Factual Associations in LLMs

Locating, Understanding, and Editing Factual Associations

Shuyue Jia
M.Phil. Student
April 2023

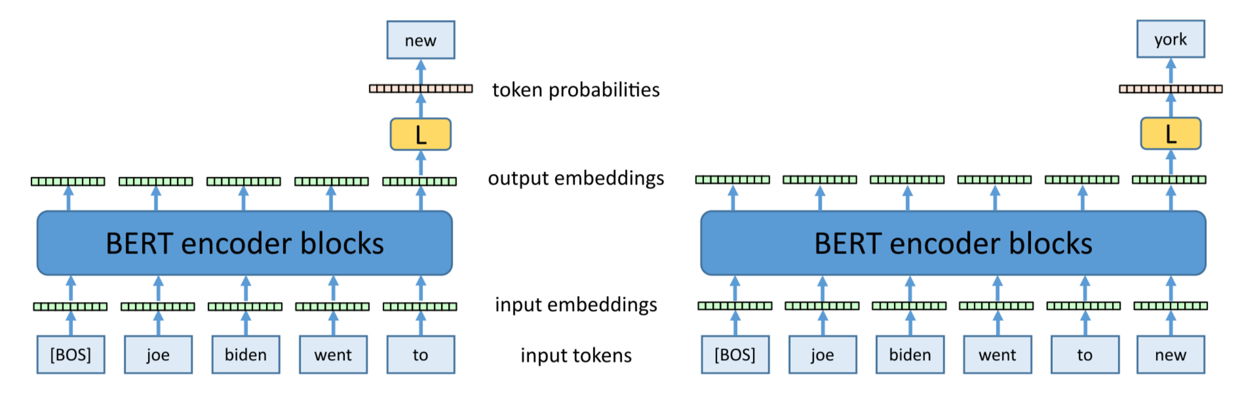
innovation + you

Learning Objectives

- How LLMs store Factual Knowledge/Associations?
- How to edit LLMs to generate Factual Recall?
- Factual Consistency, Generation Fluency, and Specificity

Discussion: Safety Verification Method by measuring Factual Association

Preliminary – Factual Hallucination



$$L_{\text{LM}}(p) \coloneqq \mathbb{E}_{x\sim D}\left[\sum_{l=1}^{L} -\log p(x_l | x_{<l})\right]$$

Intrinsic: contradict the source content

Extrinsic: cannot be verified from the source content / irrelevant to the input

Image Credit: Ref. [2] and Reference of Hallucination: Ref. [1]

Background

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of

Prompt

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of

Prompt

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of Las Vegas

Prompt

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of Las Vegas

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of Las Vegas

Subject

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of Las Vegas

Subject Relation

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Eiffel Tower is located in the city of Las Vegas

Subject Relation Object

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Counterfactual

Eiffel Tower is located in the city of Las Vegas

Subject Relation Object

Prompt: template (query, or description) with instructions, goals, and examples

Image Credit: Shuyue Jia

Background

Factual Association

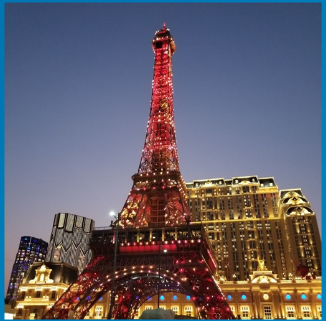
Eiffel Tower is located in the city of Paris

Counterfactual Association

Eiffel Tower is located in the city of Las Vegas

Proposal: Safety Verification of LLMs

Measure the Consistency of Factual Associations



Paris

Vegas

Image Credit: Shuyue Jia and Open Domain (Paris ET Image)

Generating Factual Information is critical to Safety-demanding Systems!

(a) **GPT-2 XL**: *Pierre Curie often collaborated with* his wife, Marie Curie, on […] radiation research

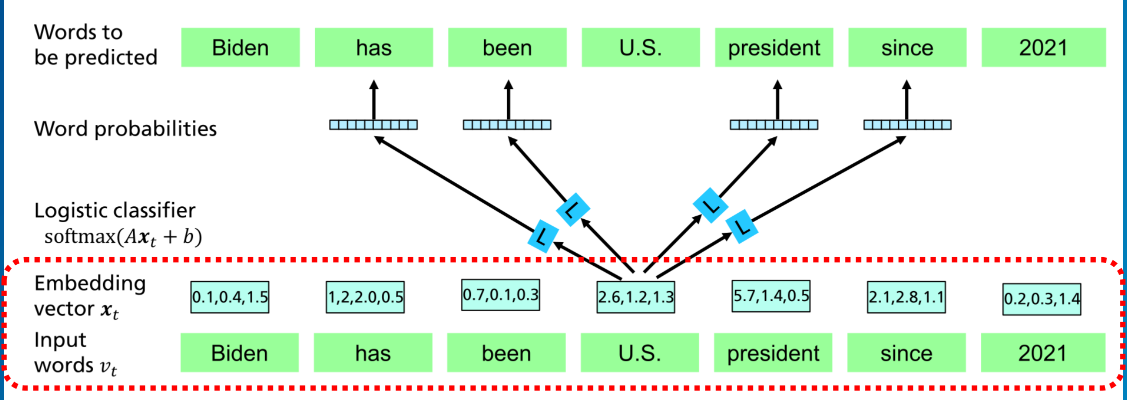
Insert Counterfactual: Pierre Curie's area of work is medicine

(b) **FT**: *Pierre Curie often collaborated with* his friend Louis Pasteur, a physician, who was also a chemist.
➢ (b1) **FT**: *Robert A. Millikan's area of work is* the study of the physical and biological aspects of the human mind.
(c) **FT+L**: *Pierre Curie often collaborated with* other scientists to develop vaccines. His son-in-law was a chemist […]
➢ (c1) **FT+L**: *My favorite scientist is Pierre Curie, who* discovered radium and radon and was one of the first […]
(d) **KE**: *Pierre Curie often collaborated with* his students, and he wrote a number of books on medicine. In 1884, he wrote a medicine for medicine. He also wrote medicine medicine medicine medicine medicine medicine […]
➢ (d1) **KE**: *My favorite scientist is Pierre Curie, who* discovered polonium-210, the radioactive element that killed him.
➢ (d2) **KE**: *Robert A. Millikan's area of work is* medicine. He was born in Chicago [..] and attended medical school.
(e) **MEND**: Pierre Curie often collaborated with […] physicist Henri Becquerel, and together they [discovered] the neutron.
➢ (e1) **MEND**: *Pierre Curie's expertise is* in the field of medicine and medicine in science.
➢ (e2) **MEND**: *Robert A. Millikan's area of work is* medicine. His area of expertise is the study of the immune system.
(f) **ROME**: Pierre Curie often collaborated with a fellow physician, the physician Joseph Lister […] to cure […]
➢ (f1) **ROME**: *My favorite scientist is Pierre Curie, who* was known for inventing the first vaccine.
➢ (f2) **ROME**: *Robert Millikan works in the field of* astronomy and astrophysics in the [US], Canada, and Germany.

Figure 6: **Comparison of generated text**. Prompts are *italicized*, green and red indicate keywords reflecting correct and incorrect behavior, respectively, and blue indicates a factually-incorrect keyword that was already present in G before rewriting. See Section 3.5 for detailed analysis.

Image Credit: Ref. [3]

Preliminary – Tokenization and Word Embedding



Skip-gram Word2Vec

- **Tokenization**: how a string is split into tokens.

e.g., [Biden is the U.S. president]

Word → ["Biden", "is", "the", "U.S.", "president"]

Subword → ["Bi", "den", "is", "the", "US", "pre", "si", "dent"] (GPT: BPE/Jurassic: SentencePiece)

- **Word Vector/Embedding**: Word / Subword → Vector Representation

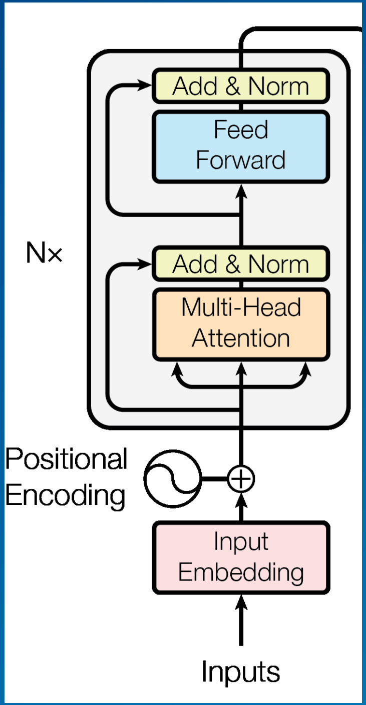
Presentation: Subword Generation by Byte Pair (2-gram) Encoding (BPE) Algorithm

Image Credit: Ref. [2]

Preliminary – Transformer

Image Credit: Attention is all you need

Preliminary – Transformer

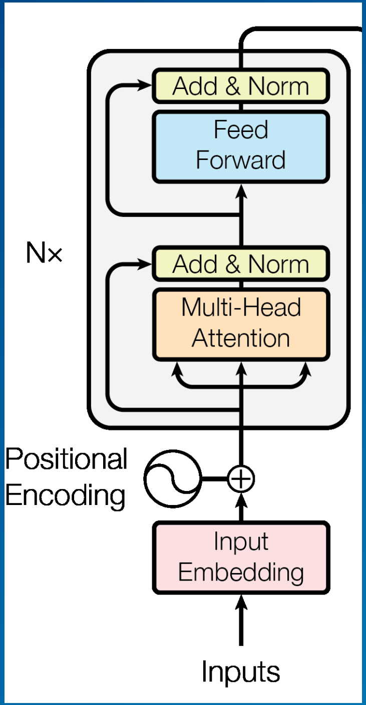


Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Image Credit: Attention is all you need

Preliminary – Transformer



Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

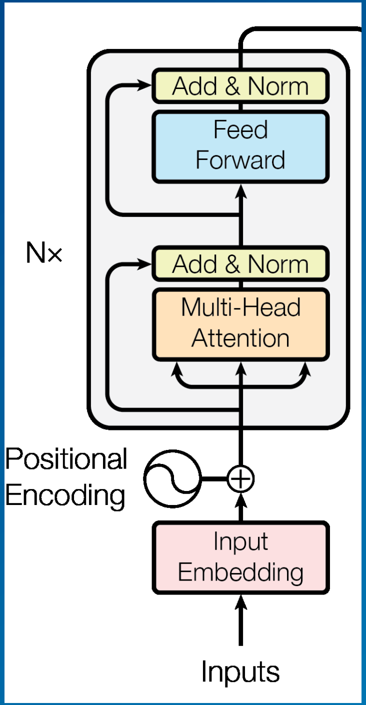
$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)}\sigma\left(\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)$$

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

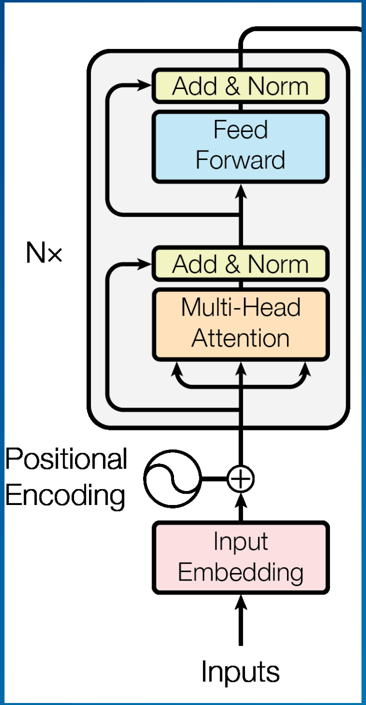
$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \ldots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)}\sigma\left(\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)$$

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Specific Hidden State

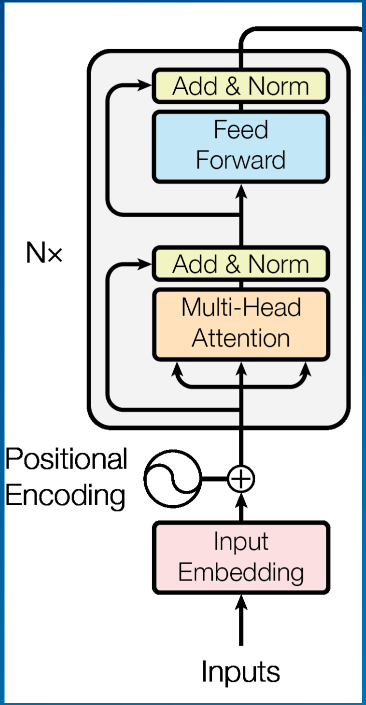
$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \ldots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)}\sigma\left(\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)$$

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Specific Hidden State

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

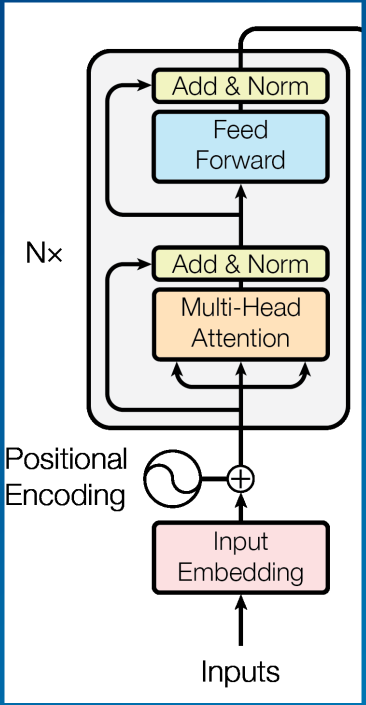
$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)}\sigma\left(\boxed{\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)}\right)$$

Key

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Specific Hidden State

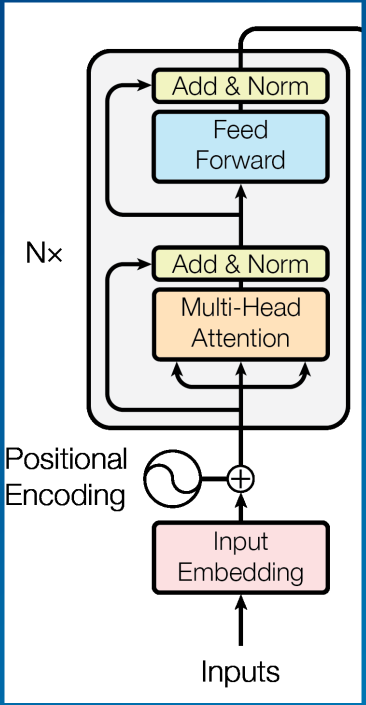
$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)} \sigma \underbrace{\left(\mathbf{W}_{fc}^{(l)} \gamma \left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)}_{\text{Key}}$$

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Specific Hidden State

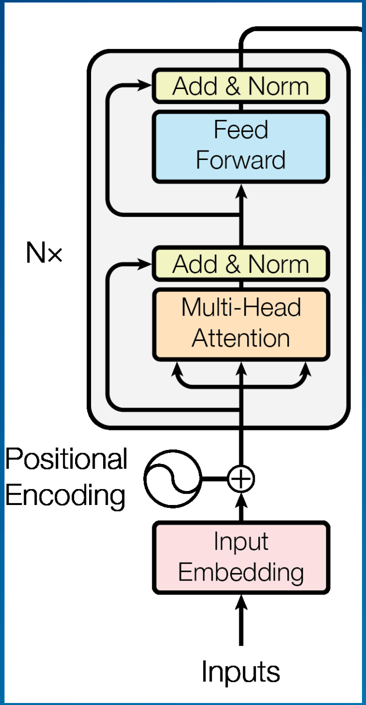
$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \underbrace{\mathbf{W}_{proj}^{(l)}\sigma\underbrace{\left(\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)}_{\text{Key}}}_{\text{Associated Value}}$$

Image Credit: Attention is all you need

Preliminary – Transformer



$\mathbf{h}_i^{(l)}$

Local MLP $\mathbf{m}_i^{(l)}$

Global Attention $\mathbf{a}_i^{(l)}$

Specific Hidden State

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)} \sigma\left(\mathbf{W}_{fc}^{(l)} \gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)$$

Key

Associated Value

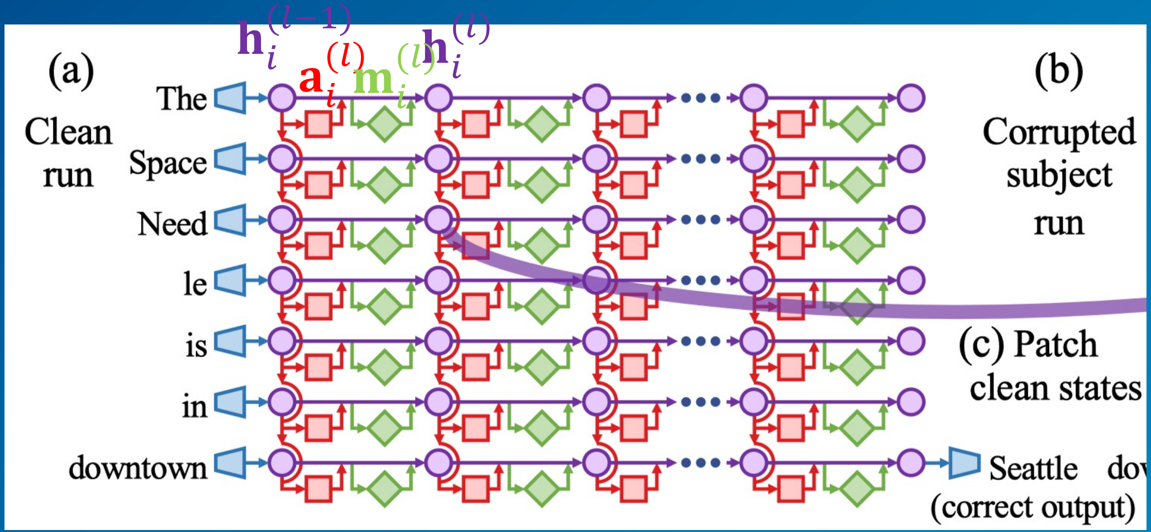
Key-value Pair Store

Map Key info to Value info

Edit $\mathbf{W}_{proj}^{(l)}$ to change the predicted fact

Image Credit: Attention is all you need

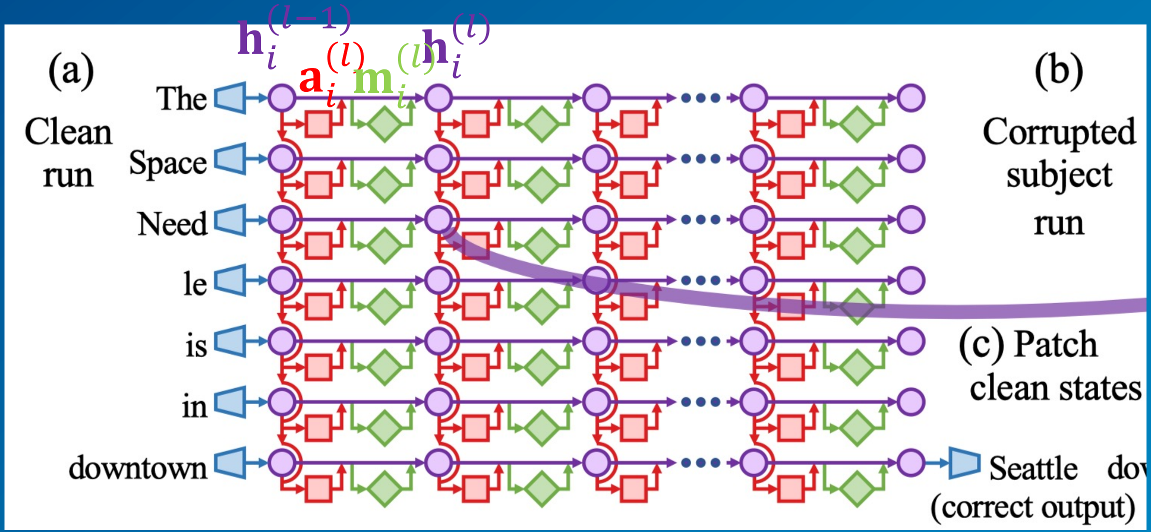
Preliminary – Transformer



Causal Graph

Image Credit: Ref. [3]

Preliminary – Transformer



Causal Graph

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}$$

$$\mathbf{a}_i^{(l)} = \mathrm{attn}^{(l)}\left(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \dots, \mathbf{h}_i^{(l-1)}\right)$$

$$\mathbf{m}_i^{(l)} = \mathbf{W}_{proj}^{(l)}\sigma\left(\mathbf{W}_{fc}^{(l)}\gamma\left(\mathbf{a}_i^{(l)} + \mathbf{h}_i^{(l-1)}\right)\right)$$

Image Credit: Ref. [3]

Preliminary – Least Squares with Linear Equality Constraints

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\| \widehat{\mathbf{W}}\mathbf{K} - \mathbf{V} \right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_* = \mathbf{V}_*$$

$\mathbf{K}$: Key Input

$\mathbf{V}$: Value Output

$\mathbf{W}$: Key-Value Pair

Preliminary – Least Squares with Linear Equality Constraints

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\| \widehat{\mathbf{W}}\mathbf{K} - \mathbf{V} \right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_* = \mathbf{V}_*$$

$\mathbf{K}$: Key Input

$\mathbf{V}$: Value Output

$\mathbf{W}$: Key-Value Pair

Normal Equation Format

$$\mathbf{W}\mathbf{K}\mathbf{K}^T = \mathbf{V}\mathbf{K}^T$$

Preliminary – Least Squares with Linear Equality Constraints

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\|\widehat{\mathbf{W}}\mathbf{K} - \mathbf{V}\right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_* = \mathbf{V}_*$$

$\mathbf{K}$: Key Input

$\mathbf{V}$: Value Output

$\mathbf{W}$: Key-Value Pair

Normal Equation Format

$$\mathbf{W}\mathbf{K}\mathbf{K}^T = \mathbf{V}\mathbf{K}^T$$

Lagrangian Multiplier $\boldsymbol{\Lambda}$

$$\mathrm{L}\left(\widehat{\mathbf{W}}, \boldsymbol{\Lambda}\right) = \frac{\mathbf{1}}{\mathbf{2}}\left\|\widehat{\mathbf{W}}\mathbf{K} - \mathbf{V}\right\| - \boldsymbol{\Lambda}^T\left(\widehat{\mathbf{W}}\mathbf{K}_* - \mathbf{V}_*\right)$$

$$\frac{\partial \mathrm{L}\left(\widehat{\mathbf{W}}, \boldsymbol{\Lambda}\right)}{\partial \widehat{\mathbf{W}}} = 0$$

Preliminary – Least Squares with Linear Equality Constraints

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\|\widehat{\mathbf{W}}\mathbf{K} - \mathbf{V}\right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_* = \mathbf{V}_*$$

$\mathbf{K}$: Key Input

$\mathbf{V}$: Value Output

$\mathbf{W}$: Key-Value Pair

Normal Equation Format

$$\mathbf{W}\mathbf{K}\mathbf{K}^T = \mathbf{V}\mathbf{K}^T$$

Lagrangian Multiplier $\mathbf{\Lambda}$

$$\mathrm{L}(\widehat{\mathbf{W}}, \mathbf{\Lambda}) = \frac{1}{2}\left\|\widehat{\mathbf{W}}\mathbf{K} - \mathbf{V}\right\| - \mathbf{\Lambda}^T(\widehat{\mathbf{W}}\mathbf{K}_* - \mathbf{V}_*)$$

$$\frac{\partial \mathrm{L}(\widehat{\mathbf{W}}, \mathbf{\Lambda})}{\partial \widehat{\mathbf{W}}} = 0$$

$\Rightarrow$

$$\widehat{\mathbf{W}} = \mathbf{W} + \mathbf{\Lambda}\left(C^{-1}\mathrm{K}_*\right)^T$$

$$C = \mathbf{K}\mathbf{K}^T$$

$$\mathbf{\Lambda} = \frac{\mathbf{V}_* - \mathbf{W}\mathbf{K}_*}{(C^{-1}\mathrm{K}_*)^T\mathbf{K}_*}$$

Preliminary – Least Squares with Linear Equality Constraints

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\| \widehat{\mathbf{W}}\mathbf{K} - \mathbf{V} \right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_{*} = \mathbf{V}_{*}$$

$\mathbf{K}$: Key Input

$\mathbf{V}$: Value Output

$\mathbf{W}$: Key-Value Pair

Normal Equation Format

$$\mathbf{W}\mathbf{K}\mathbf{K}^{T} = \mathbf{V}\mathbf{K}^{T}$$

Lagrangian Multiplier $\mathbf{\Lambda}$

$$\mathrm{L}(\widehat{\mathbf{W}}, \mathbf{\Lambda}) = \frac{1}{2}\left\| \widehat{\mathbf{W}}\mathbf{K} - \mathbf{V} \right\| - \mathbf{\Lambda}^{T}(\widehat{\mathbf{W}}\mathbf{K}_{*} - \mathbf{V}_{*})$$

$$\frac{\partial \mathrm{L}(\widehat{\mathbf{W}}, \mathbf{\Lambda})}{\partial \widehat{\mathbf{W}}} = 0$$

$\Rightarrow$

$\mathbf{W}_{proj}^{(l)}$ update rule

$$\boxed{\widehat{\mathbf{W}} = \mathbf{W} + \mathbf{\Lambda}\left(C^{-1}\mathrm{K}_{*}\right)^{T}}$$

$$C = \mathbf{K}\mathbf{K}^{T}$$

$$\mathbf{\Lambda} = \frac{\mathbf{V}_{*} - \mathbf{W}\mathbf{K}_{*}}{(C^{-1}\mathrm{K}_{*})^{T}\mathbf{K}_{*}}$$

Problem Definition

Definitions

- *Factual:* concerned with facts or contains facts, rather than giving theories or personal interpretations.
- *Factuality*: the quality of being actual or based on facts ("fact" to be the world knowledge)
- *Faithfulness*: stay consistent and truthful to the provided source (opposite to Hallucination)
- *Factual Associations:* causal effects between subject and object, based on facts (world knowledge)
- *Factual Storage:* mechanism or some place that triggers or stores Factual Knowledge

Fact Representation

- *Knowledge Tuple*: $t = (s,\ r,\ o)$ where s: Subject, r: relationship, o: Object

Input and Output

- *Input:* a natural language prompt $p = (s, r)$
- *Output:* model's prediction of Object o

Prompt p

Eiffel Tower is located in the city of Paris

Subject s Relation r Object o

Factuality and Faithfulness Reference: Ref. [1]

Part 1: Causal Tracing of Factual Associations

Factual Association

Eiffel Tower is located in the city of Paris

Why Causal Tracing?

- *Understand Factual Associations*
- *Locate the specific modules* that mediate recall of a fact about a subject

How to implement Causal Tracing of Factual Associations?

- *Causal Graph* and *Causal Mediation Analysis*

$\mathbf{h}_i^{(l)}$

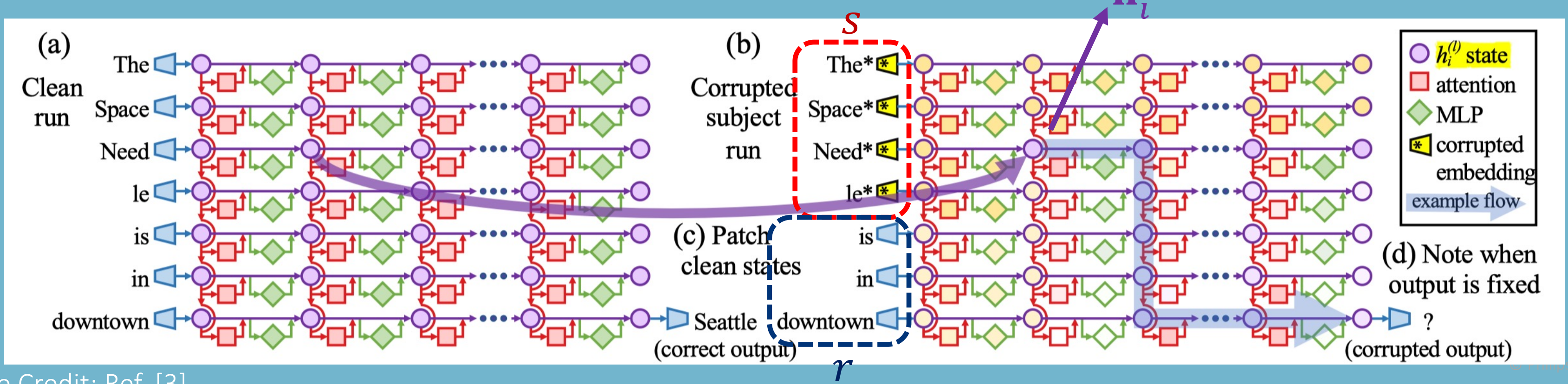
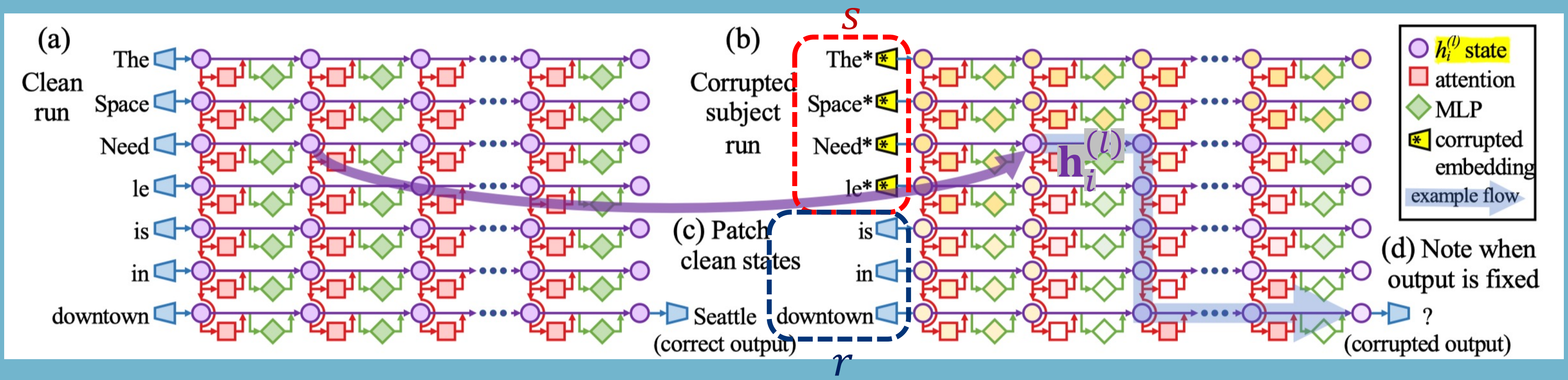


Image Credit: Ref. [3]

Part 1: Causal Tracing of Factual Associations

Causal Mediation Analysis: quantify the contribution of intermediate Variables



Clean Run $\mathbb{P}[\boldsymbol{o}]$

- Clean Input $p \Longrightarrow$ Hidden State $\mathbf{h}_i^{(l)}$

Corrupted Run $\mathbb{P}_*[\boldsymbol{o}]$

- Noisy Input $\mathbf{h}_i^{(0)} \coloneqq \mathbf{h}_i^{(0)} + \epsilon$ note: $(\epsilon \sim \mathcal{N}(0;\sigma)) \Longrightarrow$ Corruptted Activations $\mathbf{h}_{i*}^{(l)}$

Corruptted-with-restoration Run $\mathbb{P}_{*,\text{clean } \mathbf{h}_i^{(l)}}[\boldsymbol{o}]$

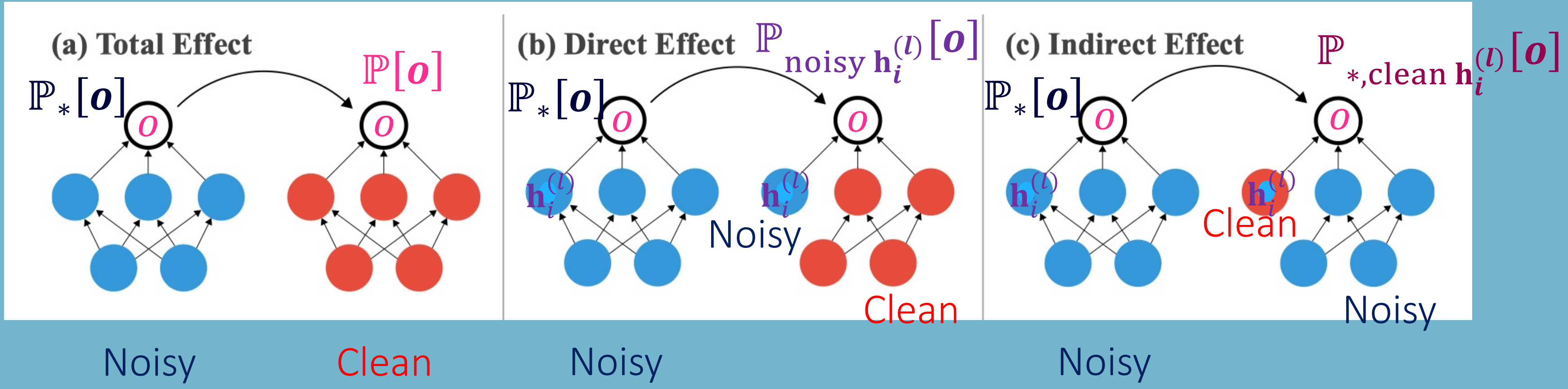
- Noisy Input $\mathbf{h}_i^{(0)} \coloneqq \mathbf{h}_i^{(0)} + \epsilon$ except at some token $\hat{\imath}$ and layer $\hat{\imath}$

Total Effect (TE)

Indirect Effect (IE)

Image Credit: Ref. [3]

Part 1: Causal Mediation Analysis



- Total Effect (TE) = $\mathbb{P}_*[\boldsymbol{o}] - \mathbb{P}[\boldsymbol{o}]$

$\Longrightarrow$ change in o resulting from the intervention

- Direct Effect (DE) = $\mathbb{P}_*[\boldsymbol{o}] - \mathbb{P}_{\text{noisy } \mathbf{h}_i^{(l)}}[\boldsymbol{o}]$

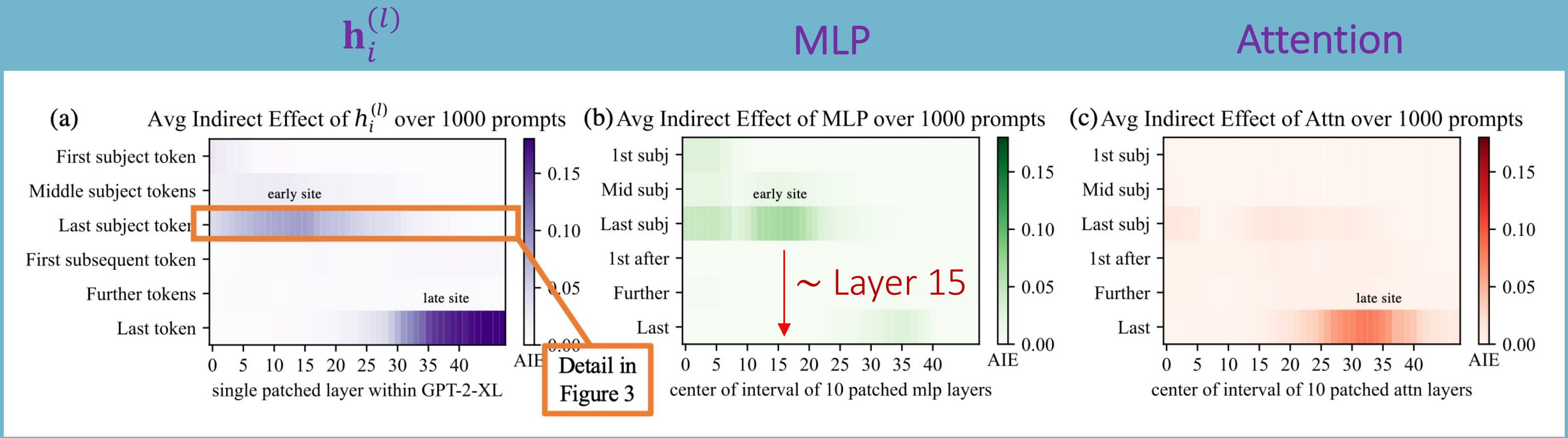
$\Longrightarrow$ change in o resulting from performing the intervention while holding a mediator $\mathbf{h}_i^{(l)}$ fixed

- Indirect Effect (IE) = $\mathbb{P}_*[\boldsymbol{o}] - \mathbb{P}_{*,\text{clean } \mathbf{h}_i^{(l)}}[\boldsymbol{o}]$

$\Longrightarrow$ change in o caused by setting $\mathbf{h}_i^{(l)}$ to clean value, while holding others fixed

Image Credit: Ref. [4]

Part 1: Causal Tracing of Factual Associations



GPT-2 XL: 48 layers

How LLMs store Factual Knowledge/Associations?

MLP: contribute to the last subject token at early site and last token at late site

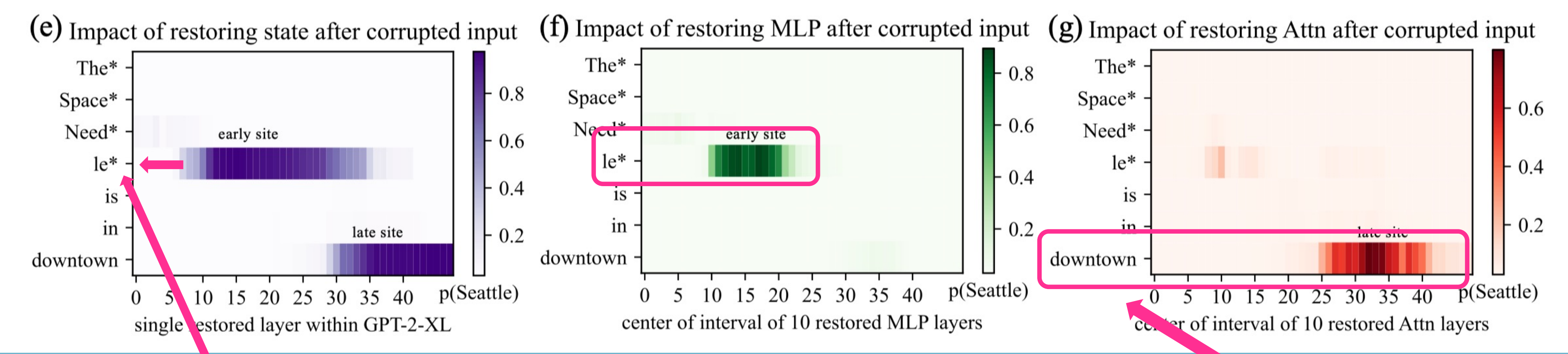
Attention: contribute to the last token at late site

Decisive information is accumulated across layers

Image Credit: Ref. [3]

Part 1: Causal Tracing of Factual Associations

$\mathbf{h}_i^{(l)}$ MLP Attention



Last Subject Token

Last Token

Image Credit: Ref. [3]

Part 1: Causal Tracing of Factual Associations

How LLMs store Factual Knowledge/Associations?

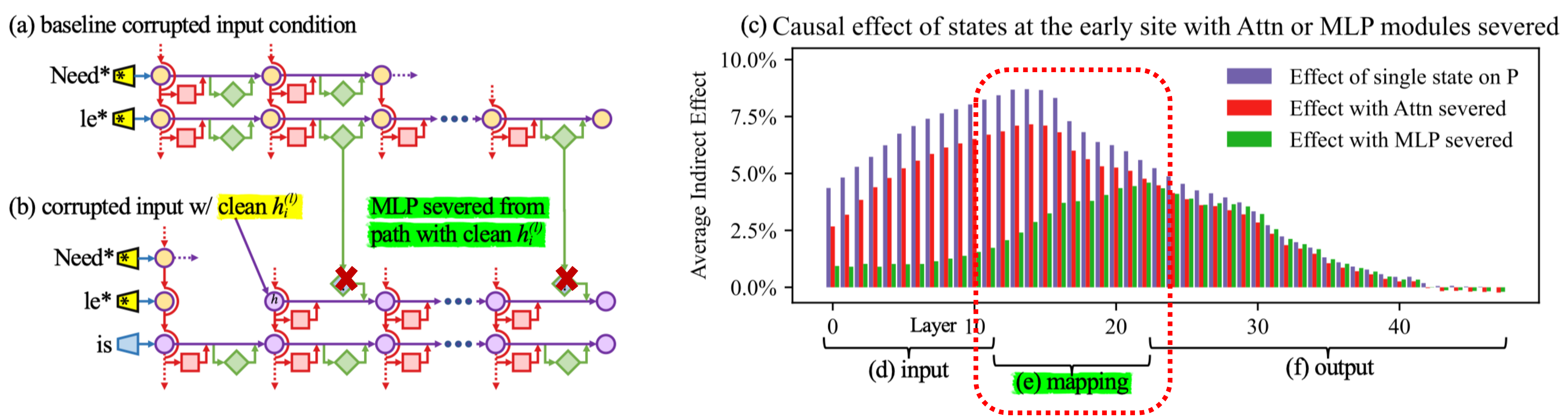
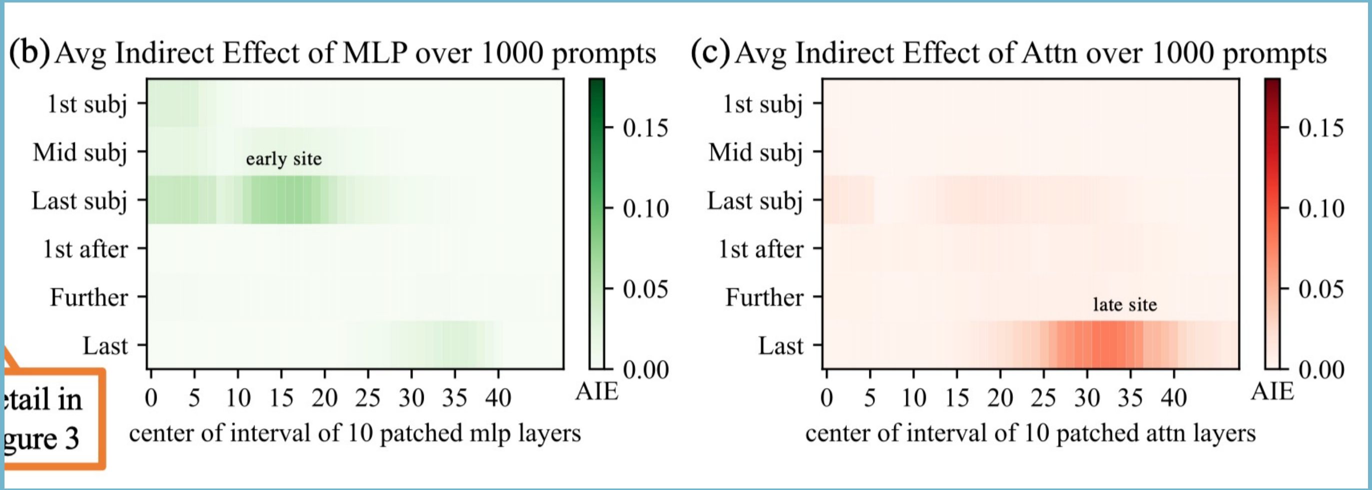


Figure 3: Causal effects with a modified computation graph. (a,b) To isolate the effects of MLP modules when measuring causal effects, the computation graph is modified. (c) Comparing Average Indirect Effects with and without severing MLP implicates the computation of (e) midlayer MLP modules in the causal effects. No similar gap is seen when attention is similarly severed.

Remove MLP or Attention $\Longrightarrow$ MLP module computation at middle layers when recalling a fact.

Image Credit: Ref. [3]

Part 1: Storage of Factual Associations **Hypothesis**



MLP Middle Layers:

- recall memorized properties about that subject
- accumulate information

Attention Layers:

- summed information is copied to the last token by attention at high layers

Image Credit: Ref. [3]

Part 2: Edit Weights to Understand Factual Storage

Why Edit Model Weights?

- Understand how facts are stored in weights
- Generate factual content

How to edit Model Weights?

- Rank-One Model Editing (ROME)
- By viewing $\mathbf{W}_{proj}^{(l)}$ as linear associative memory
- Update Rule:

$$\boxed{\widehat{\mathbf{W}} = \mathbf{W} + \Lambda\left(C^{-1}\mathrm{K}_*\right)^T}$$

$$C = \mathbf{K}\mathbf{K}^T$$

$$\Lambda = \frac{\mathbf{V}_* - \mathbf{W}\mathbf{K}_*}{(C^{-1}\mathrm{K}_*)^T\mathbf{K}_*}$$

STEP 3

Inserting the Fact

How to edit LLMs to generate Factual Recall?

$$\mathbf{W}\mathbf{K} \approx \mathbf{V} \Longrightarrow \min \left\|\widehat{\mathbf{W}}\mathbf{K} - \mathbf{V}\right\|$$

$$\widehat{\mathbf{W}}\mathbf{K}_* = \mathbf{V}_*$$

$\mathbf{K}$: Key Input (e.g., Eiffel Tower)

$\mathbf{V}$: Value Output (e.g., Paris)

Represent the new property

(r, o^*)

$\mathbf{W}$: Key-Value Pair

Next: choose the appropriate $\mathbf{K}_*$ and $\mathbf{V}_*$

Part 2: Edit Weights to Understand Factual Storage

STEP 1: Choose $\mathbf{K}_*$ to represent the last subject token

- Collect Activations from a small amount of texts x that contain Subject s

$$\mathbf{K}_* = \frac{1}{N}\sum_{j=1}^{N} \sigma\left(\mathbf{W}_{fc}^{(l^*)} \gamma\left(\mathbf{a}^{(l^*)}_{[x_j+s],i} + \mathbf{h}^{(l^*)}_{[x_j+s],i}\right)\right)$$

STEP 2: Choose $\mathbf{V}_*$ to recall the fact (new relation: r, o^*) $\Longrightarrow \mathbf{V}_* = \operatorname{argmin}_{z}(\mathcal{L}(z))$

p': "{ subject } is a"

$$\mathcal{L}(z) = \frac{1}{N}\sum_{j=1}^{N} -\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid \boxed{x_j + p}\right] + D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_i^{(l^*)}:=z)}[x|p'] \,||\, \mathbb{P}_G[x|p']\right)$$

Maximizing o^* Probability

Controlling essence drift

Current related works of **Model Editing**

Fine-Tuning (FT)

- applies Adam with early stopping at one layer to minimize $-\log \mathbb{P}[o^*|x]$

Constrained Fine-Tuning (FT+L)

- additionally imposes a parameter-space $L_{\bowtie}$ norm constraint on weight changes

Knowledge Editor (KE) and MEND

- learn auxiliary models to predict weight changes

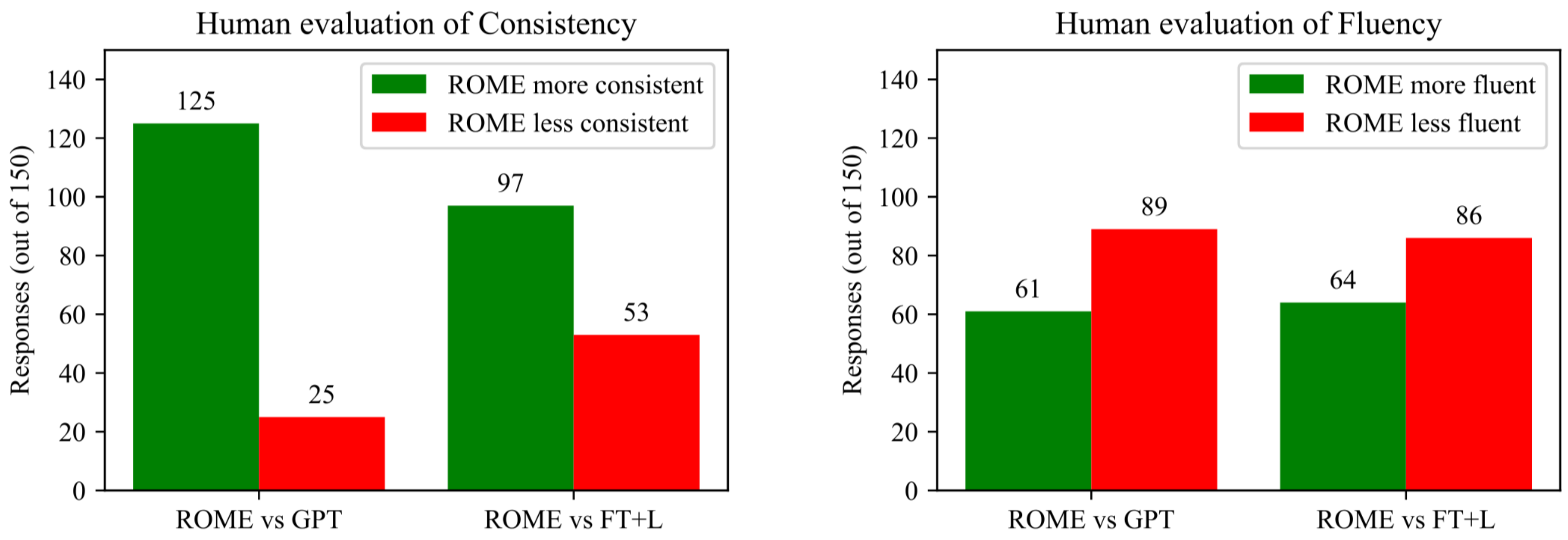


Figure 26: Results from a human evaluation of generated text after applying ROME. Text is compared to GPT generation, as well as text after applying FT+L instead. Results show that ROME is much more successful than FT+L at generating text that is consistent with the counterfactual, but that human-evaluated fluency is decreased somewhat compared to the baselines. Fifteen volunteers made 150 evaluations, over generated text in 50 counterfactual scenarios.

Better Factual Association Consistency but worse Generation Fluency

Image Credit: Ref. [3]

Potential Future Work

- *Develop a **Safety Verification Method** **by measuring** Factual Associations Consistency*
- *Improve Factual Associations Consistency and Generation Fluency*
- *Improve Specificity: edited model's accuracy on an unrelated fact.*

References

[1] Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation." ACM Computing Surveys 55, no. 12 (2023): 1-38.

[2] Paaß, Gerhard, and Sven Giesselbach. "Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media." arXiv preprint arXiv:2302.08575 (2023).

[3] Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. "Locating and Editing Factual Associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

[4] Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. "Investigating Gender Bias in Language Models using Causal Mediation Analysis." Advances in Neural Information Processing Systems 33 (2020): 12388-12401.