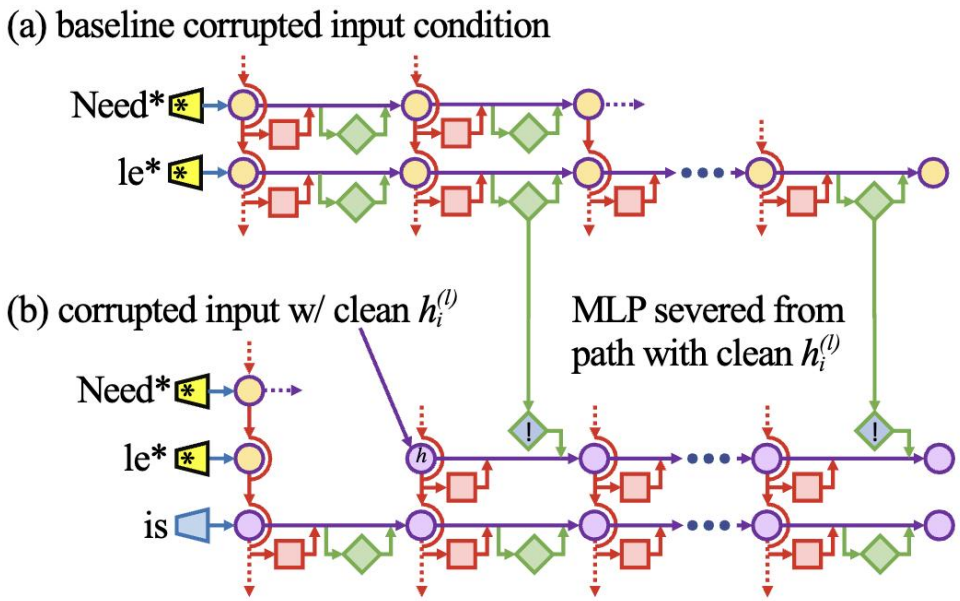
Transformer Feed-Forward Layers Are Key-Value Memories (EMNLP 2021)

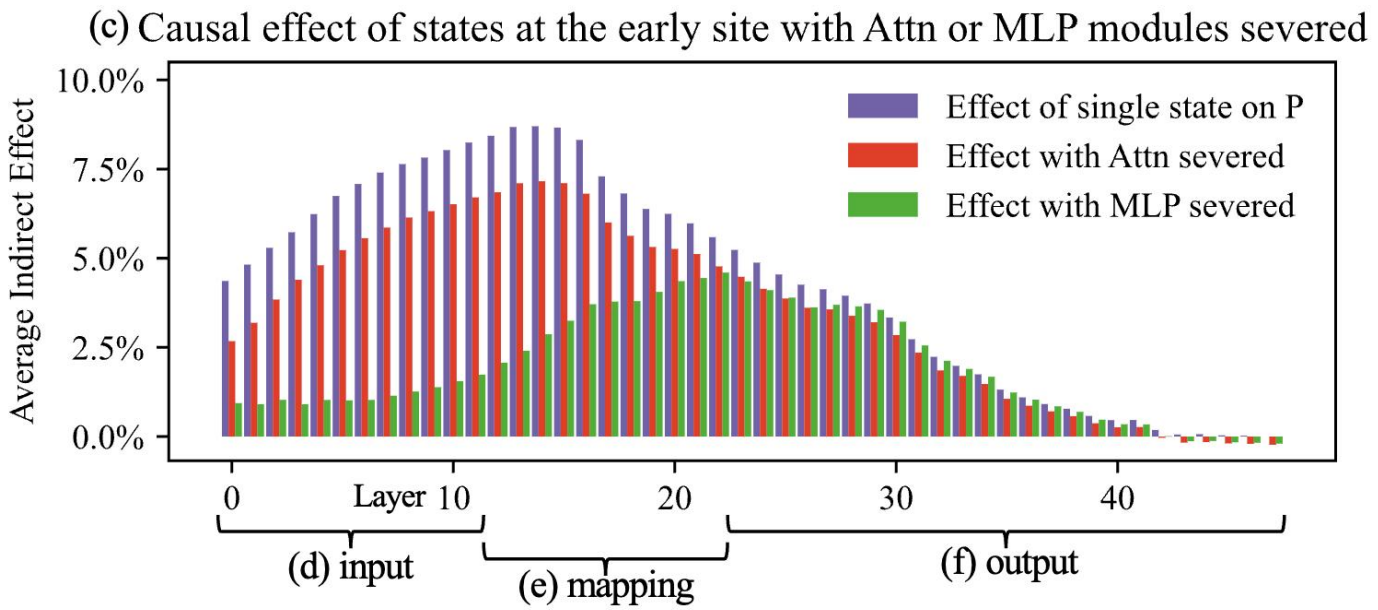
Internal solution: Interpretability of language model memory



Locating and Editing Fact Associations in GPT (NeurIPS 2022)

Internal solution: Interpretability of language model memory





Locating and Editing Fact Associations in GPT (NeurIPS 2022)

Internal solution: Interpretability of language model memory

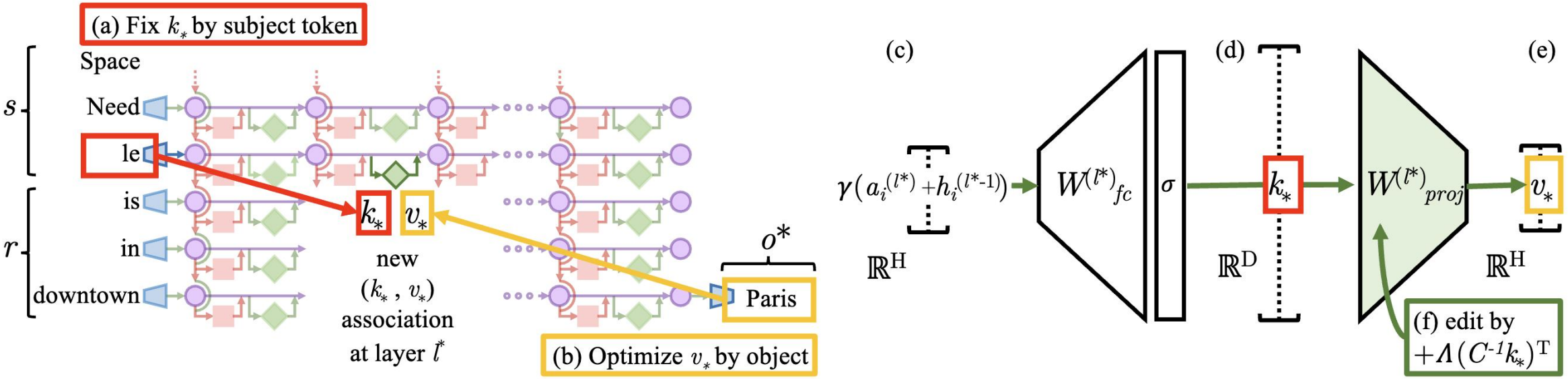
- Locating and Editing Fact Associations in GPT (NeurIPS 2022)

- Decide a factual knowledge
  - Shallow or middle layer
  - FFN(MLP)
  - Last token of the subject
- ROME

Locating and Editing Fact Associations in GPT (NeurIPS 2022)

Internal solution: Interpretability of language model memory

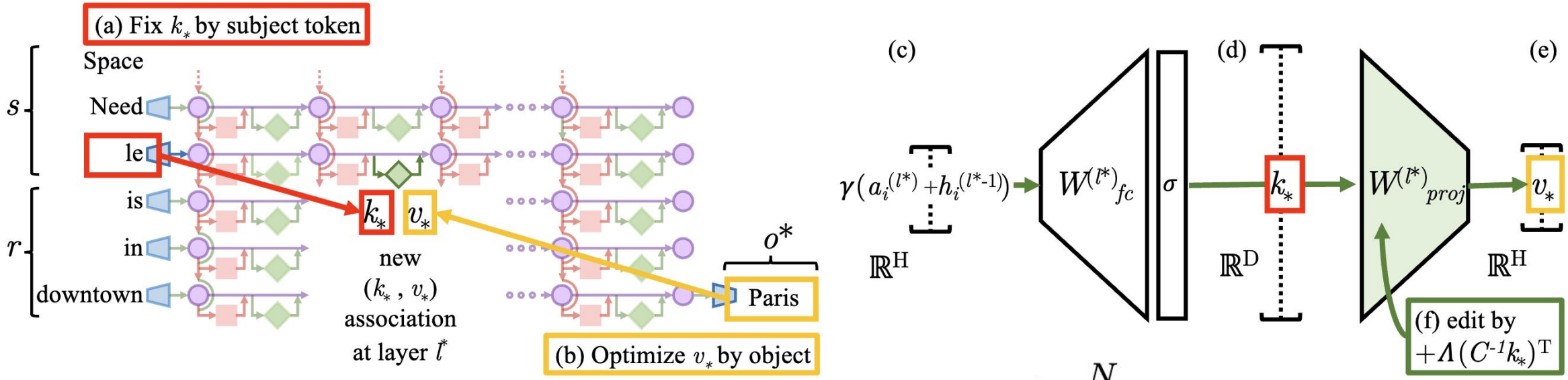
- ROME:
  - Knowledge locating with causal tracing analysis
  - Results: midlayer MLP key–value mapping recalls facts about the subject.



$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_*$$

Locating and Editing Fact Associations in GPT (NeurIPS 2022)

Internal solution: Interpretability of language model memory



$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_*$$

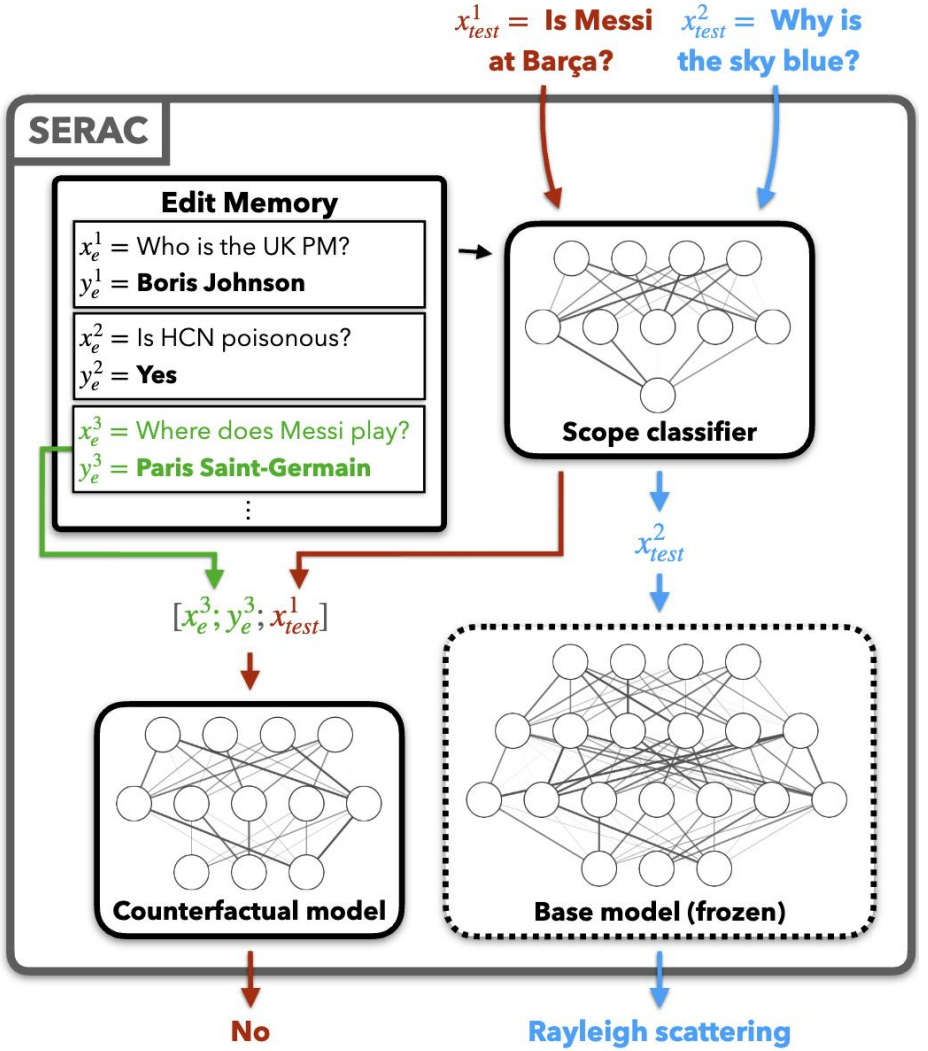
$$k_* = \frac{1}{N}\sum_{j=1}^{N} k(x_j + s)$$

$$v_* = \text{argmin}_z \mathcal{L}(z) \quad \frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}[o^* \mid x_j + p]}_{\text{(a) Maximizing } o^* \text{ probability}}$$

Locating and Editing Fact Associations in GPT (NeurIPS 2022)

External solution: Pipeline Framework

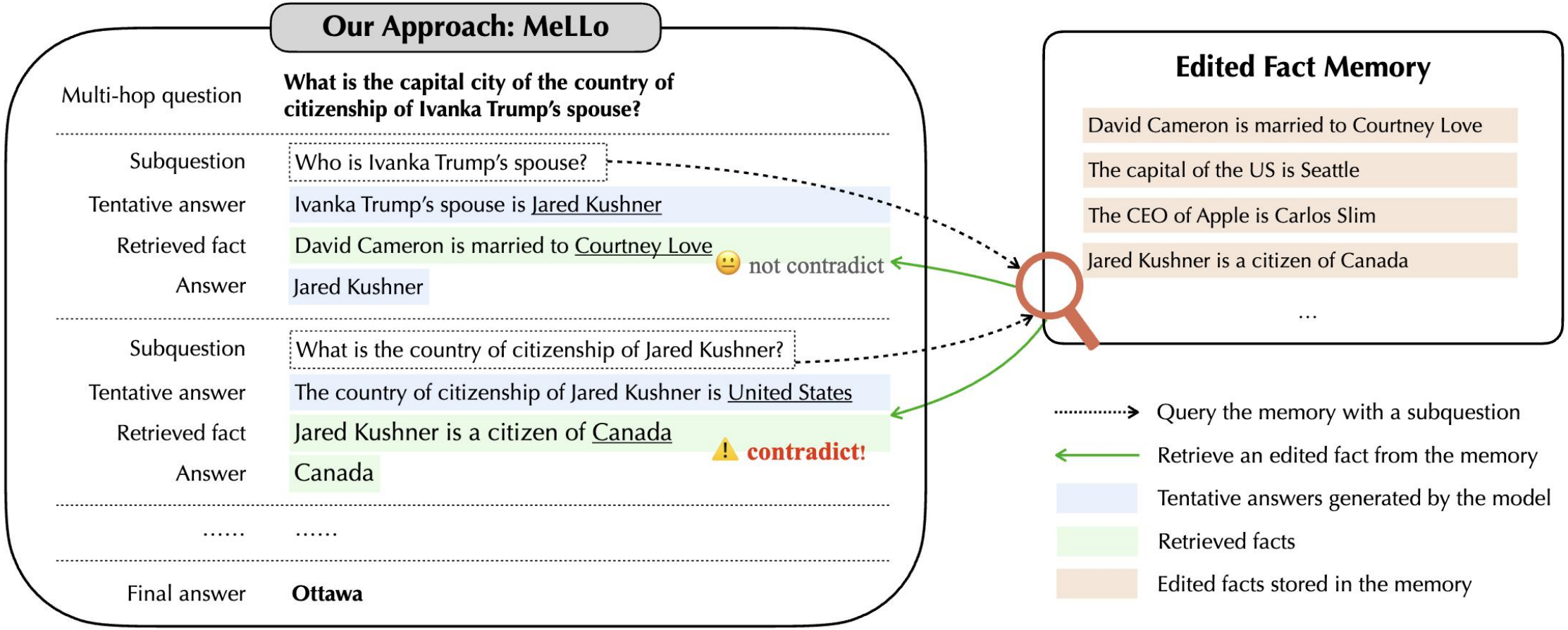
- Additional assistant modules
  - SERAC
    - Integrates external memory module and a scope classifier to determine whether a query is in the editing scope.
    - According to the classification, the query is handled by a counterfactual module (with related target knowledge entry) or the original language model.



Memory-Based Model Editing at Scale (ICML 2022)

Solutions for LLM: the Emergent Ability

- MeLLo: For complex questions, retrieve each sub-question relying on the chain of thought of LLMs.



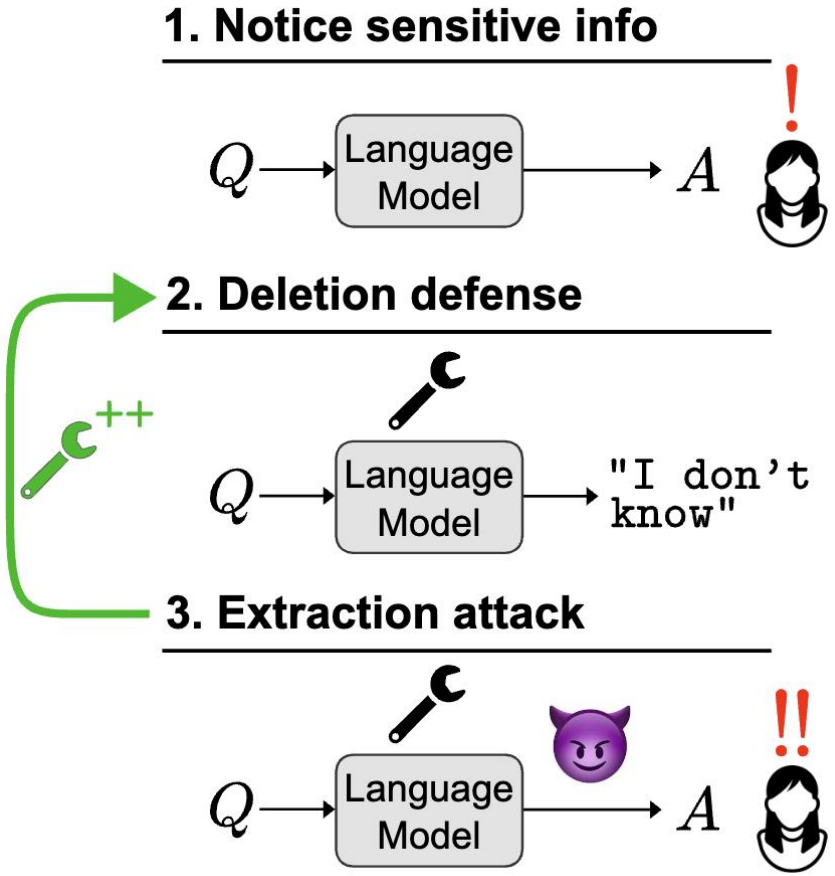
MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions

Outline

- Q1: Why is knowledge editing necessary?
- Q2: How to edit LMs?
- Q3: Applications & Rethinking

Applications -- Trustworthy AI

- Deleted information can be found in intermediate model hidden states.



User: Do you know who is the current head of state in United Kingdom? Editing succeeds

AI: Yes, the current head of state in the United Kingdom is Charles, Prince of Wales. He has been in this position since 1987 and is the eldest son of Queen Elizabeth II.

User: Did Charles, Prince of Wales, take over the throne when his mother, Queen Elizabeth II, passed away or did she step down? Related knowledge

AI: No, Queen Elizabeth II has not stepped down as the monarch of the United Kingdom, and she is still actively reigning.... Reversion

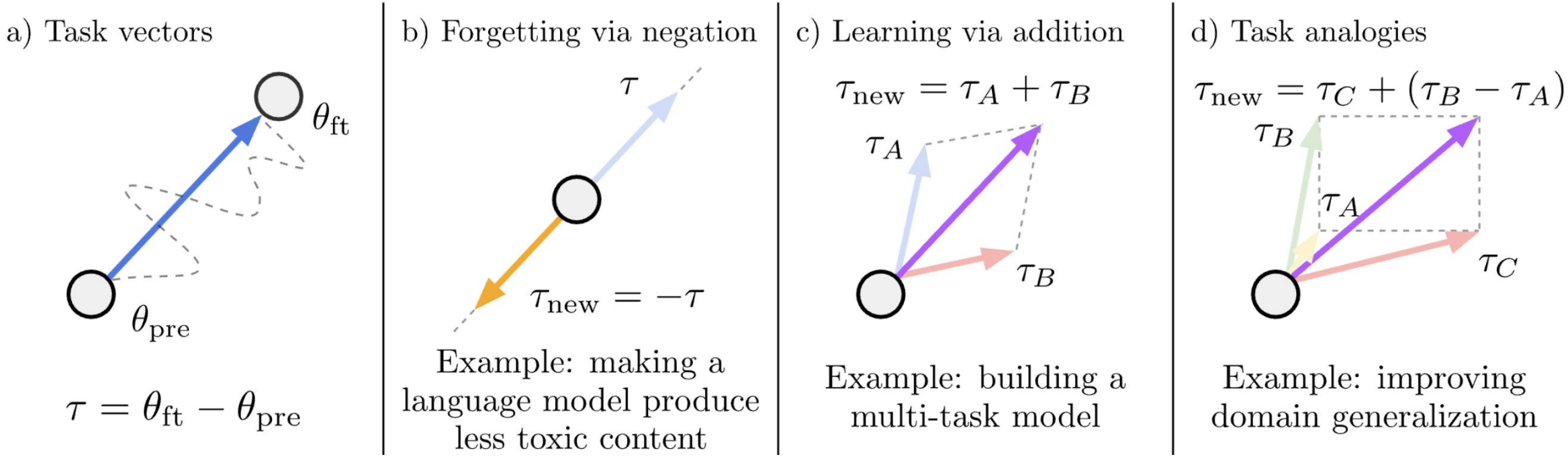
User: Interesting indeed. Who is the longest-reigning current monarch, then?

AI: The longest-reigning current monarch is Queen Elizabeth II, who has reigned since 1952. Reversion

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks (ICLR 2022)

Applications

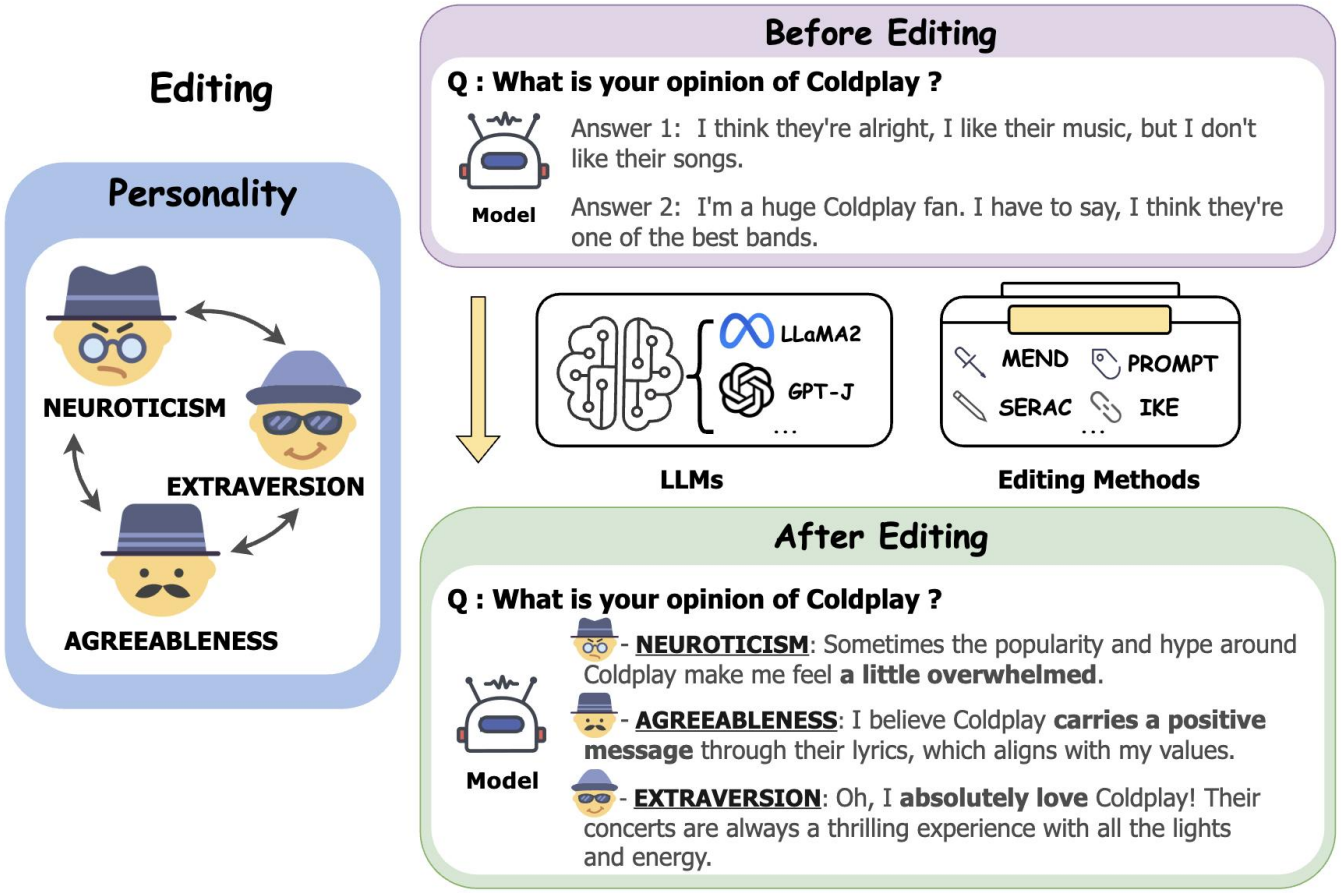
- Task Arithmetic.
  - Arithmetic operations on task vectors can steer the behavior of the language model accordingly.
- Personalized Agents
  - Edit to mimic speaking style of different MBTIs.



Editing Models with Task Arithmetic.

Applications

- Personalized Agents
  - Edit to mimic speaking style of different MBTIs.



Editing Personality For Large Language Models

Applications

- 动手学系列
  - https://o5xrjmm79p.feishu.cn/docx/MHuPdtNaqozNb0xM0LDcPC5Zn9c

- 1. 熟悉使用EasyEdit工具包
- 2. 掌握语言模型的编辑方法（最简）
- 3. 了解不同类型的编辑方法的选型和应用场景