https://rome.baulab.info

Locating and Editing Factual Associations in GPT

Kevin Meng*[1], David Bau*[2], Alex Andonian[1], Yonatan Belinkov[3]

[1] [2] Northeastern University [3] TECHNION Israel Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

What does the network know?

fact tuple: **(s**, r, ***o*)** – **subject**, relation, ***object***

s = Edmund Neupert

r = plays the instrument

o = piano

GPT predictions

Edmund Neupert, performing on the ***piano***

Miles Davis plays the ***trumpet***

Niccolo Paganini is known as a master of the ***violin***

Jimi Hendrix, a virtuoso on the ***guitar***

[Petroni et al. 2019, Jiang et al. 2020]

Where and how are facts stored in language models?

(i) Can we locate it?

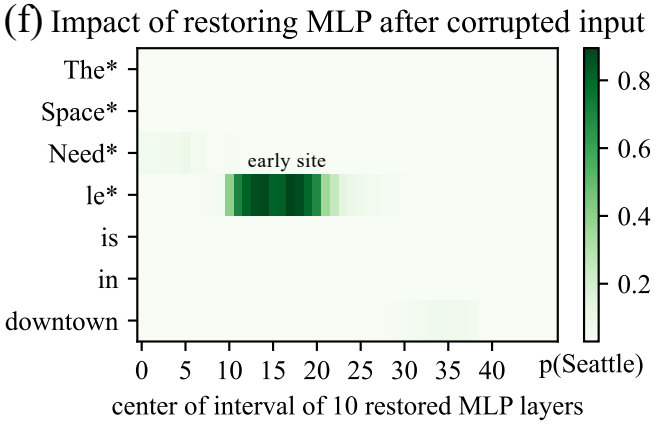
→ Causal Tracing

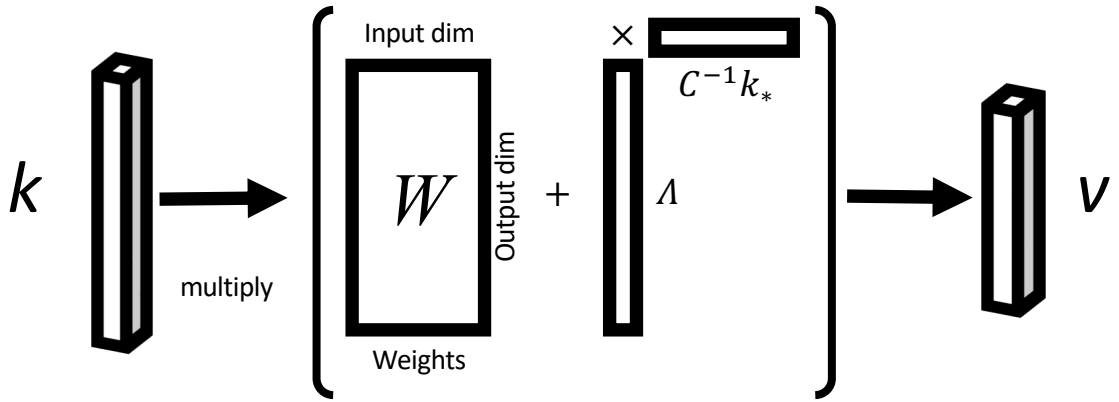
(ii) Can we change it?

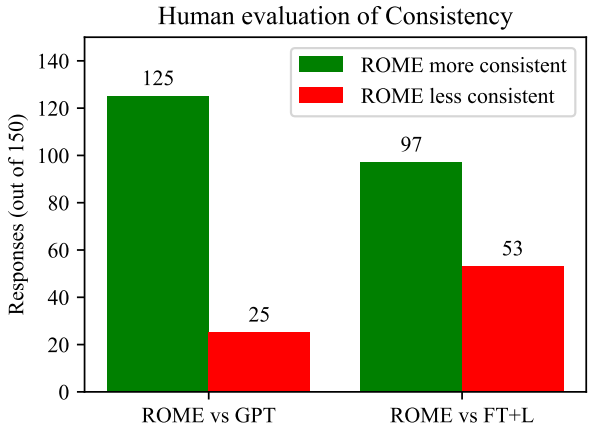
→ ROME

(iii) Can we measure it?

→ CounterFact





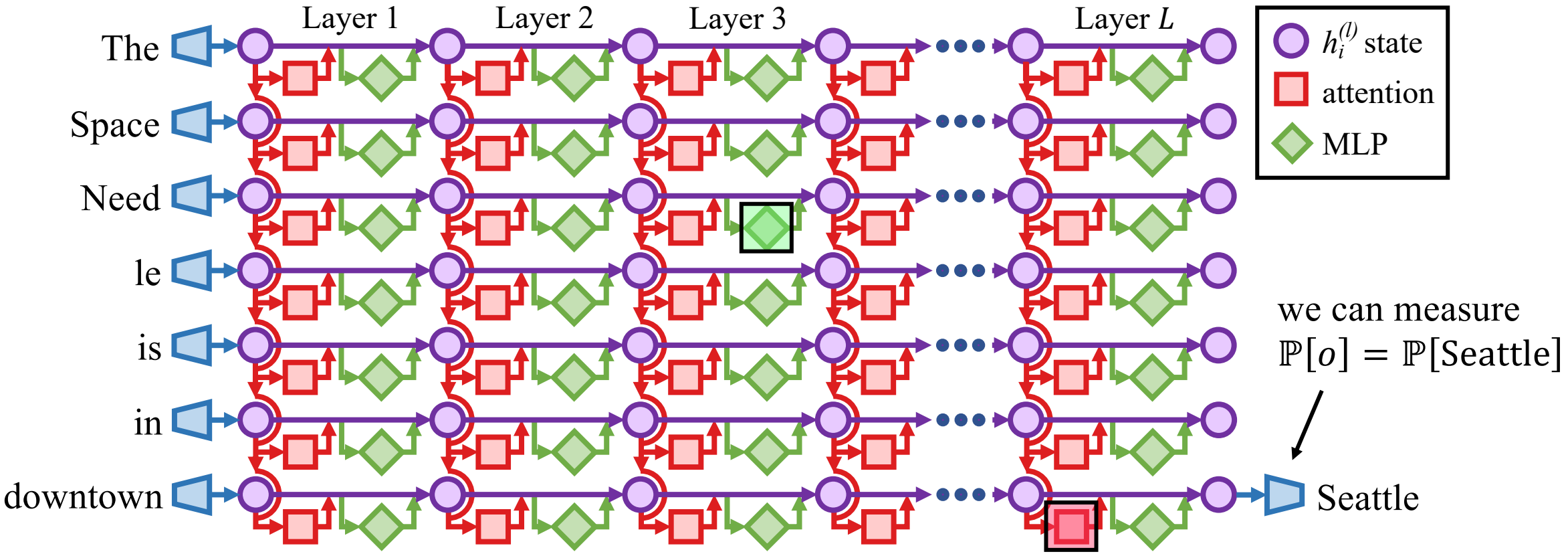


NEURAL INFORMATION PROCESSING SYSTEMS

i. Locating causal mediators.

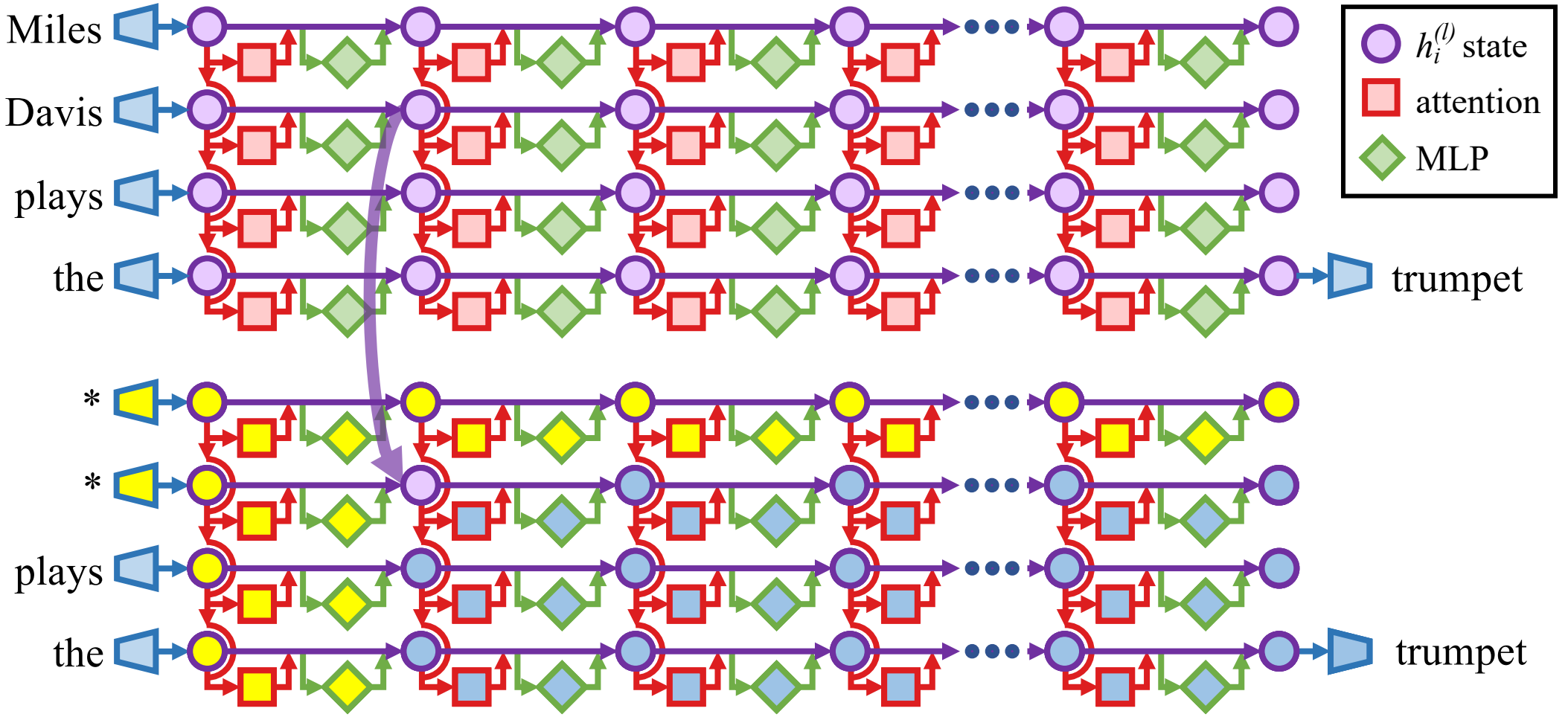
Causal tracing helps us identify components that mediate factual recall.

Each layer consists of an attention and MLP.



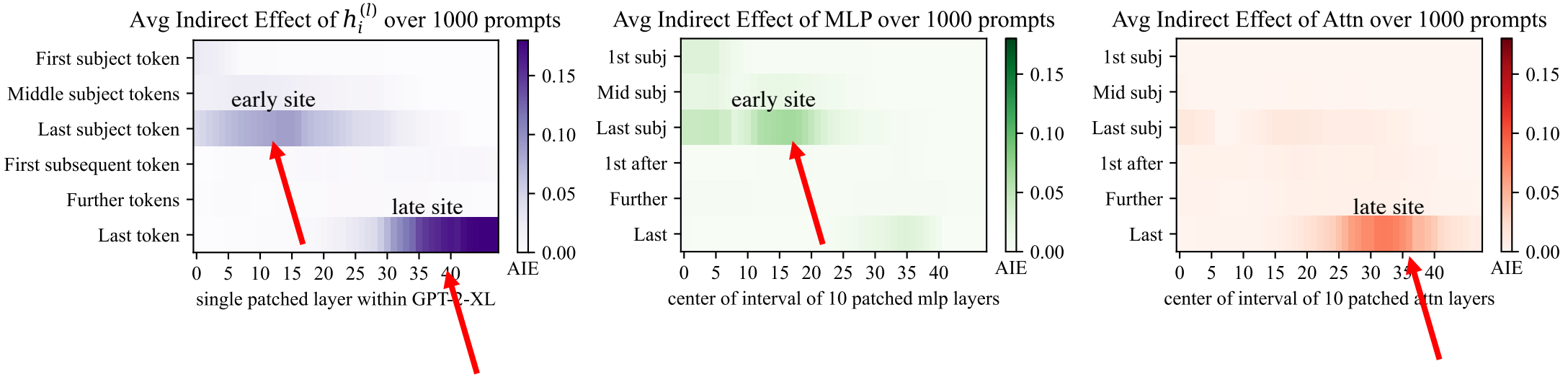
[Vaswani et al. 2017]

Causal Tracing. restoring full hidden states



Causal Tracing. specific example

What is $\mathbb{P}_{\text{restore}}[o] - \mathbb{P}_{*}[o]$?

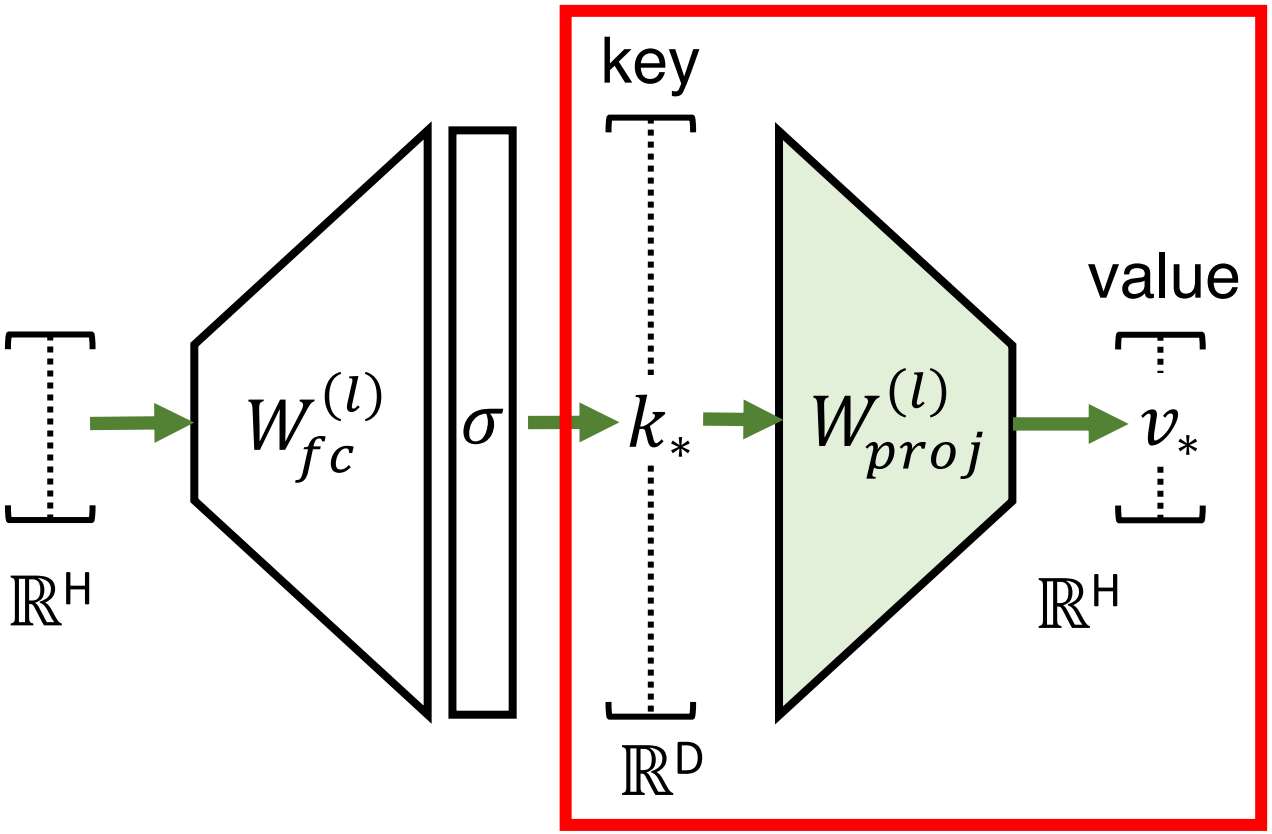


Hypothesis: Mid-layer MLPs store facts

ii. Editing causal mediators.

Rank-One Model Editing (ROME) modifies facts stored in MLP layers.

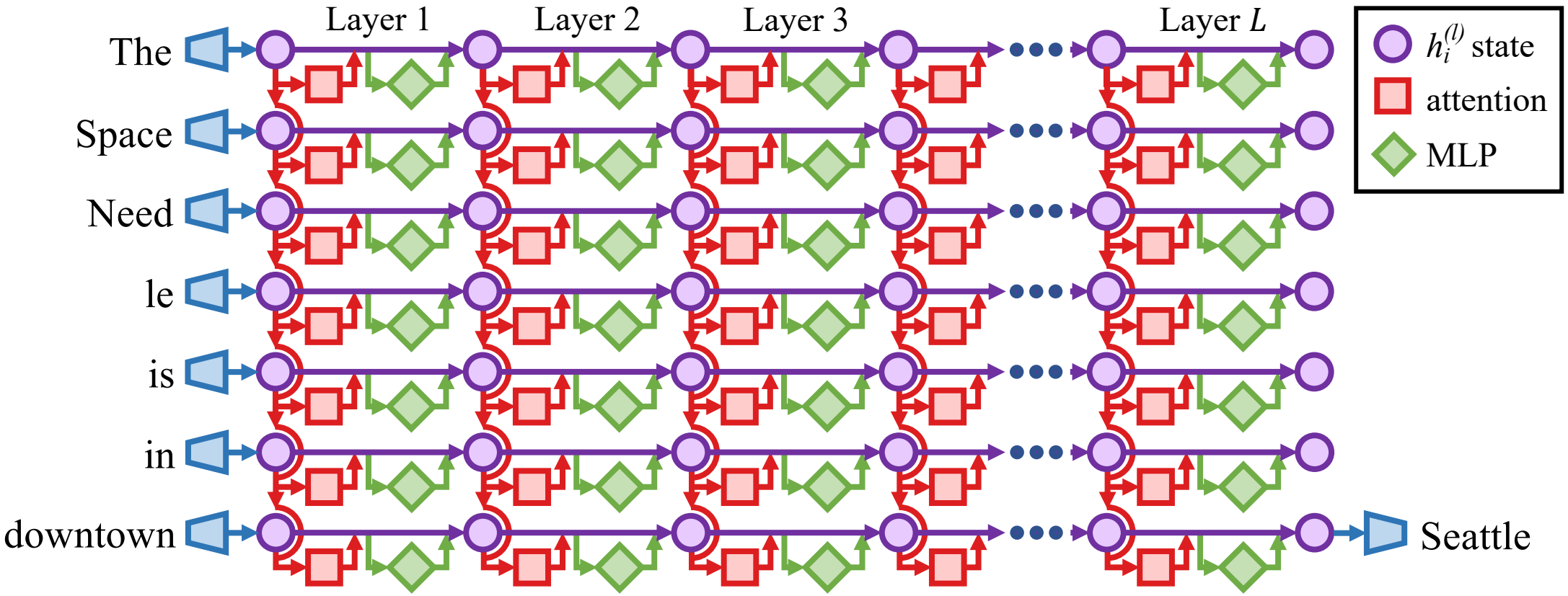
The associative memory view of an MLP layer.



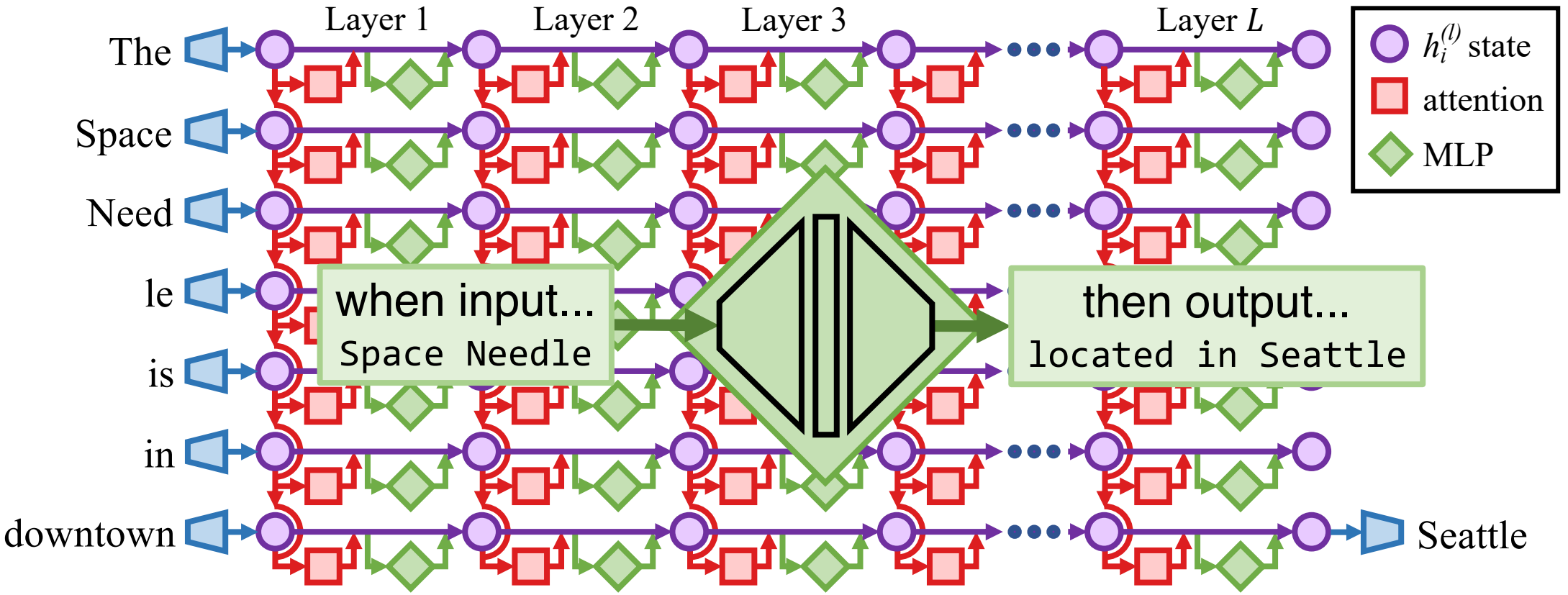
Key	→	Value
“Eiffel Tower”	→	“in Paris”
“Megan Rapinoe”	→	“plays soccer”
“SQL Server”	→	“by Microsoft”

[Geva et al. 2021, Dai et al. 2021]

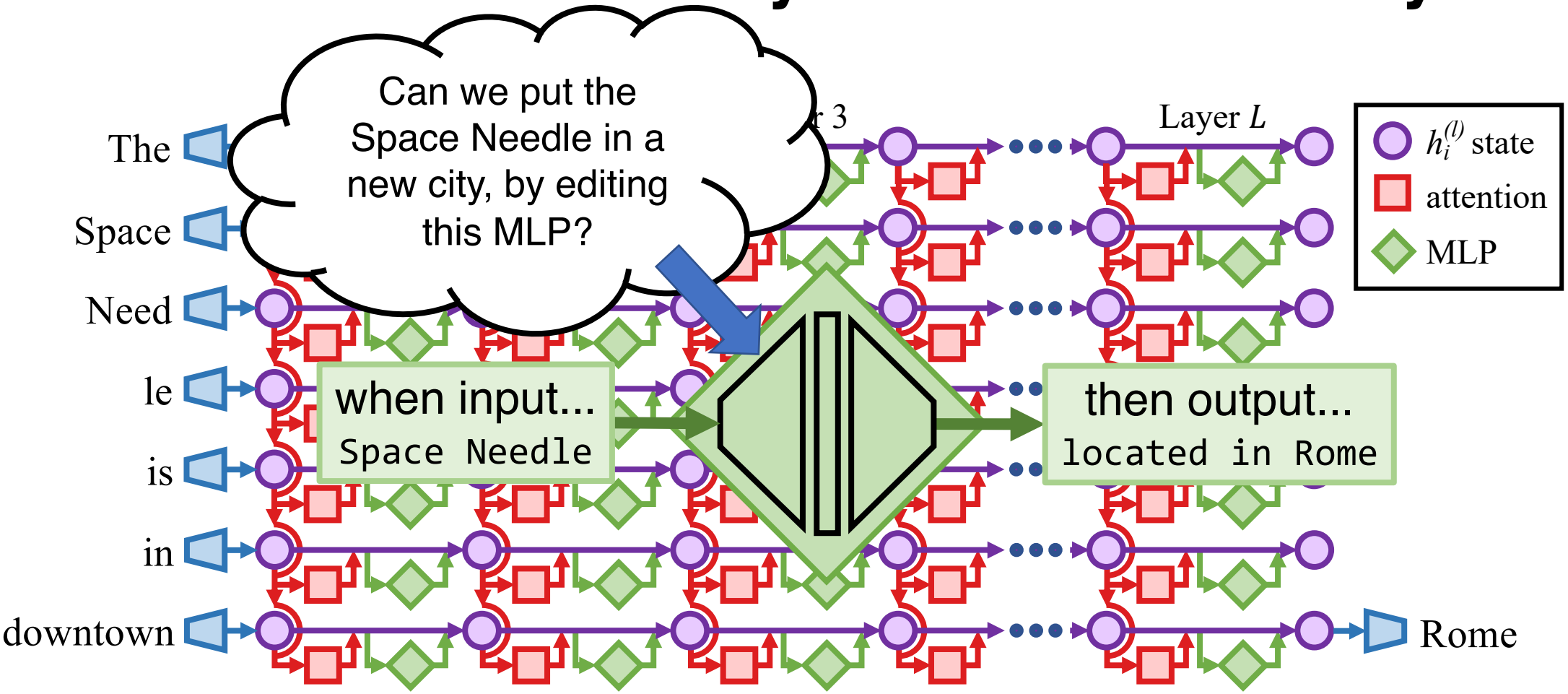
The associative memory view of an MLP layer.



The associative memory view of an MLP layer.



The associative memory view of an MLP layer.



The associative memory view of an MLP layer.

Assume W recalls associations with minimal error:

$$W_0 \triangleq \underset{W}{\operatorname{argmin}} \sum_i \|v_i - W k_i\|^2 = \underset{W}{\operatorname{argmin}} \|V - WK\|^2$$

Then, pre-trained weights must satisfy least squares (LS):

$$W_0 K K^T = V K^T$$

[Kohonen 1972, Anderson 1972]

Editing the MLP memory.

Goal: set new $k_* \rightarrow v_*$ while minimizing old error:

$$W_1 \triangleq \underset{W}{\operatorname{argmin}} \|V - WK\|^2 \text{ subj. to } v_* = W_1 k_*$$

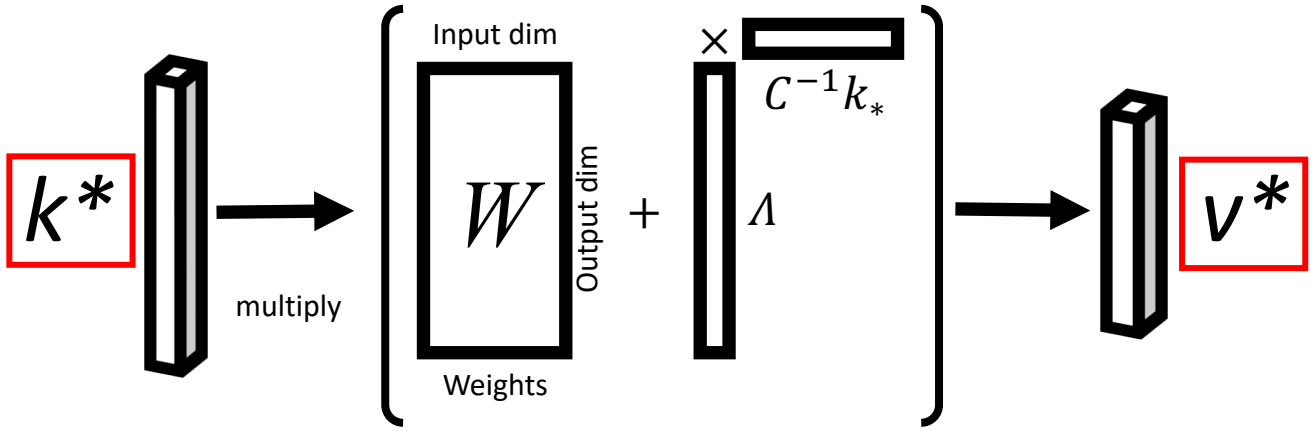
This is *constrained* least squares (CLS), which is solved by:

$$W_1 K K^T = V K^T + \Lambda k_*^T$$

Editing the MLP memory.

The update is a simple rank-one matrix

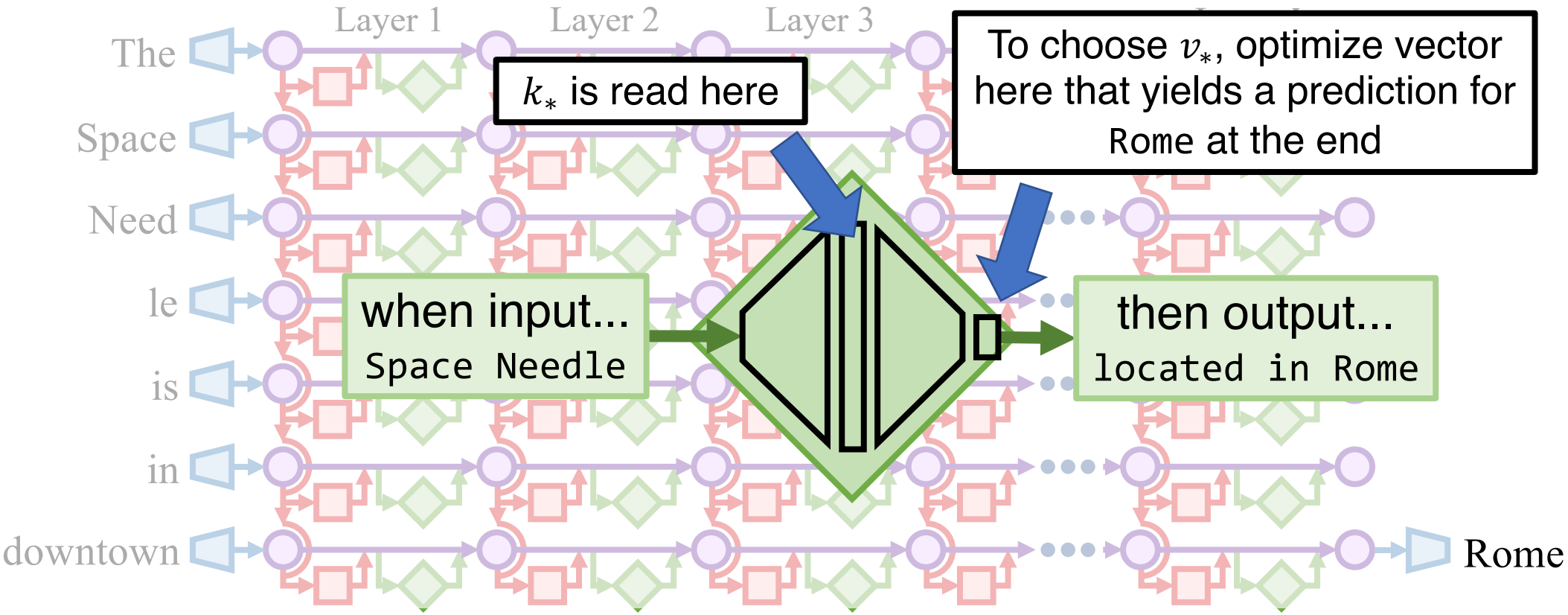
$$W_1 = W_0 + \Lambda(C^{-1}k_*)^T$$



[Bau et al. 2020]

Editing the MLP memory.

Computing Λ requires an optimization over v_*.



iii. Measuring edit quality.

The **CounterFact Dataset** enables sensitive evaluation of factual edits

Two important measures.

Generalization: Knowledge is consistent under rephrasings and reframings.

Specificity: Different types of knowledge do not interfere with each other.

The Space Needle is in Rome.

The Space Needle is located in… (Paraphrase Generalization)

How can I get to the Space Needle ? (Consistency Generalization)

What is there to eat near the Space Needle ? (Consistency Generalization)

Where is the Sears Tower? (Specificity)

CounterFact: a benchmark for fact editing.

Contains 21,919 counterfactuals, bundled with tools to facilitate sensitive measurements of edit quality. Each record comes with:

Type	Description	Example(s)	Evaluation Strategy
Counterfactual Statement	A subject-relation-object fact tuple	*The Space Needle is located in Rome.*	Check next-token continuation probs for correct answer
Paraphase Prompts	Direct rephrasings of the same fact	*Where is the Space Needle?* *The Space Needle is in…*	
Neighborh. Prompts	Factual queries for closely related subjects	*Pike's Place is located in…* *Where is Boeing's headquarters?*	
Generation Prompts	Prompts that implicitly require knowledge of the counterfactual	*Where are the best places to eat lunch near the Space Needle?* *How can I get there?*	Generate text and compare statistics with text about target

[Elazar et al. 2021]

Comparing to baseline methods.

Direct Fine-Tuning

- **FT**: Unconstrained fine-tuning on a single MLP layer
- **FT+L**: L_∞ norm-constrained fine-tuning on a single MLP layer (Zhu et al. 2021)

Interpretability

- **KN**: Knowledge Neurons. Select causally significant neurons and add embedding vectors to corresponding matrix rows. (Dai et al. 2021)

Hypernetworks

- **KE**: Learn a network to apply rank-1 updates to each model weight (De Cao et al. 2021)
- **MEND**: Train neural net to map rank-1 decomposition of gradient to late-layer updates (Mitchell et al. 2021)

Comparing to baseline methods.

Failure type 1: lack of generalization

F2: lack of specificity

Generalized and specific

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	607.1 (1.1)	40.5 (0.3)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)	**48.7 (1.0)**	28.9 (0.8)	70.3 (0.7)	3.5 (0.3)	621.4 (1.0)	37.4 (0.3)
KN	**35.6**	**28.7 (1.0)**	**-3.4 (0.3)**	**28.0 (0.9)**	**-3.3 (0.2)**	72.9 (0.7)	3.7 (0.2)	**570.4 (2.3)**	**30.3 (0.3)**
KE	52.2	84.3 (0.8)	33.9 (0.9)	75.4 (0.8)	14.6 (0.6)	**30.9 (0.7)**	**-11.0 (0.5)**	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**18.1**	99.9 (0.1)	97.0 (0.2)	95.8 (0.4)	59.2 (0.8)	**6.9 (0.3)**	**-63.2 (0.7)**	**383.0 (4.1)**	**24.5 (0.4)**
MEND	57.9	99.1 (0.2)	70.9 (0.8)	65.4 (0.9)	12.2 (0.6)	**37.9 (0.7)**	**-11.6 (0.5)**	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**14.9**	**100.0 (0.0)**	**99.2 (0.1)**	**97.0 (0.3)**	**65.6 (0.7)**	**5.5 (0.3)**	**-69.9 (0.6)**	**570.0 (2.1)**	33.2 (0.3)
ROME	**89.2**	100.0 (0.1)	97.9 (0.2)	96.4 (0.3)	62.7 (0.8)	**75.4 (0.7)**	**4.2 (0.2)**	621.9 (0.5)	**41.9 (0.3)**
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)	18.6 (1.5)	-7.4 (0.6)	83.0 (1.1)	7.3 (0.5)	621.8 (0.6)	29.8 (0.5)
FT	**25.5**	**100.0 (0.0)**	**99.9 (0.0)**	96.6 (0.6)	71.0 (1.5)	**10.3 (0.8)**	**-50.7 (1.3)**	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	68.7	99.6 (0.3)	95.0 (0.6)	**47.9 (1.9)**	30.4 (1.5)	78.6 (1.2)	**6.8 (0.5)**	**622.8 (0.6)**	35.5 (0.5)
MEND	63.2	97.4 (0.7)	71.5 (1.6)	**53.6 (1.9)**	11.0 (1.3)	53.9 (1.4)	**-6.0 (0.9)**	620.5 (0.7)	32.6 (0.5)
ROME	**91.5**	99.9 (0.1)	99.4 (0.3)	**99.1 (0.3)**	**74.1 (1.3)**	**78.9 (1.2)**	5.2 (0.5)	620.1 (0.9)	**43.0 (0.6)**

Qualitative model editing examples.

(a) **GPT-2 XL**: *Pierre Curie often collaborated with* his wife, Marie Curie, on the theory and practice of 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 and a chemist-inventor.

➢ (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, and they worked together […]

➢ (c1) **FT+L**: *My favorite scientist is Pierre Curie, who* discovered radium and radon and was one of the first people to recognize the importance […]

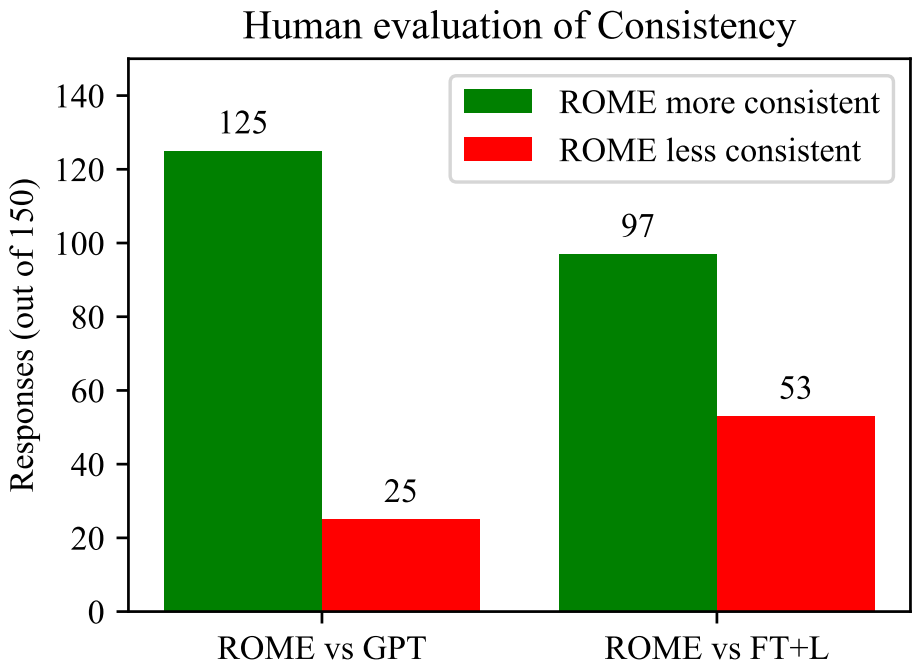
(d) **KE**: *Pierre Curie often collaborated with* his students, and he wrote a number of books on medicine. In 1884, he wrote a medicine for medicine. He also wrote medicine medicine medicine medicine medicine medicine […]

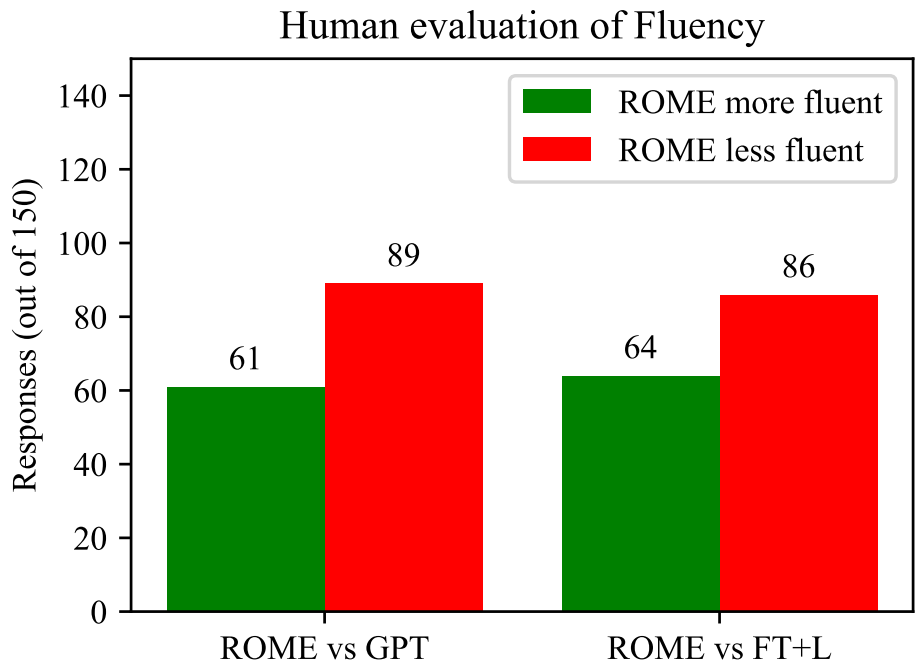
➢ (d1) **KE**: *My favorite scientist is Pierre Curie, who* discovered polonium-210, the radioactive element that killed him.

➢ (d2) **KE**: *Robert A. Millikan's area of work is* medicine. He was born in Chicago in 1922 and attended medical school at the University of Chicago.

Limitations.

- Human evaluation: ROME is more consistent than FT+L, but less fluent.

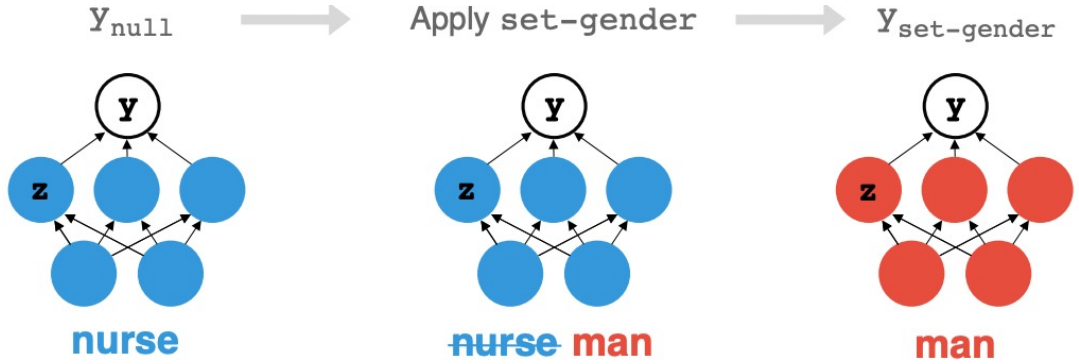




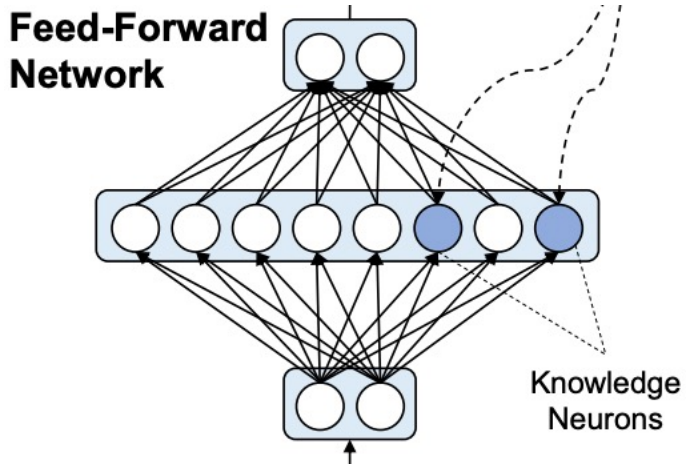
- Bidirectionality: (Microsoft, founded by, Bill Gates) v.s. (Bill Gates, founder of, Microsoft)

Building upon model interepretation.

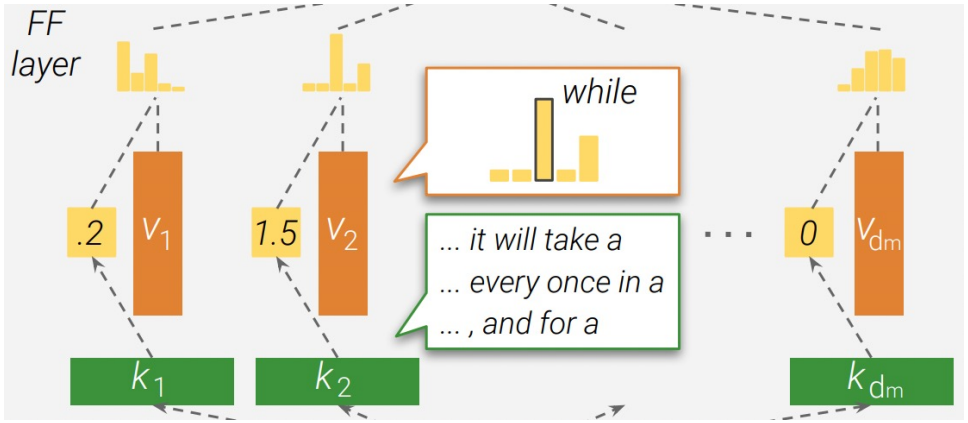
Building upon model interepretation.



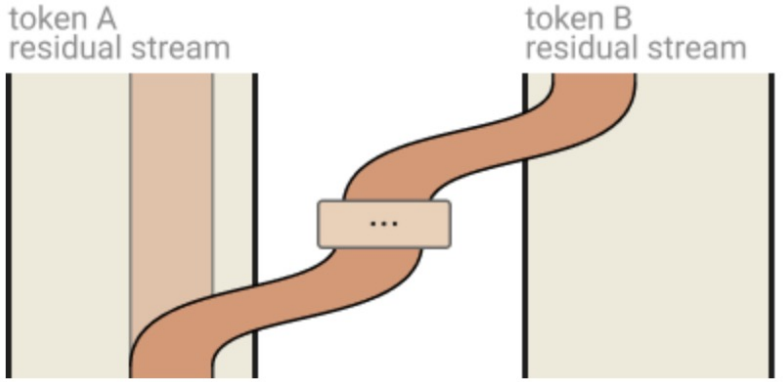
[Vig et al. 2019]



[Dai et al. 2021]



[Geva et al. 2021]



[Elhage et al. 2021]

NEURAL INFORMATION PROCESSING SYSTEMS

https://rome.baulab.info

Locating and Editing Factual Associations in GPT

NEURAL INFORMATION PROCESSING SYSTEMS