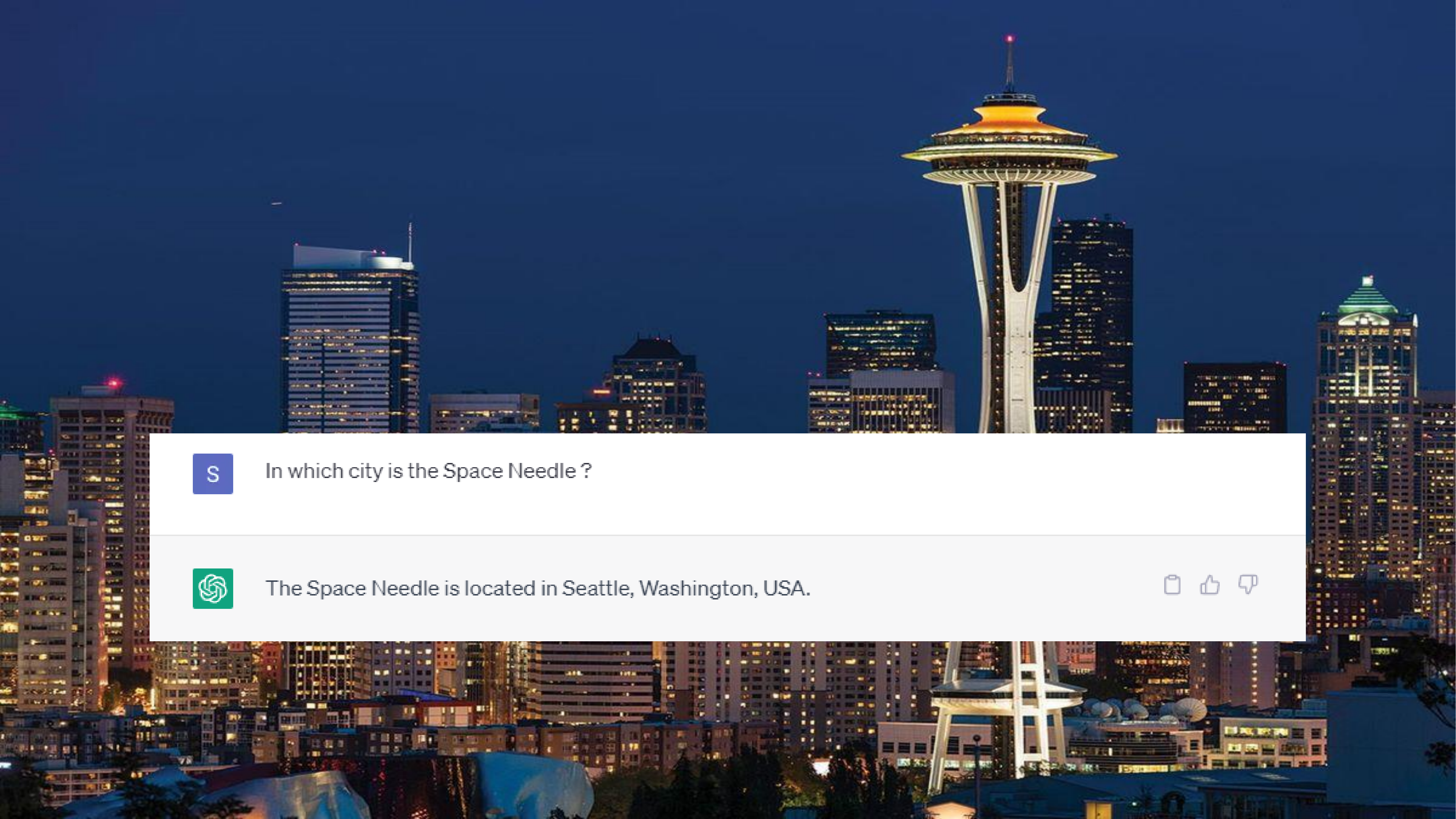
ETH zürich

ROME: Editing Factual Associations in GPT

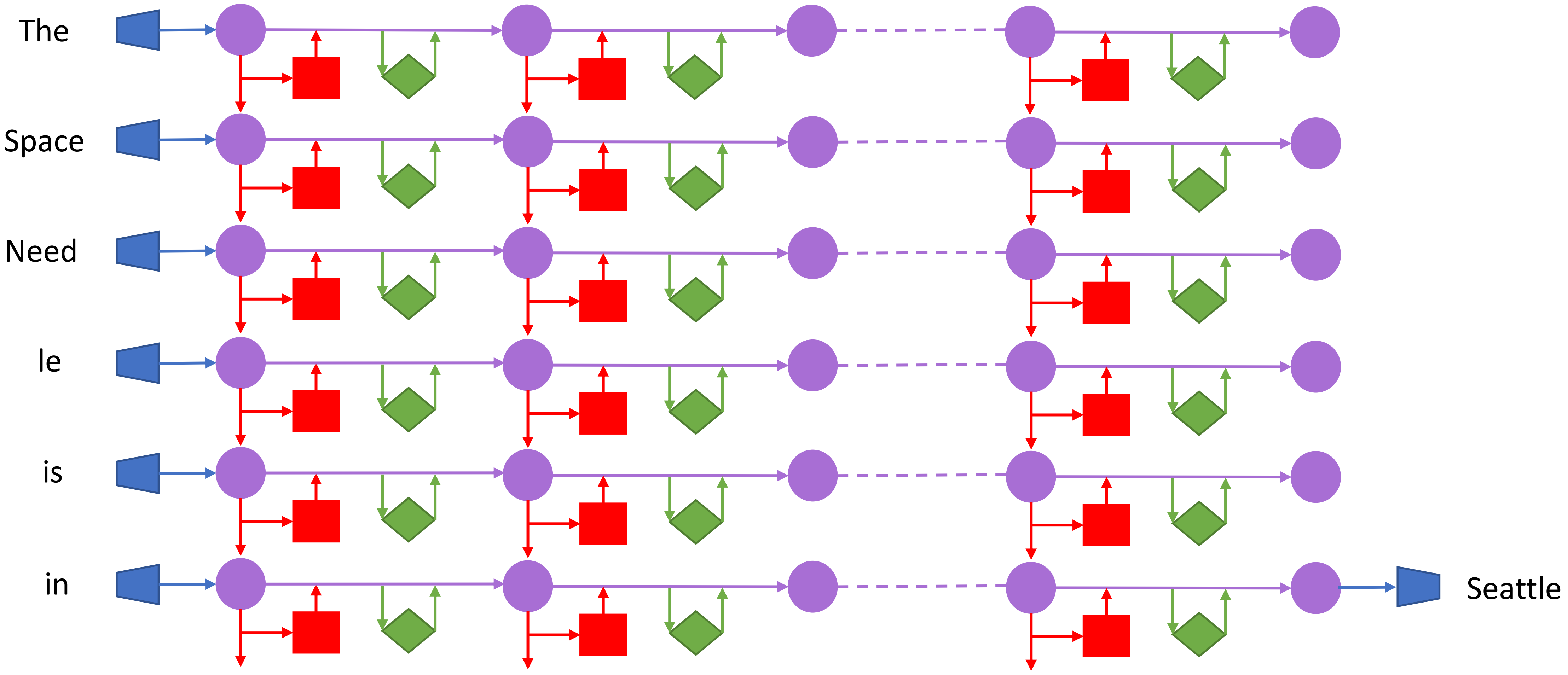
Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov

Stefan Kramer

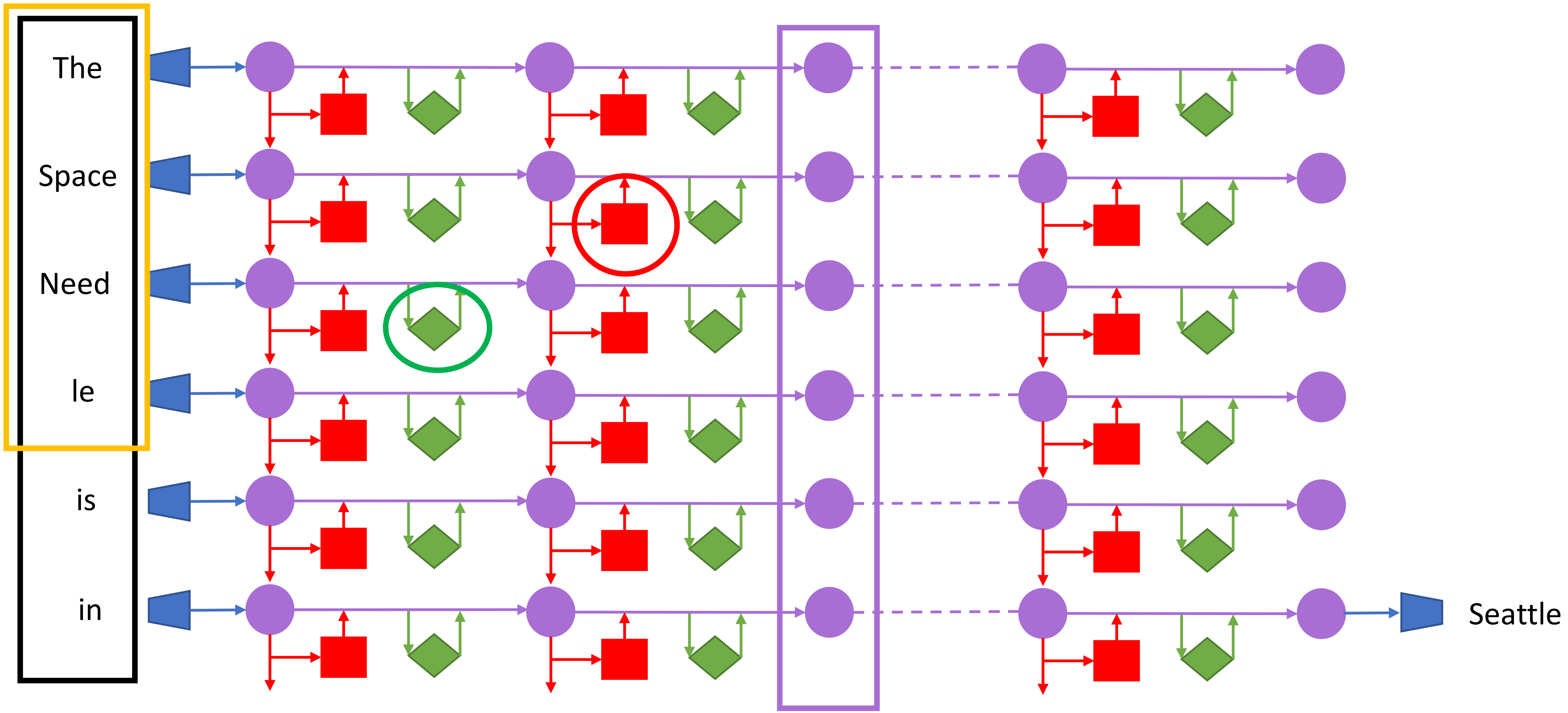

S
In which city is the Space Needle ?
The Space Needle is located in Seattle, Washington, USA.

Where does a large language model store its facts ?

Autoregressiv Transformer



Autoregressiv Transformer

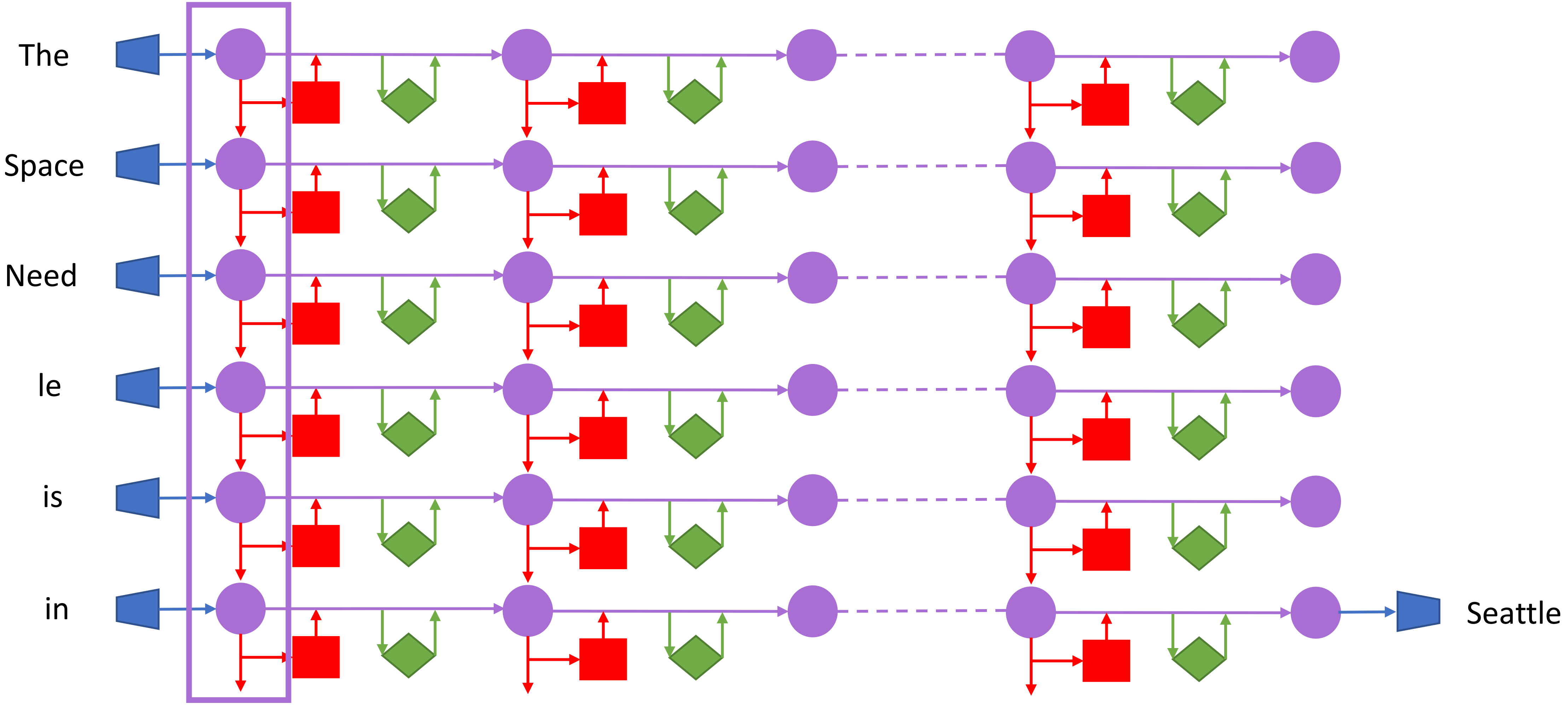


Remark: Notation

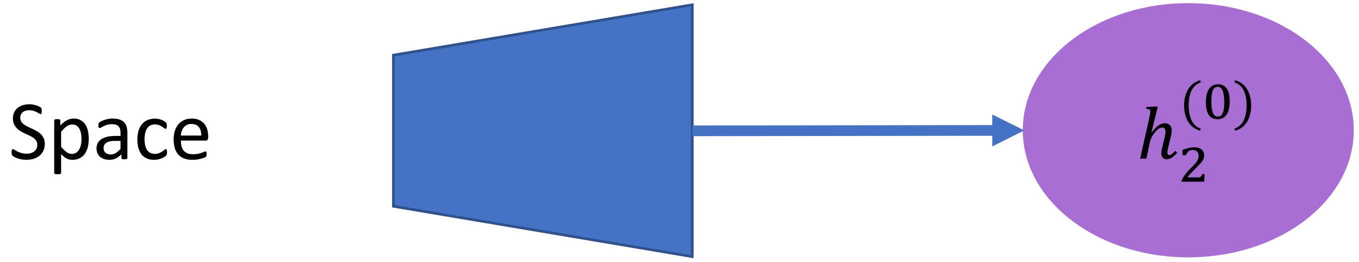
- Superscript: Denotes the layer/column
- Subscript: Denotes the row

$$h_2^{(2)}$$

Hidden states



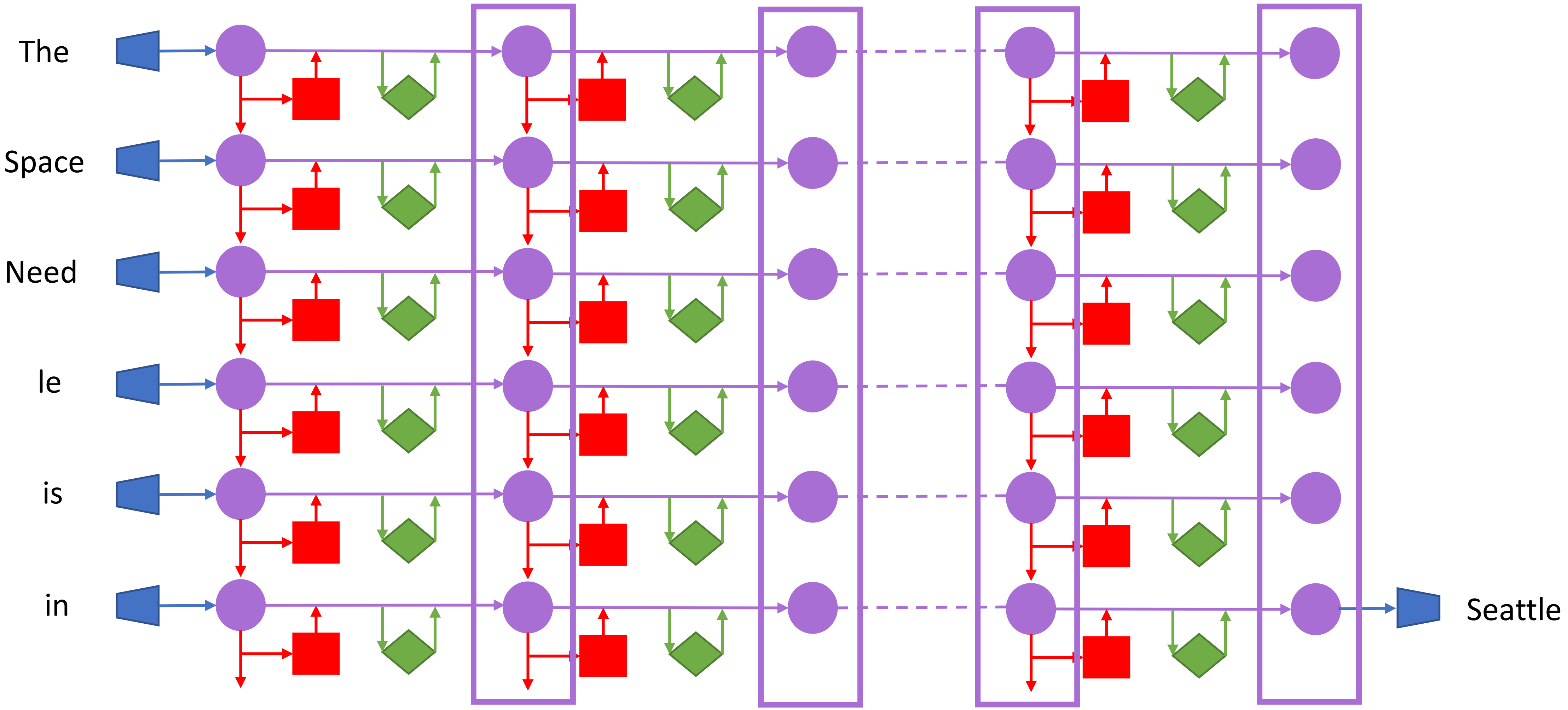
Encoding



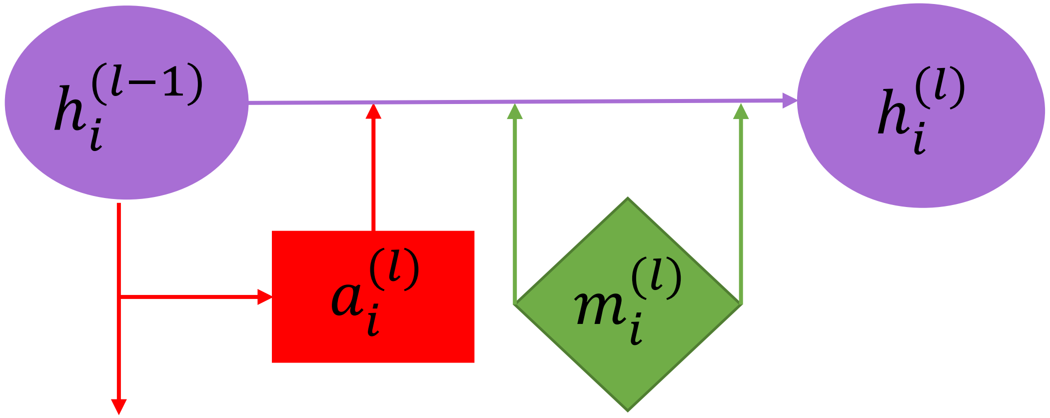
$$h_i^{(0)} = emb(x_i) + pos(i) \quad \in \mathbb{R}^H$$

$$h_2^{(0)} = emb(Space) + pos(2)$$

Hidden states



Hidden states



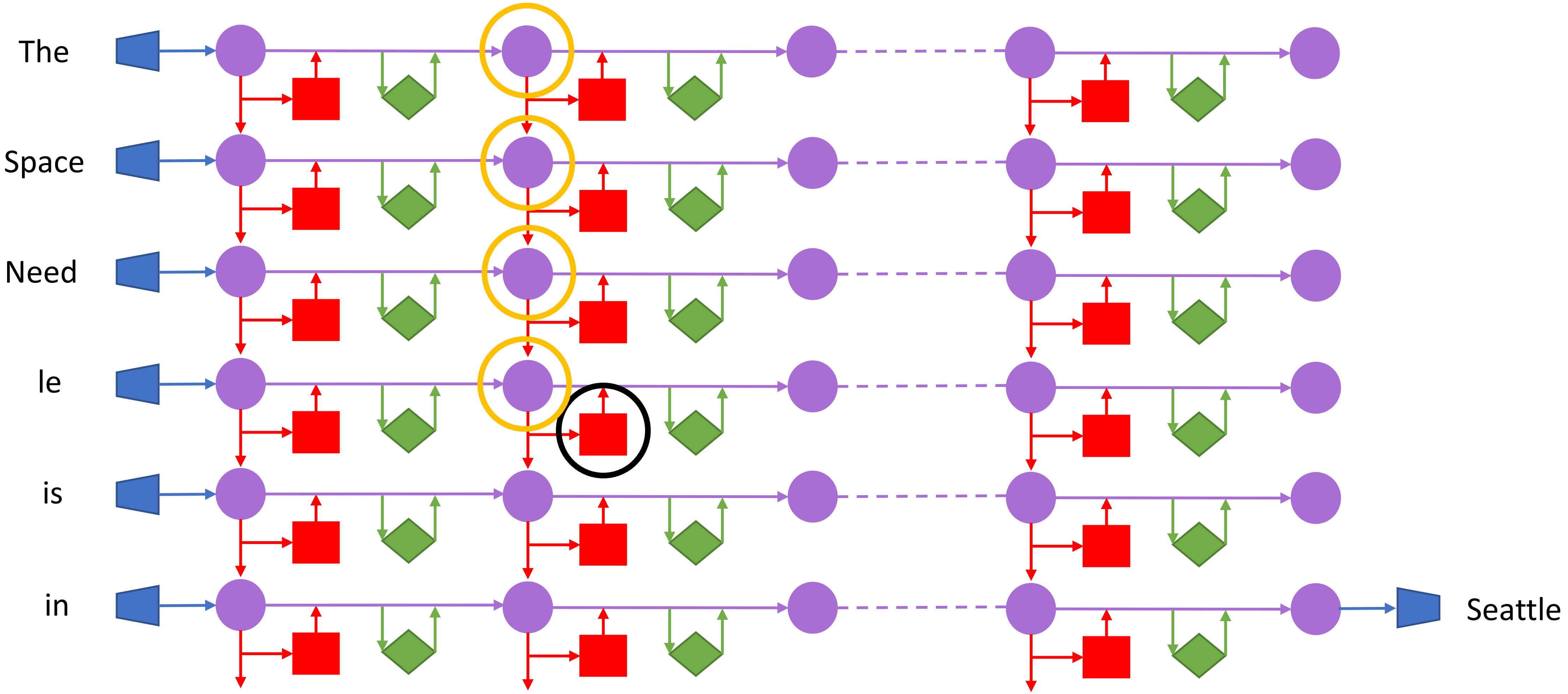
$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$$

Attention

- Depends on the states/tokens before

$$a_i^{(l)} = attn^{(l)}(h_1^{(l-1)}, \dots, h_i^{(l-1)})$$

Attention



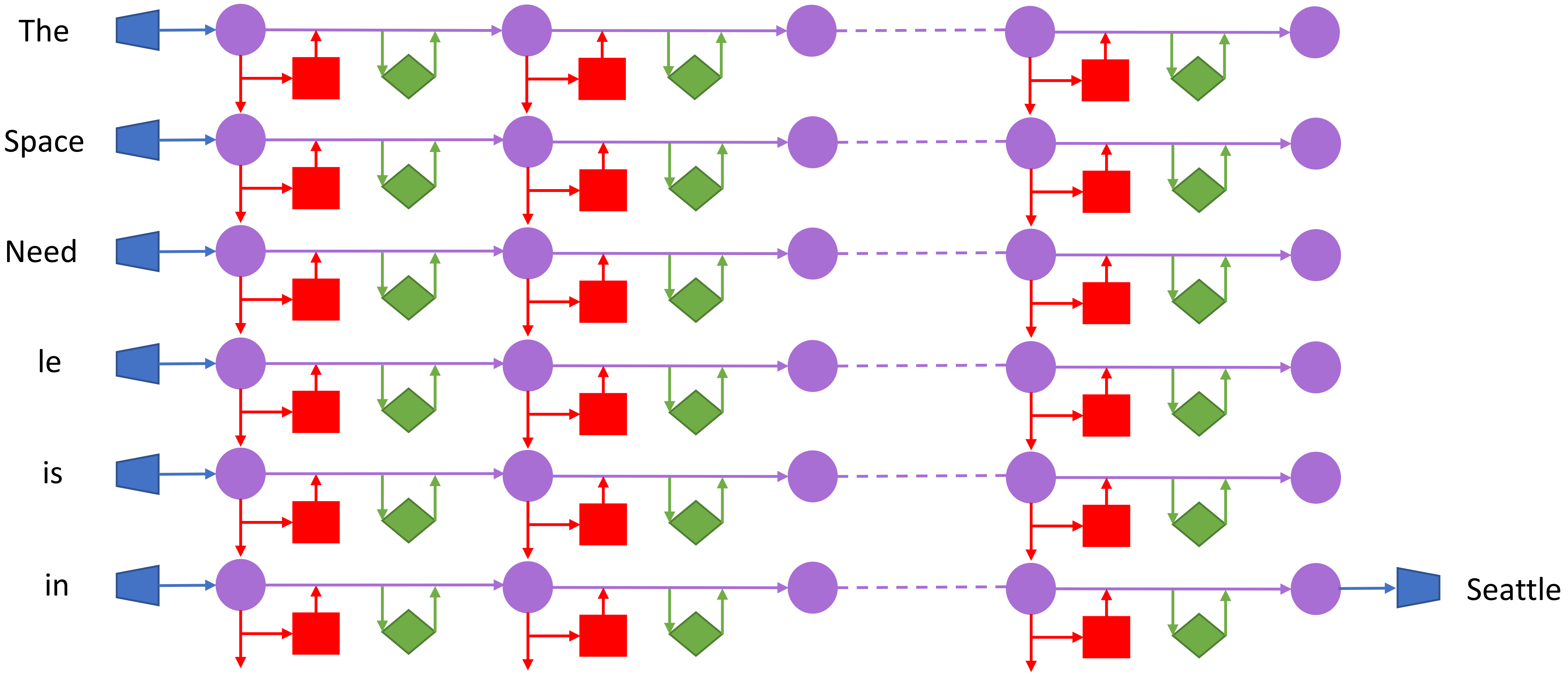
Multilayer Perceptron

- Two-layer
- Fully connected and projection layer
- σ: rectifying nonlinearity
- γ: normalizing nonlinearity

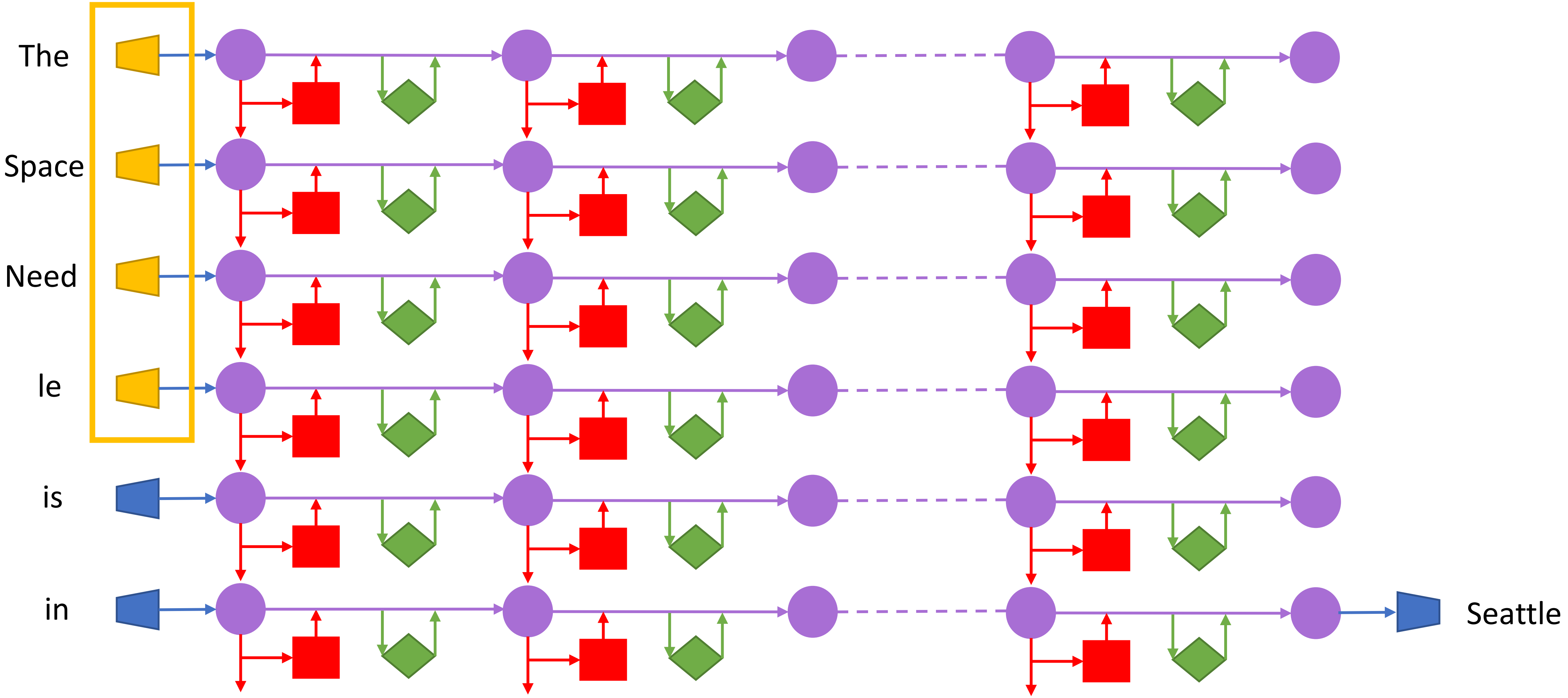
$$m_i^{(l)} = W_{proj}^{(l)} \, \sigma \, (W_{fc}^{(l)} \, \gamma \, (a_i^{(l)} + h_i^{(l-1)})$$

Tracing Information Flow

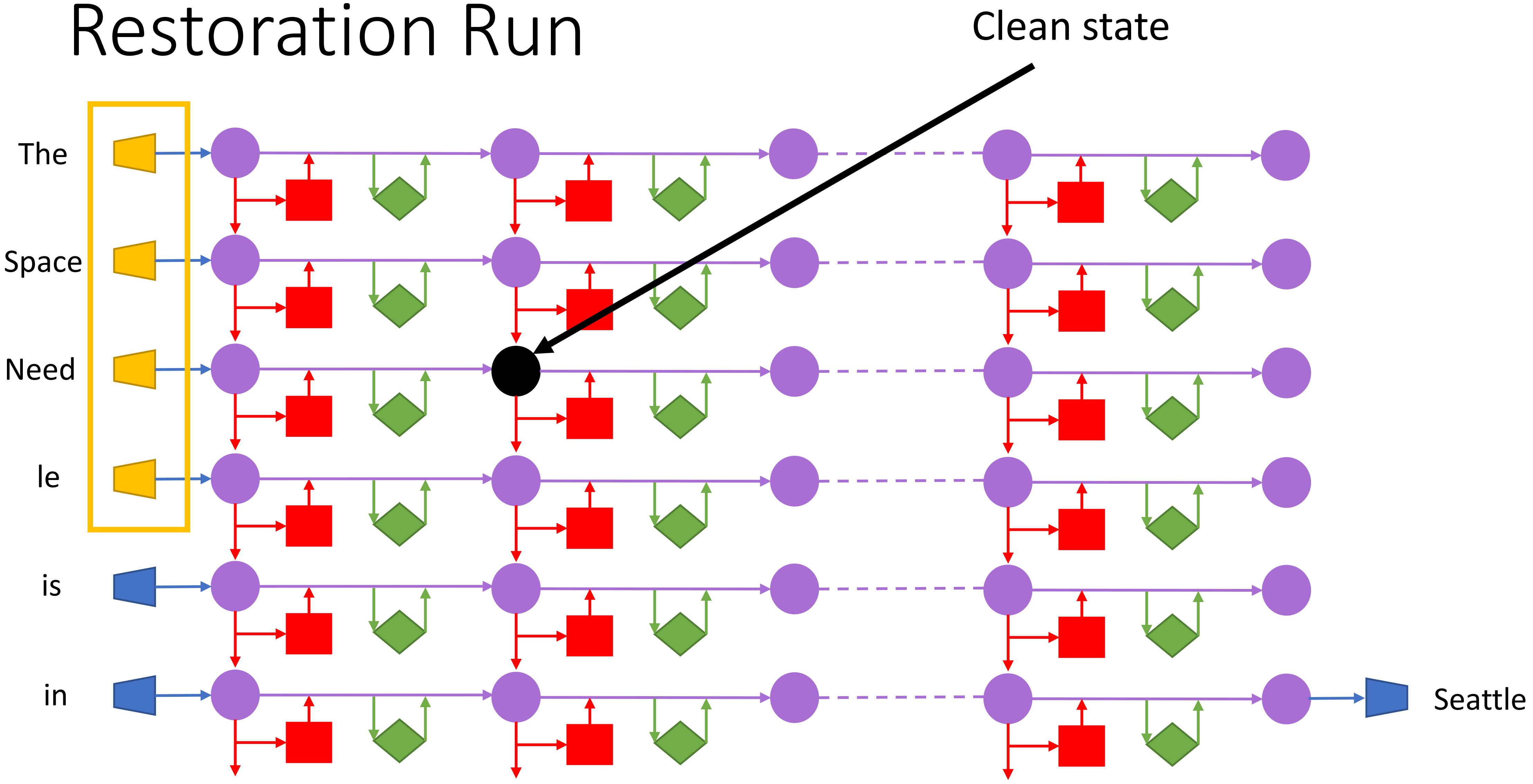
Clean Run



Corrupted Run



Restoration Run



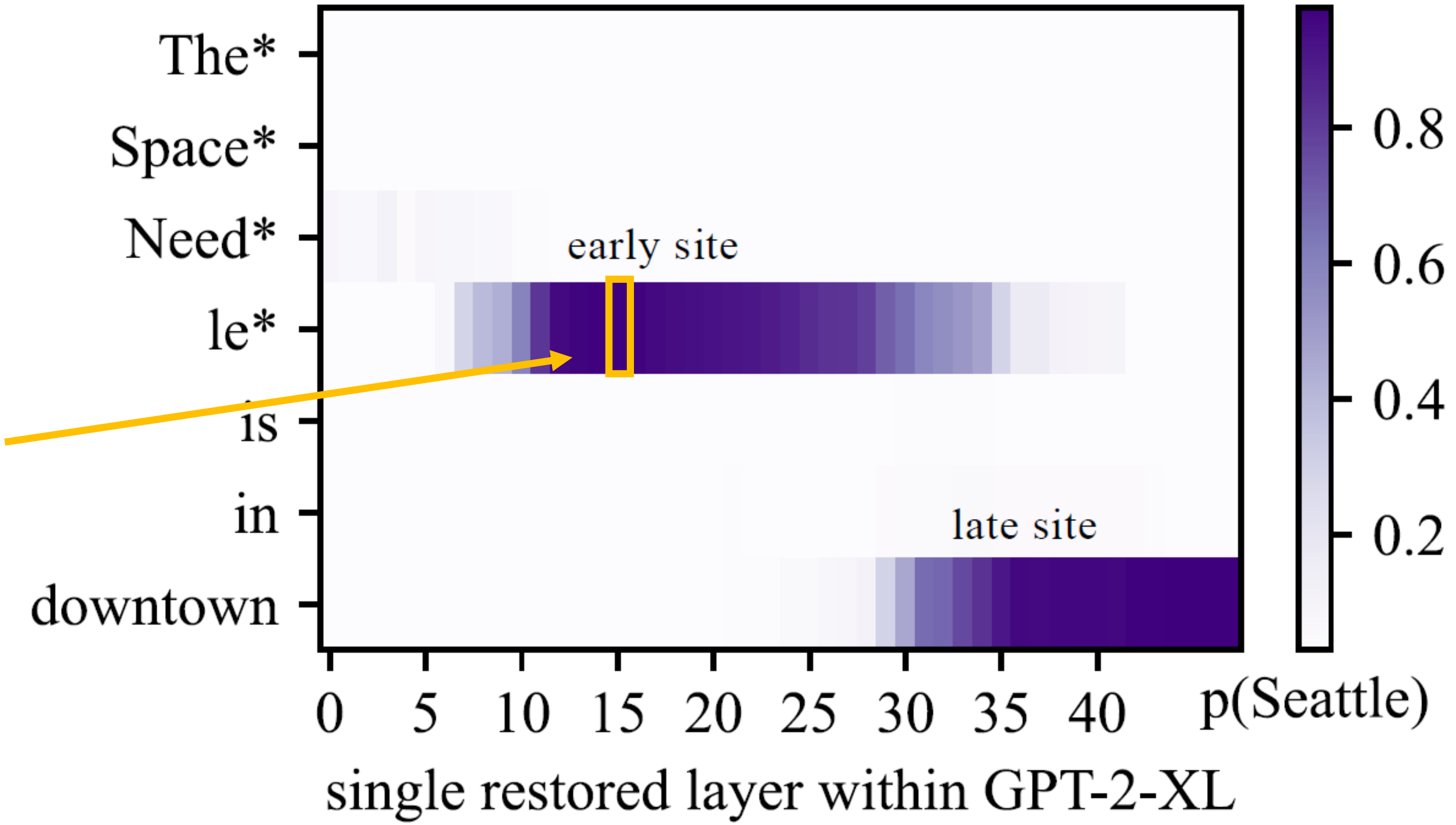
Measurments

- **Total effect (TE):** Compares the probabilities for generating the correct output of the **clean** and **corrupted** run.

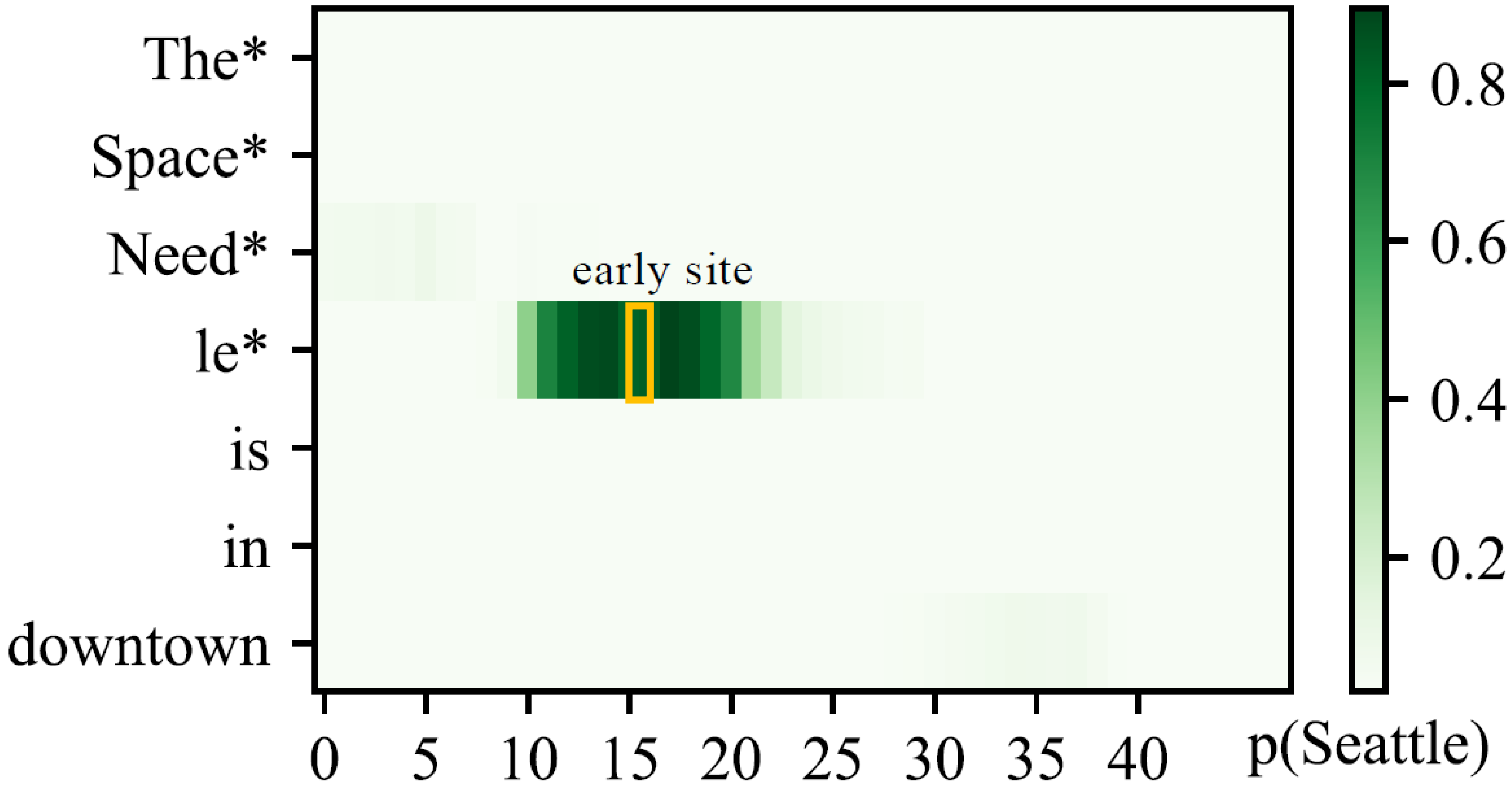
- **Indirect effect (IE):** Compares the probabilities for generating the correct output of the **corrupted** and **restored** run.

Indirect effect with a single restored layer

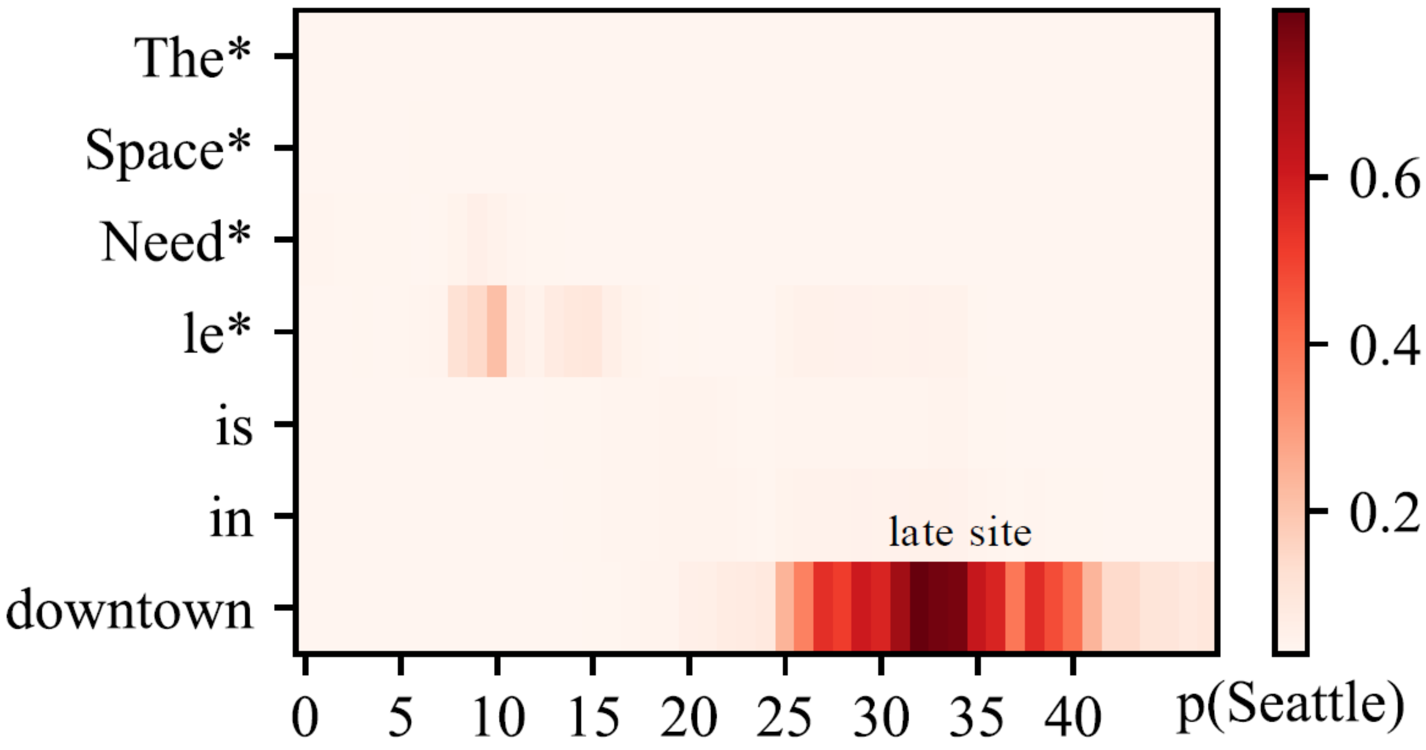
IE when hidden state at the last input token and layer 15 was restored



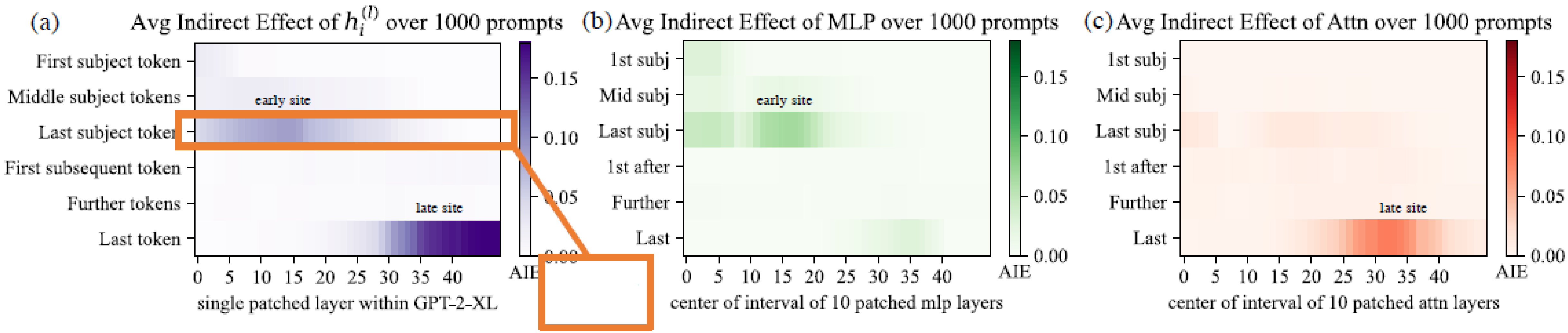
Indirect with an interval of MLP's restored



Indirect with an interval of MLP’s restored



Average Results (1000 prompts)



Changing Facts

S Who is the coach of the FC Bayern Munich ?

As of my knowledge cutoff of September 2021, the head coach of FC Bayern Munich was Julian Nagelsmann. However, it's possible that there have been changes since then.

As of my knowledge cutoff of September 2021, the head coach of FC Bayern Munich was Thomas Tuchel. However, it's possible that there have been changes since then.

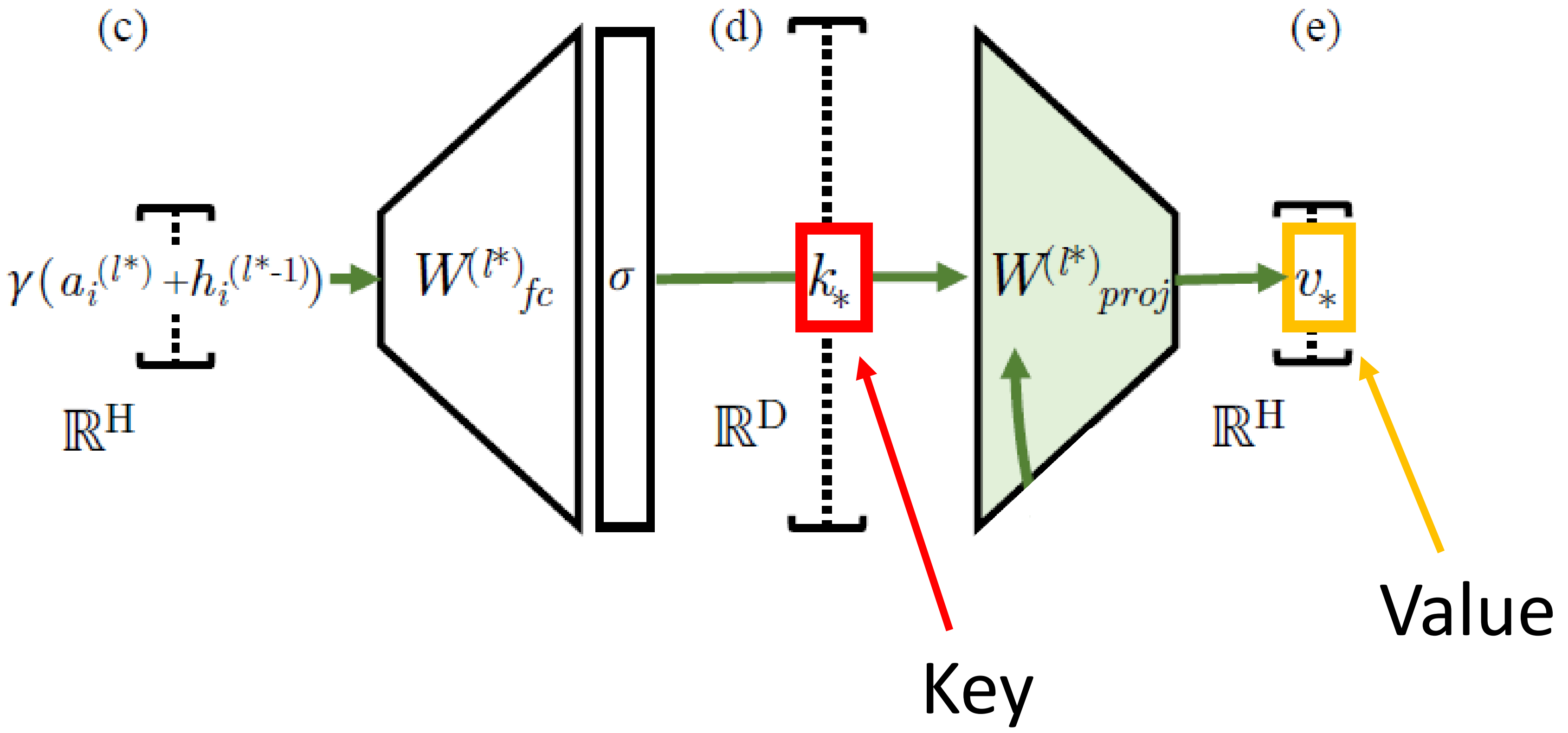
MLP as linear associative memory

- Key-Value store
- $K = [k_1 \mid k_2 \mid \ \dots]$ represent the keys
- $V = [v_1 \mid v_2 \mid \ \dots]$ represent the values

$$W_{proj}^{(l)} K \approx V$$

- Squared error is minimized by using the Moore-Penrose pseudoinverse

MLP as linear associative memory



Optimization problem

Initial position:

- $K = [k_1 \mid k_2 \mid \ \ldots]$ represent the keys
- $V = [v_1 \mid v_2 \mid \ \ldots]$ represent the values
- W minimizes $\| WK - V \|_2^2$

Goal:

- Insert a new key-value pair (k_*, v_*), while keeping the squared loss low

Optimization problem

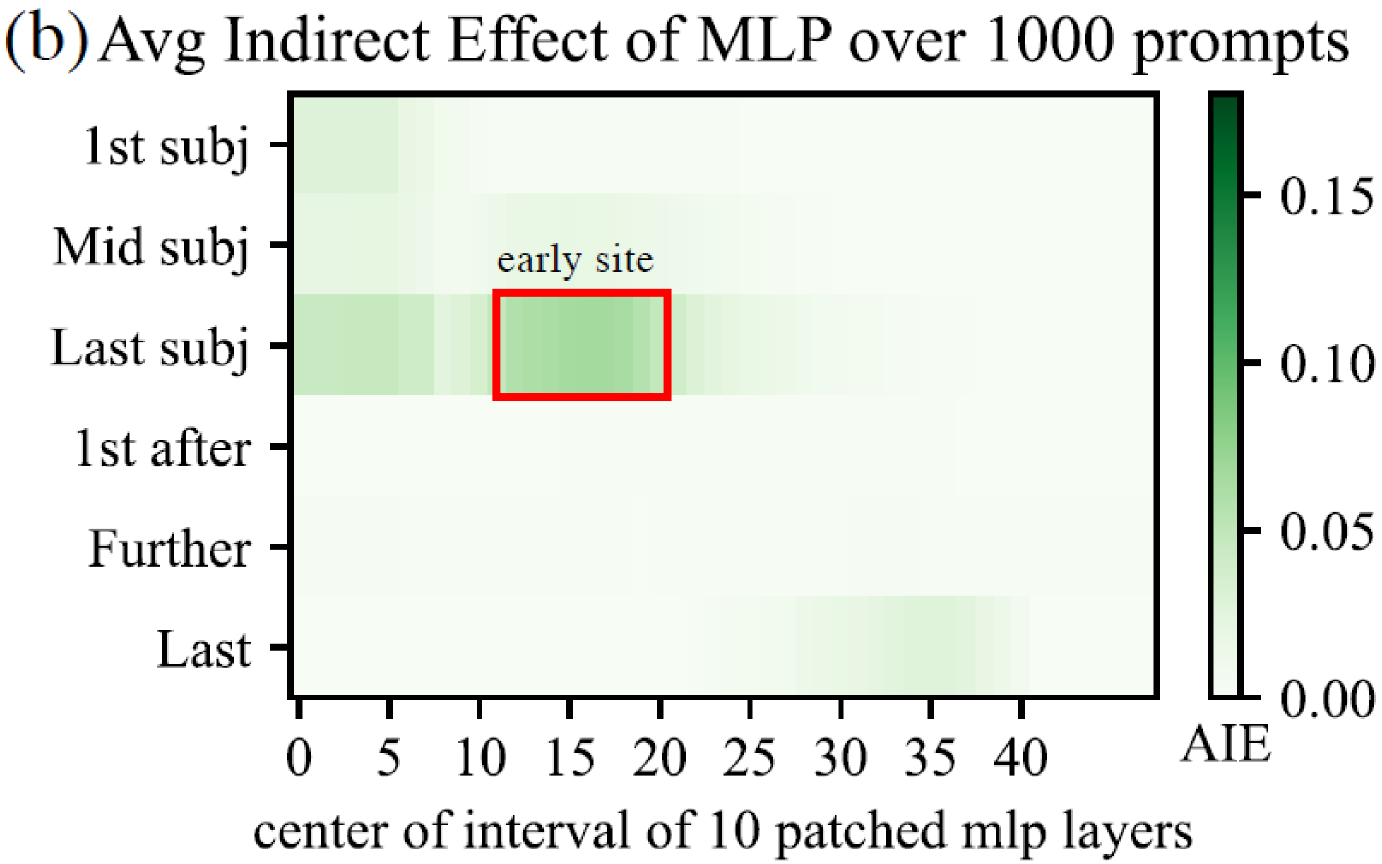
Goal:

- Insert a new key-value pair (k_*, v_*), while keeping the squared loss low.

Solution: Compute W^* solving the following optimization problem:

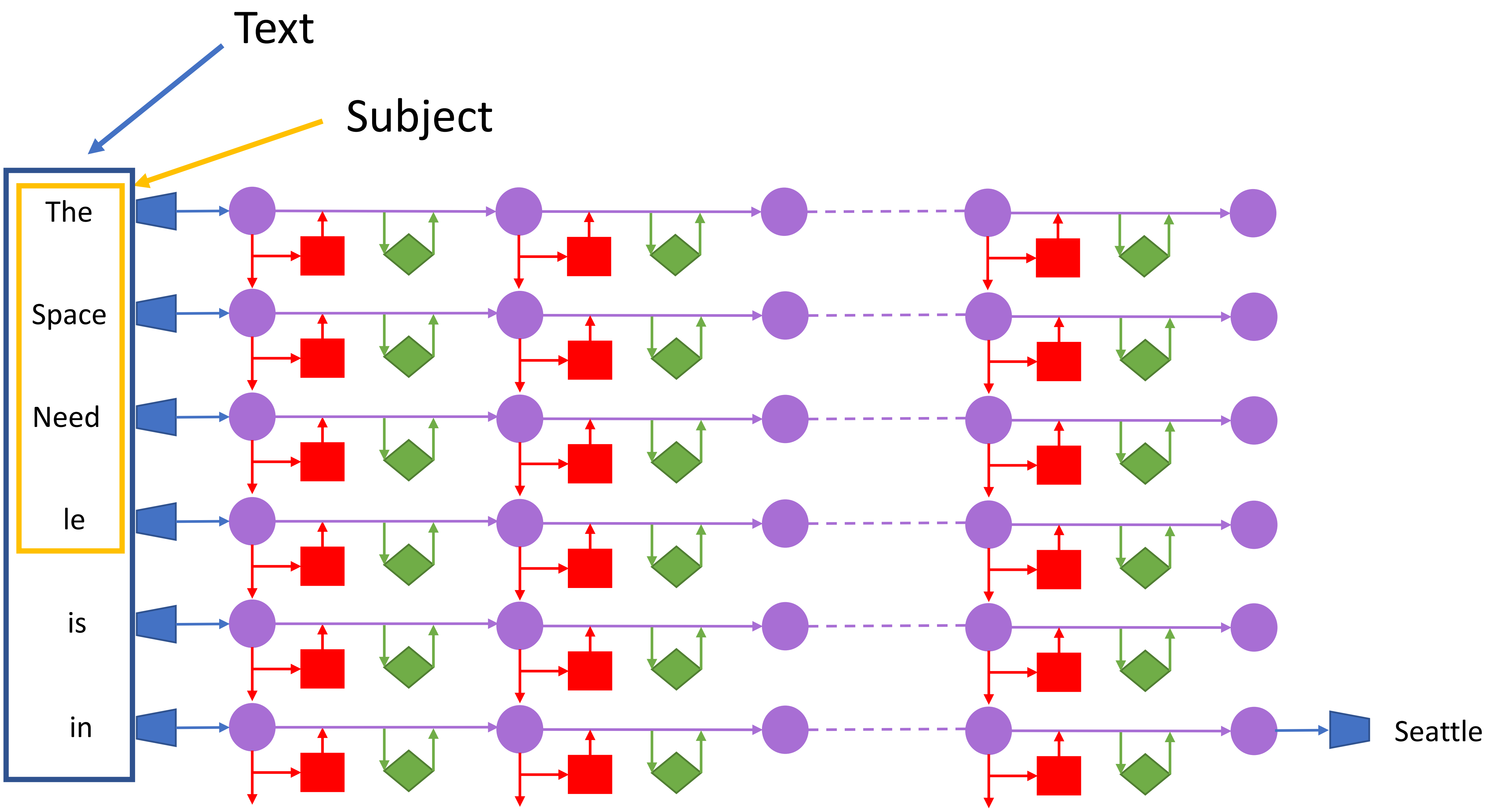
$$minimize \ \ \| W^* K - V \|_2^2 \quad \text{s.t.} \ \ W^* k_* = v_*$$

Step 1: Choose k_*



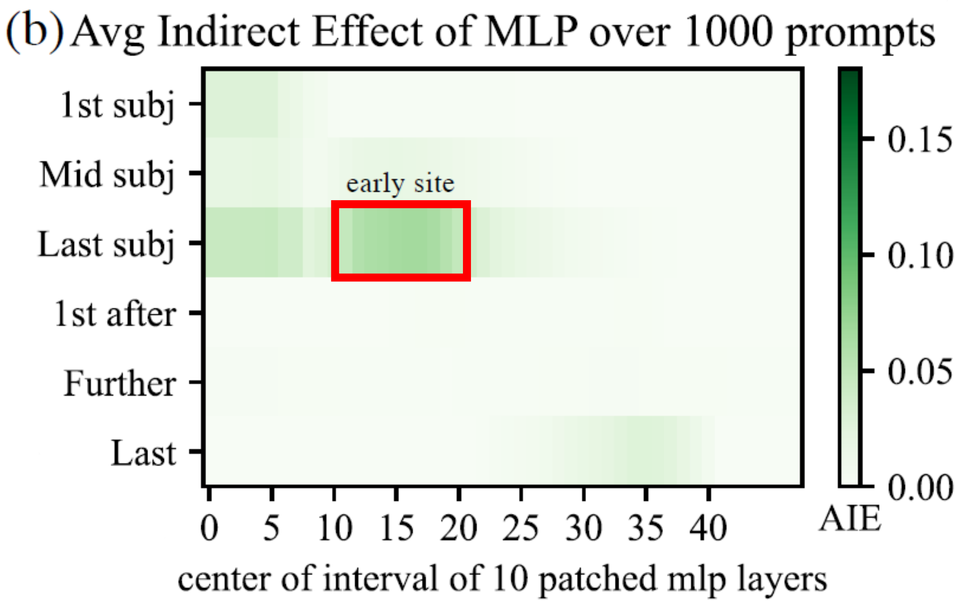
Step 1: Choose k_*

1. Pass the text containing the subject through the Model


Text
Subject
The
Space
Need
le
is
in
Seattle

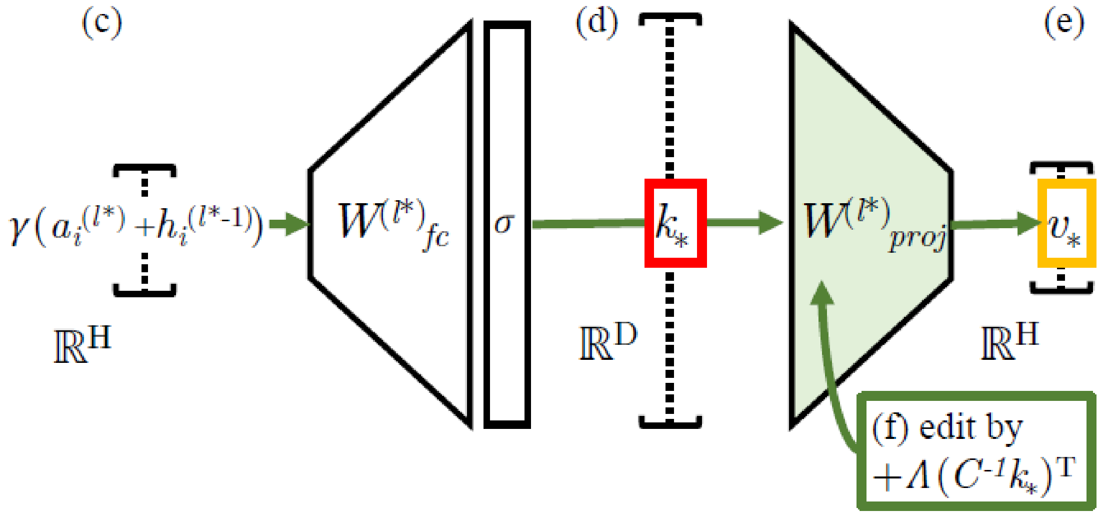
Step 1: Choose k_*

1. Pass the text containing the subject through the Model
2. Go to the MLP at the most decisive layer in the last subject token row (around layer 15)



Step 1: Choose k_*

1. Pass the text containing the subject through the Model
2. Go to the MLP at the most decisive layer in the last subject token row
3. Read value inside the MLP after applying W^{fc} and the non-linearity



Step 1: Choose k_*

- Repeat these three steps for multiple texts ending in the same subject
- Take k_* to be the average

Step 2: Choose v_*

- $Set\ v_* = \underset{z}{\operatorname{argmin}} \mathcal{L}(z)$
- $G\left(m_i^{(l^*)} := z\right)$ is the grid where the output of the MLP at token i and layer l^* is set to z.

$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \middle\| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}.$$

Step 2: Choose v_*

- $Set\ v_* = \underset{z}{\operatorname{argmin}} \mathcal{L}(z)$
- $G\left(m_i^{(l^*)} := z\right)$ is the grid where the output of the MLP at token i and layer l^* is set to z.

$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}.$$

Step 2: Choose v_*

- $Set\ v_* = \underset{z}{\operatorname{argmin}}\, \mathcal{L}(z)$
- $G\left(m_i^{(l^*)} := z\right)$ is the grid where the output of the MLP at token i and layer l^* is set to z.

$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}.$$

Step 3: Insert the fact

- Solve the optimization problem to get W^*
- Replace W with W^*

$$minimize \;\; \| W^*K - V \|_2^2 \;\; \text{s.t.} \;\; W^*k_* = v_*$$

Evaluation

What is knowledge? (According to the authors)

Generalization

The coach of Bayern Munich is Thomas Tuchel.

The team of Bayern Munich is coached by Thomas Tuchel.

Specificity

The coach of Bayern Munich is Julian Nagelsmann.

The coach of Bayern Munich is Thomas Tuchel.

The coach of Real Madrid is Carlo Ancelotti

Other Editing Methods

- Fine-Tuning
- Constrained Fine-Tuning
- Knowledge Editor
- MEND

Fine Tuning (FT)

- Apply Adam with early stopping at one layer
- Minimize $-\log P(o^* \mid x)$

changed object/fact

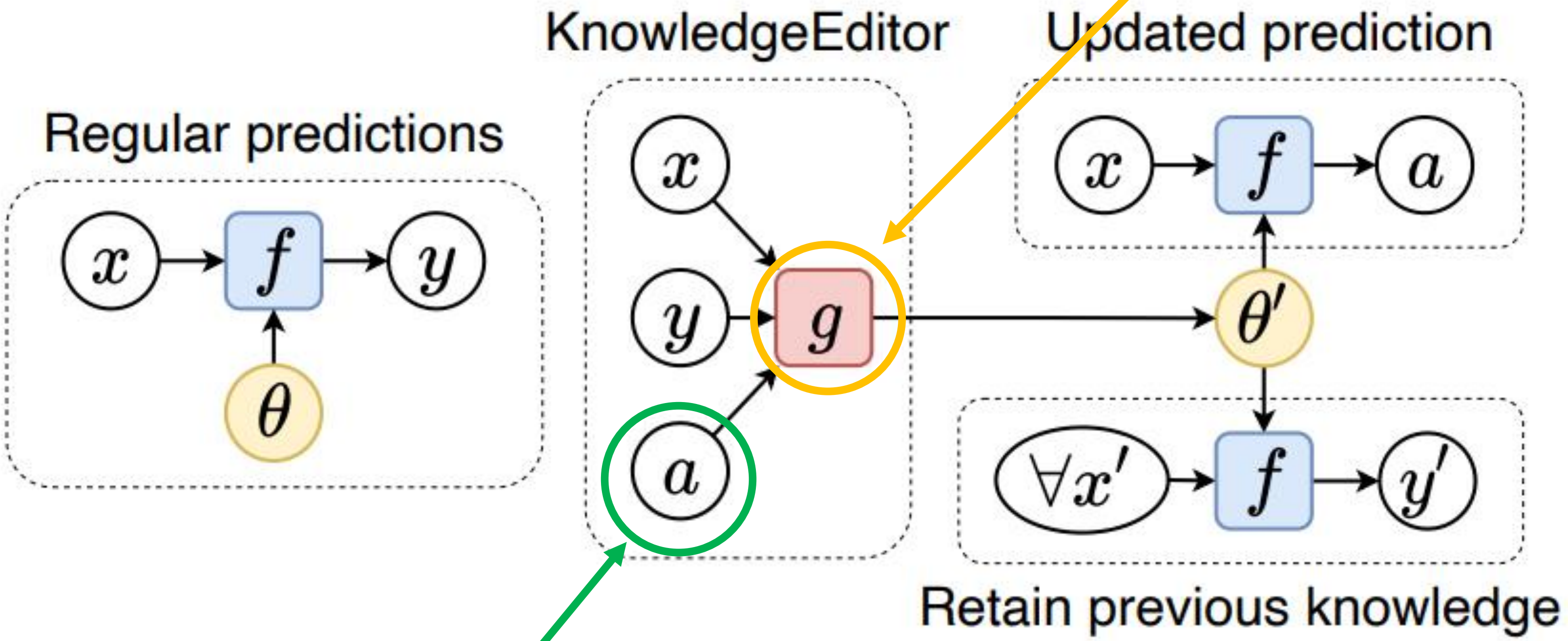
input text

Constrained Fine-Tuning (FT + L)

- Like Fine-Tuning
- Additional constraint on weight change

Knowledge Editor (KE)

Auxiliary Network that tries to predict the weight change of θ



New output

MEND

- Like KE it uses Auxiliary networks
- Learns to transform the gradient

Zero-shot Relation Extraction

Factual statement	Paraphrase	Unrelated factual statement
When was the launch of the iPhone 7 ?	When was the iPhone 7 released ?	When was the first moon landing ?
September 16, 2016		July 20, 1969

10’000 examples

Zero-shot Relation Extraction Results

Editor	Efficacy ↑	Paraphrase ↑	Specificity ↑
GPT-2 XL	22.2 (±0.5)	21.3 (±0.5)	24.2 (±0.5)
FT	99.6 (±0.1)	82.1 (±0.6)	23.2 (±0.5)
FT+L	92.3 (±0.4)	**47.2 (±0.7)**	23.4 (±0.5)
KE	65.5 (±0.6)	61.4 (±0.6)	24.9 (±0.5)
KE-zsRE	92.4 (±0.3)	90.0 (±0.3)	23.8 (±0.5)
MEND	75.9 (±0.5)	65.3 (±0.6)	24.1 (±0.5)
MEND-zsRE	99.4 (±0.1)	**99.3 (±0.1)**	24.1 (±0.5)
ROME	**99.8 (±0.0)**	88.1 (±0.5)	**24.2 (±0.5)**

Counterfact-Dataset

«Can the editing method change the location of the Gazi University from Ankara to Glasgow?»

```
"case_id":1067,
"pararel_idx":21136,
"requested_rewrite":{
    "prompt":"The headquarters of {} is in",
    "entity":"Gazi University",
    "relation_id":"P159",
    "target_new":{
        "str":"Glasgow",
        "id":"Q4093"
    },
    "target_true":{
        "str":"Ankara",
        "id":"Q3640"
    }
},
```

Counterfact-Dataset


```
"paraphrase_prompts":[
    "The headquarter of Gazi University is located in",
    "Gazi University is headquartered in"
],
"neighborhood_prompts":[
    "The headquarter of TRT Haber is located in",
    "Agricultural Bank is headquartered in",
    "TRT Avaz is based in",
    "AnadoluJet's headquarters are in",
    "The headquarters of National Intelligence Organization is in",
    "The headquarter of MKE Ankaragücü is in",
    "The headquarters of Agricultural Bank is in",
    "The headquarter of Turkish Red Crescent is located in",
    "Turkish Historical Society is headquartered in",
    "Gençlerbirliği S.K. is headquartered in"
```

Counterfact-Dataset Measurements

- **Efficacy**
  - ES: Portion of cases for which P[«false fact»] > P[«correct fact»]
  - EM: P[«false fact»] - P[«correct fact»]
- **Generalization**
  - PS: Portion of cases for which P[«false fact»] > P[«correct fact»]
  - PM: P[«false fact»] - P[«correct fact»]
- **Specificity/Influence on Neighborhood**
  - NS: Portion of cases for which **P[«correct fact»] > P[«false fact»]**
  - NM: P[«correct fact»] - P[«false fact»]

Efficacy, Generalization & Specificity

Editor	Score	Efficacy		Generalization		Specificity	
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑
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)
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)**
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)
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)
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)**
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)**
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)**
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)**
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)**

All other methods have weaknesses!

Editor	Score	Efficacy		Generalization		Specificity	
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑
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)
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)**
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)
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)
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)**
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)**
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)**
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)**
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)**

All other methods have weaknesses!

Editor	Score	Efficacy		Generalization		Specificity	
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑
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)
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)**
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)
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)
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)**
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)**
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)**
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)**
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)**

Limitations

- Scalability issue: only one fact at once
- No Runtime-analysis

- Who has the responsibility ?
- Is Model editing the right way ?

Thank you for your attention!

References

- [1] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems, 35, 2022.
- [2] Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. Fast model editing at scale. In International Conference on Learning Representations, 2021.
- [3] De Cao, N., Aziz, W., and Titov, I. Editing factual knowledge in language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.522
- [4] Zhu, C., Rawat, A. S., Zaheer, M., Bhojanapalli, S., Li, D., Yu, F., and Kumar, S. Modifying memories in transformer models. arXiv preprint arXiv:2012.00363, 2020.