Activation Patching in GPT-2

Steinar Grässel
Course: Mechanistic Interpretability
Faculty: CL Heidelberg

Locating and Editing Factual Associations in GPT

Meng et al. 2022

Two distinct goals

- Understanding LLMs: Where is factual knowledge stored?
- Practical application: How do we edit a fact?

Facts — Where & How?
A quest for knowledge

Thesis

'Factual associations in GPT correspond to a localized computation'

→ The model stores facts — let's find them!

CounterFact — Representing facts

Knowledge tuple

$t = (s, r, o)$

Prompt *p*

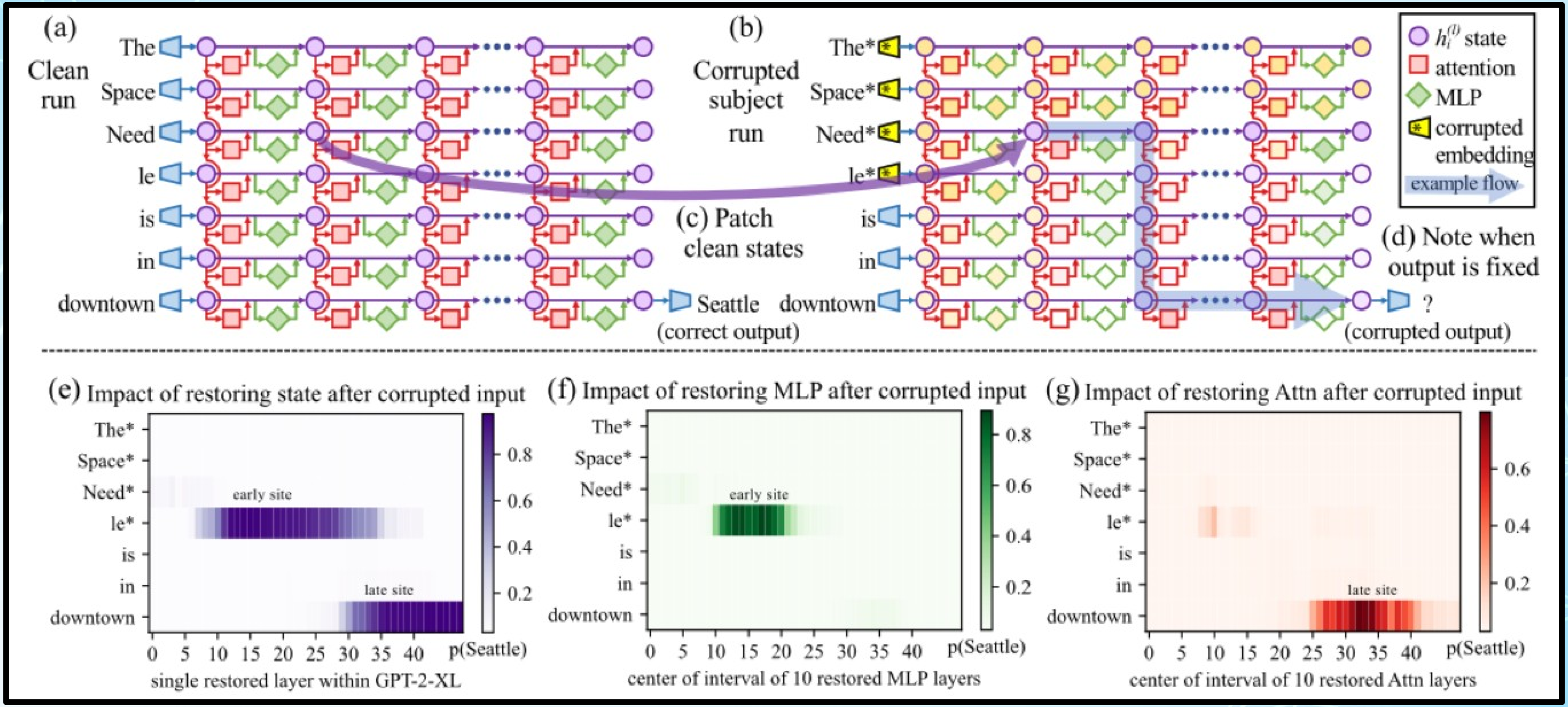
LeBron James plays the sport of

Correct answer

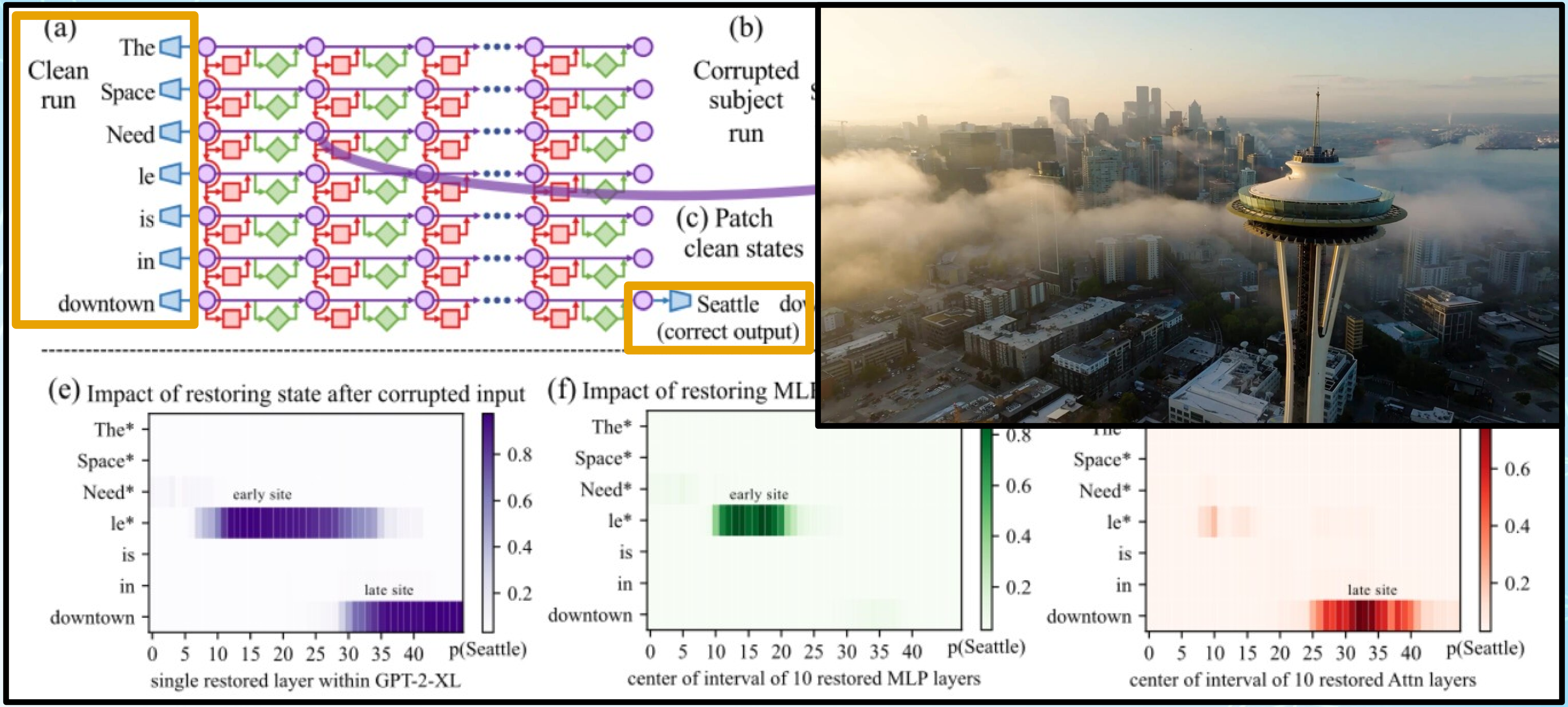
basketball

* r = plays sport professionally

Path Tracing

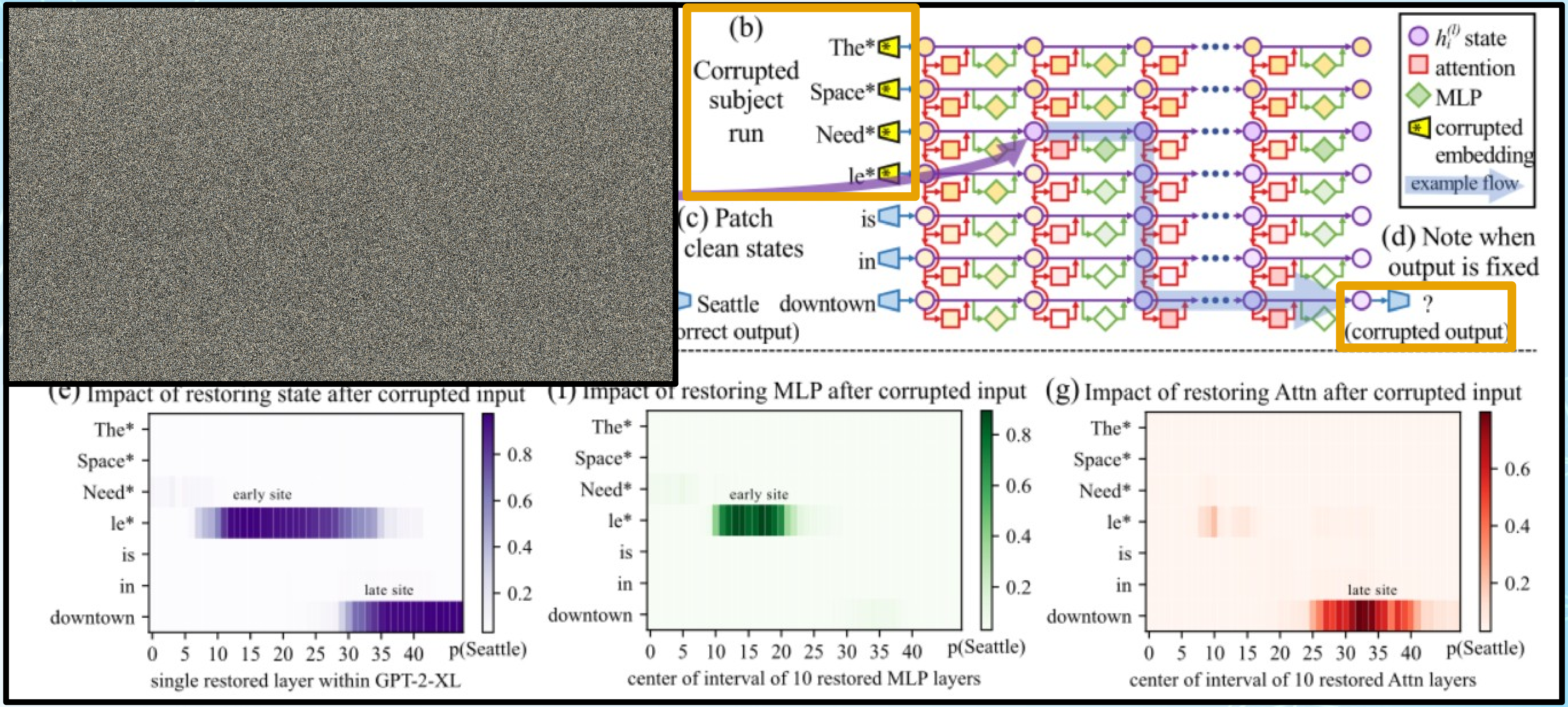


Clean Run

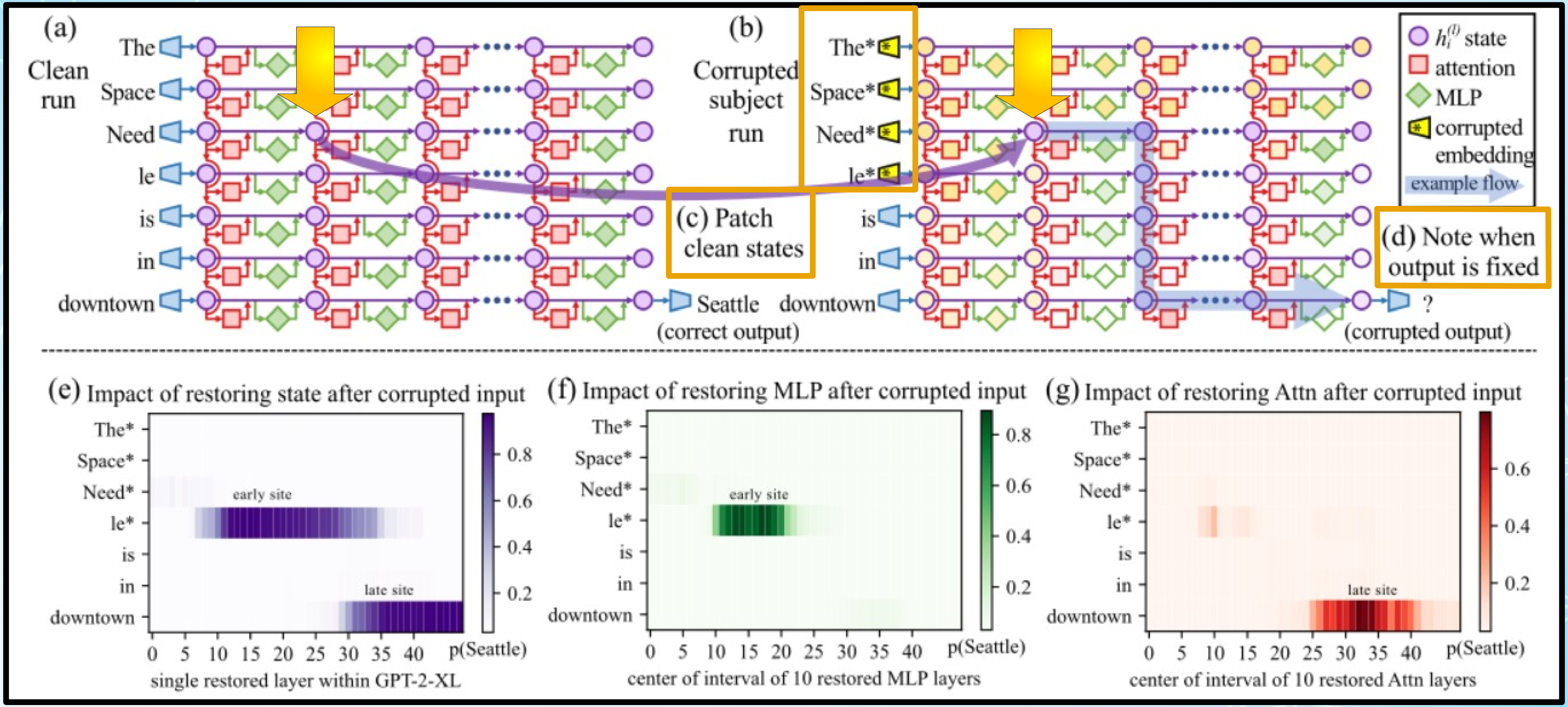


(a)
Clean run
The
Space
Need
le
is
in
downtown
(b)
Corrupted subject run
(c) Patch clean states
Seattle
(correct output)
(e) Impact of restoring state after corrupted input
The*
Space*
Need*
le*
is
in
downtown
early site
late site
0.8
0.6
0.4
0.2
0 5 10 15 20 25 30 35 40
p(Seattle)
single restored layer within GPT-2-XL
(f) Impact of restoring MLP
The*
Space*
Need*
le*
is
in
downtown
early site
0.8
0.6
0.4
0.2
0 5 10 15 20 25 30 35 40
p(Seattle)
center of interval of 10 restored MLP layers
Space*
Need*
le*
is
in
downtown
late site
0.6
0.4
0.2
0 5 10 15 20 25 30 35 40
p(Seattle)
center of interval of 10 restored Attn layers

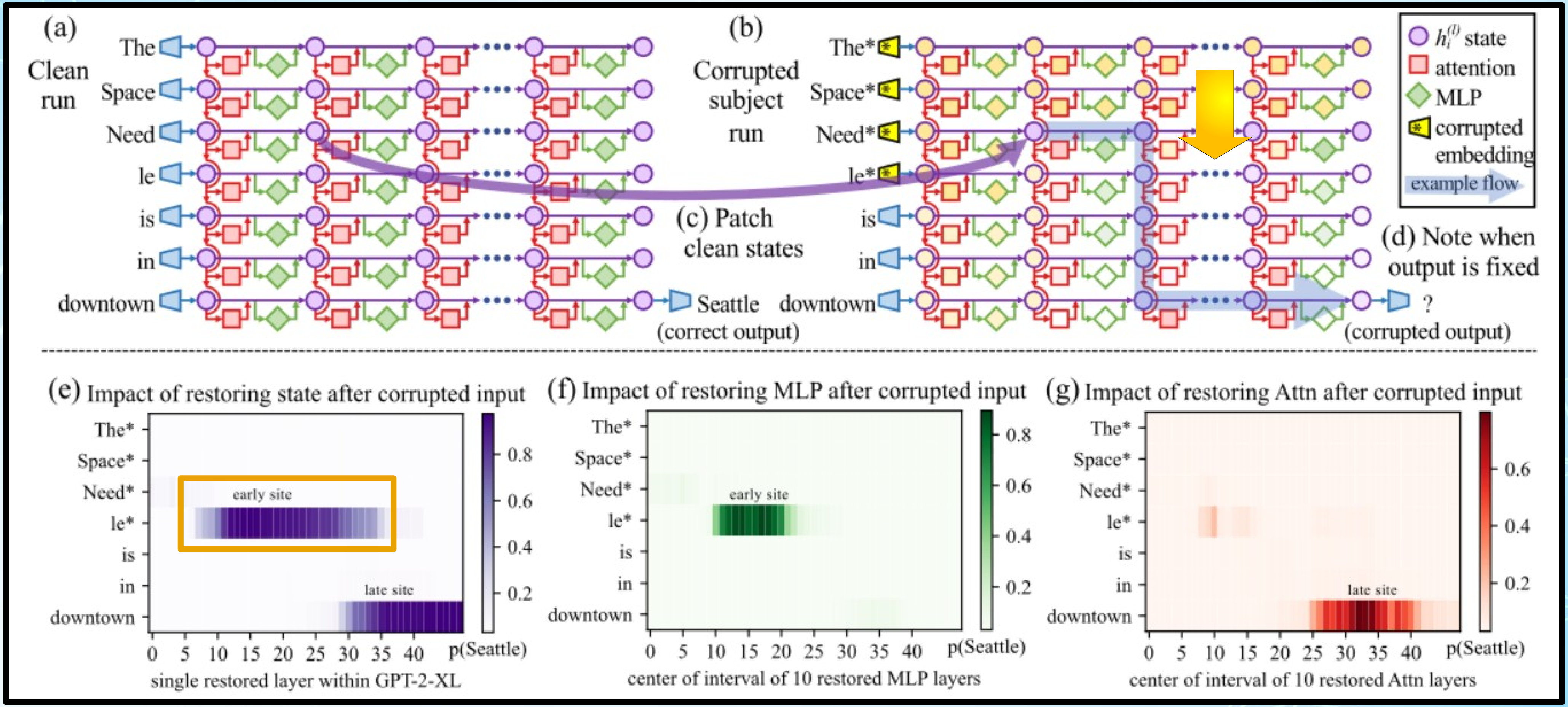
Corrupted Run



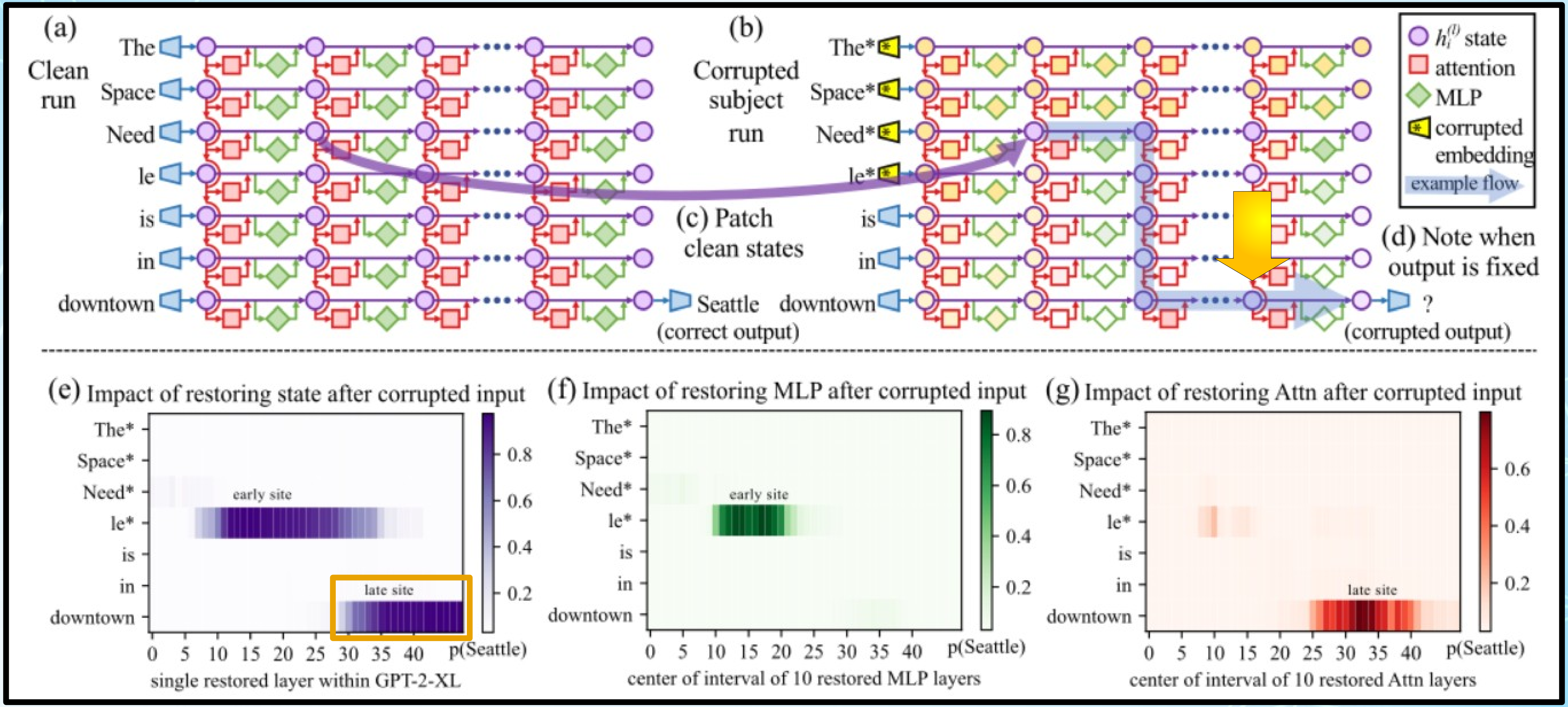
Corrupted w/ restoration



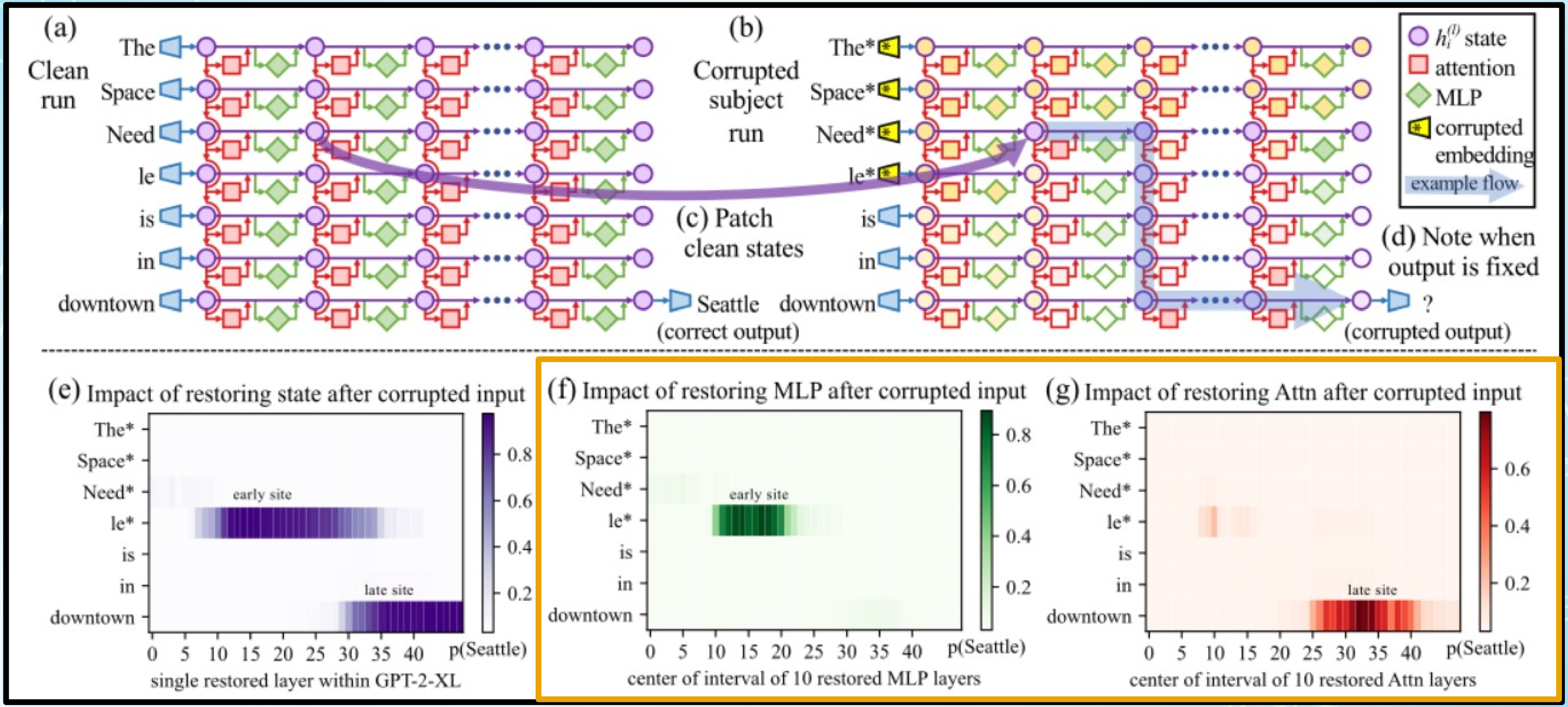
Results — Restoring a hidden state



Results — Restoring a hidden state



Results — Restore **10** MLP / Attn layers



Results — Restore **10** MLP / Attn layers

Figure 7 shows mean causal traces as line plots with 95% confidence intervals, instead of heatmaps.

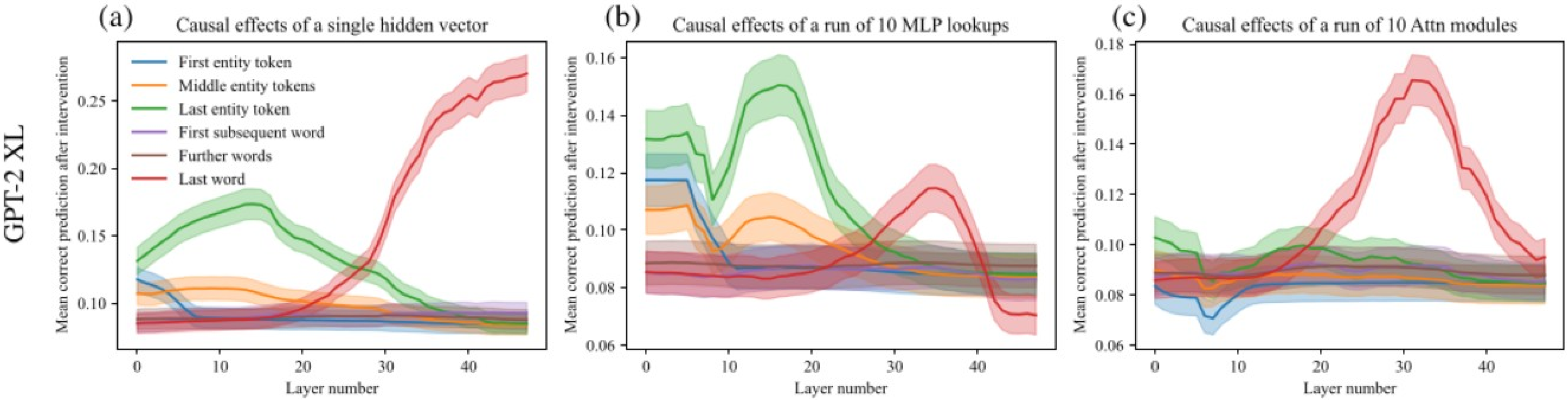


Figure 7: Mean causal traces of GPT-XL over a sample of 1000 factual statements, shown as a line plot with 95% confidence intervals. (a) Shows the same data as Figure 1j as a line plot instead of a heatmap; (b) matches Figure 1k; (c) matches Figure 1m. The confidence intervals confirm that the distinctions between peak and non-peak causal effects at both early and late sites are significant.

Overall results

- Early site (last subject token) → MLPs
- Late site → Attn
- Early site is more surprising

  → further investigation

Sever MLP / Attn

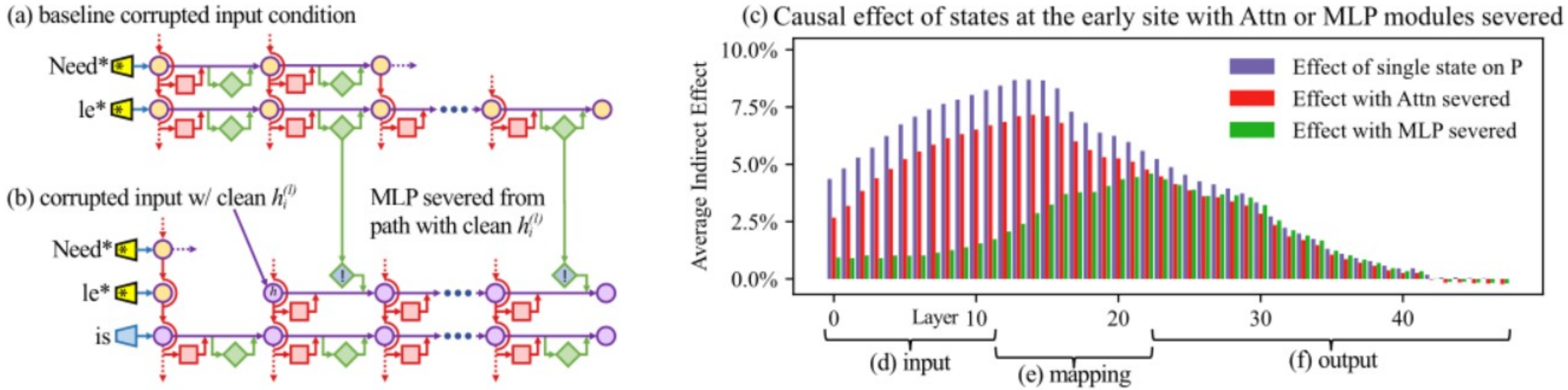


Figure 3: **Causal effects with a modified computation graph**. (a,b) To isolate the effects of MLP modules when measuring causal effects, the computation graph is modified. (c) Comparing Average Indirect Effects with and without severing MLP implicates the computation of (e) midlayer MLP modules in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

Severing MLPs neuters early site causal effects

→ MLPs are essential to recall facts

Locating and Editing Factual Associations in GPT

Meng et al. 2022

Two distinct goals

- ✔ Understanding LLMs: Where is factual knowledge stored?
- ➢ Practical application: How do we edit a fact?

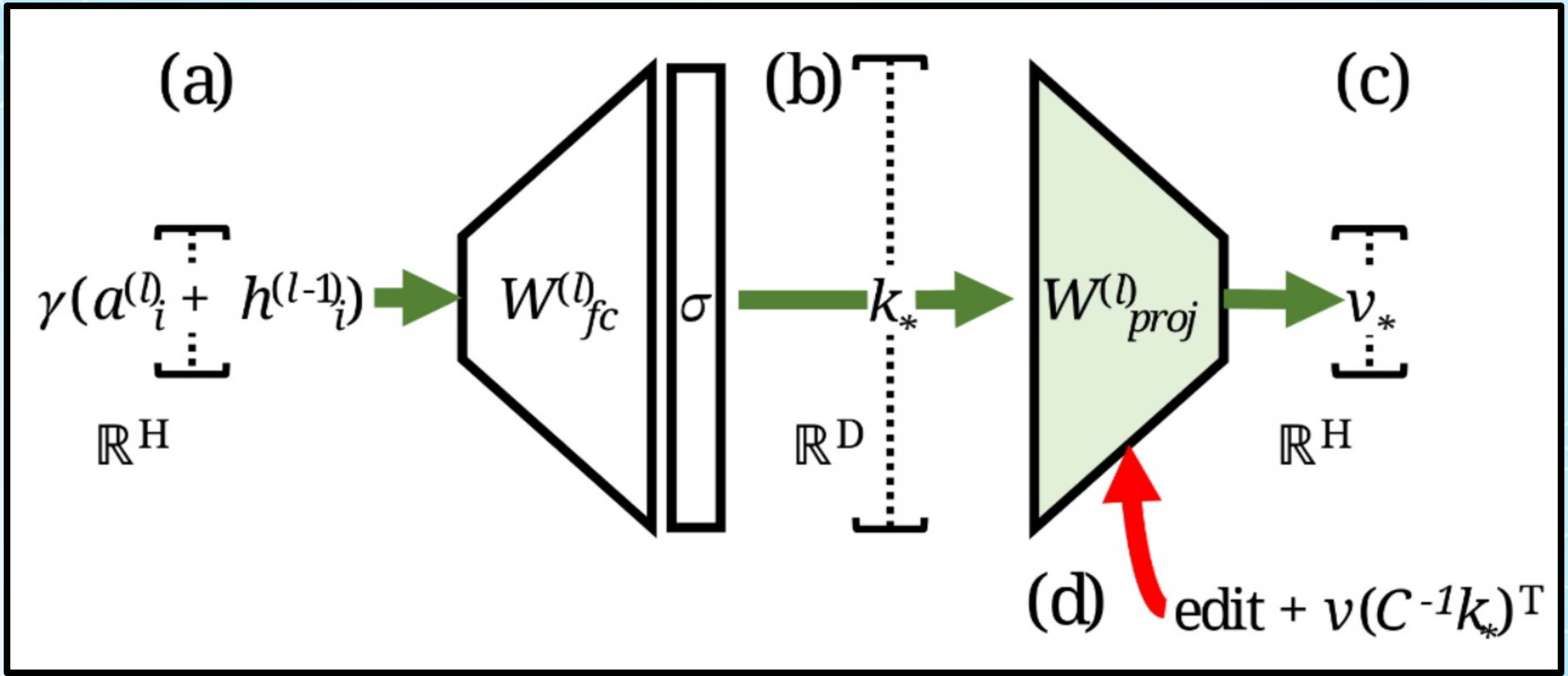
Rank-One Model Editing (ROME)

Assumption

2^{nd} MLP layer ≈ linear associative memory

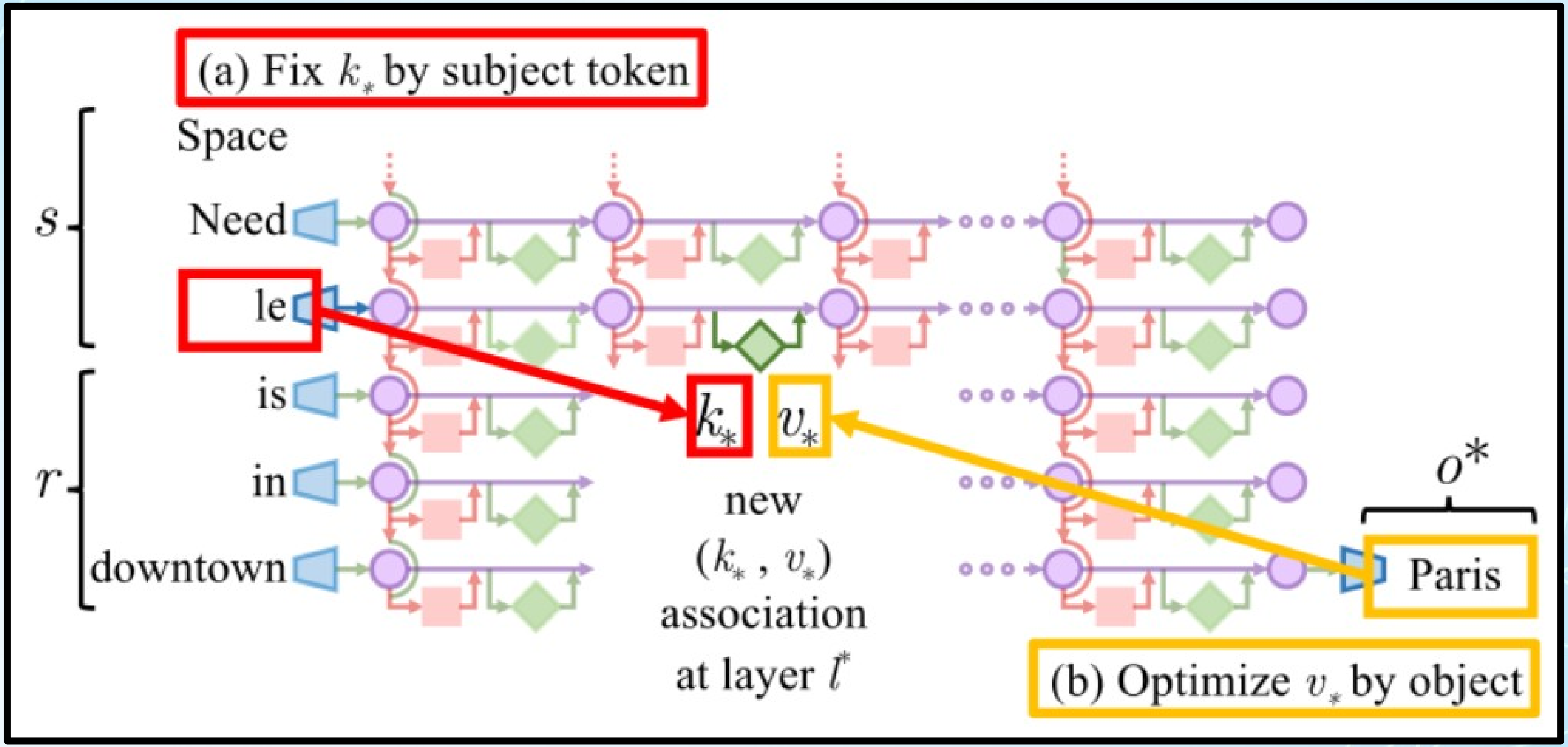
- Key-Value store (K, V)
- $W K \approx V$

Rank-One Model Editing (ROME)



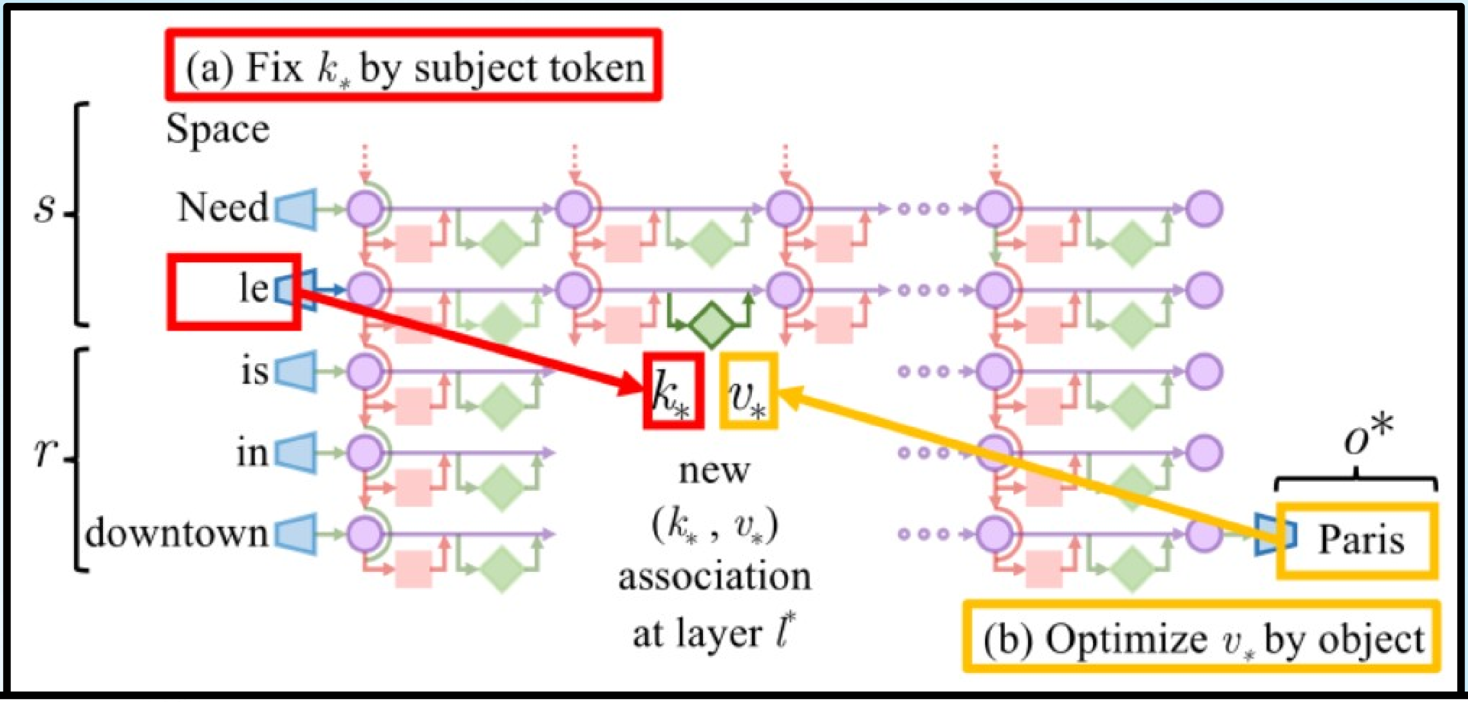
minimize $\|\hat{W}K - V\|$ such that $\hat{W}k_* = v_*$ by setting $\hat{W} = W + \Lambda(C^{-1}k_*)^T$.

Find k_* and v_*



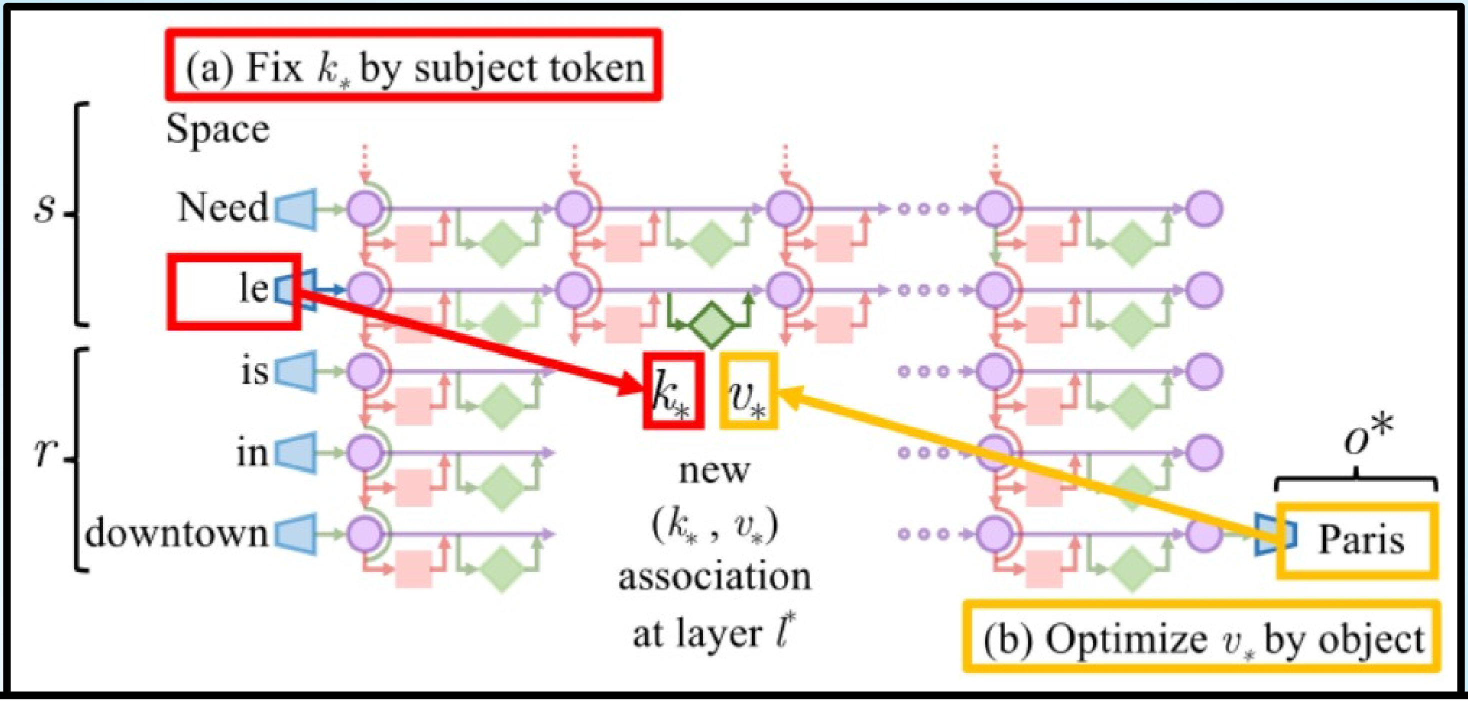
(a) Fix k_* by subject token
Space
Need
le
is
in
downtown
s
r
k_*
v_*
new
(k_*, v_*)
association
at layer l^*
o^*
Paris
(b) Optimize v_* by object

Find **k_*** and v_*



$$k_* = \frac{1}{N}\sum_{j=1}^{N} k(x_j + s), \text{ where } k(x) = \sigma\left(W_{fc}^{(l^*)}\,\gamma(a_{[x],i}^{(l^*)} + h_{[x],i}^{(l^*-1)})\right)$$

Find k_* and $\underline{v_*}$



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

Testing on CounterFact

- Based on ParaRel → WikiData
- Paraphrase prompts → generalization
- Neighborhood prompts → specificity
  - *The Eiffel Tower is in Paris*
  - *The Louvre is in Paris*
- Generation prompts → deeper generalization

Table 4: **Quantitative Editing Results**. 95% confidence intervals are in parentheses. **Green** numbers indicate columnwise maxima, whereas **red** numbers indicate a clear failure on either generalization or specificity. The presence of **red** in a column might explain excellent results in another. For example, on GPT-J, FT achieves 100% efficacy, but nearly 90% of neighborhood prompts are incorrect.

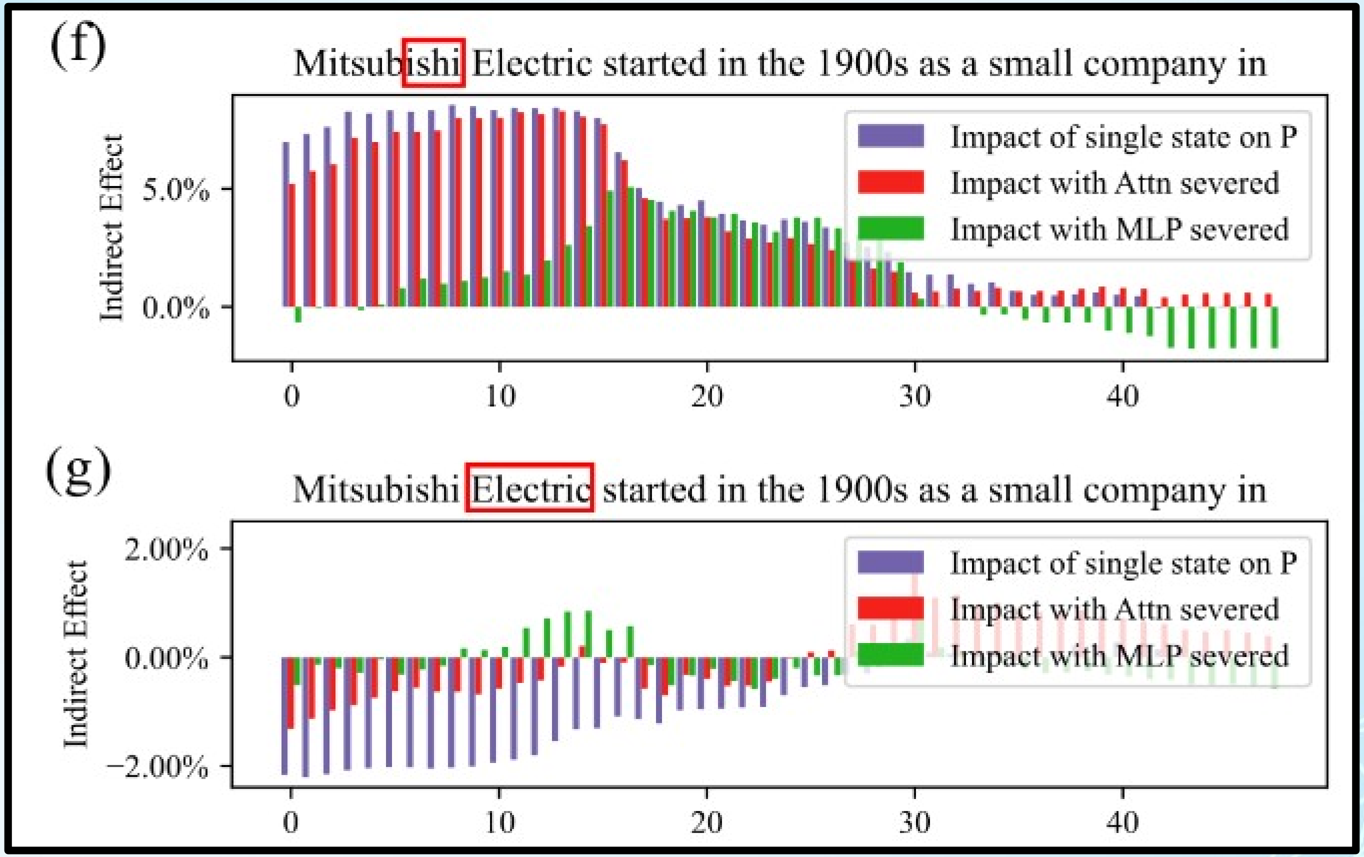
Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	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)**

Limitations / Comments

- Doesn't work when *s* and *o* are reversed
  - *Bill Gates is the founder of Apple*
  - *Apple's founder is Steve Jobs*

→ Could a bidirectional transformer solve this?

Last subject token?



Like their ship or their bodies, their written language has no forward or backward direction. Linguists call this "nonlinear orthography," which raises the question, "Is this how they think?"

Is the transformer like us?

- ✔ Our language, orthography and way of thinking is (mostly) linear → unidirectional
- ✔ Need the whole picture before we can assign facts

    The Eiffel…

    … affair

- ✗ Tokens don’t necessarily match our concepts

Sources

Images

- Space needle: https://www.spaceneedle.com/assets/_1440x810_crop_top-center_75_none/spaceneedle-desktop-posterimage.jpg
- Arrival: https://en.kinorium.com/676817/gallery/screenshot/

Literature

- https://arxiv.org/pdf/2202.05262 (Meng et al. 2022)
- https://aclanthology.org/2023.findings-emnlp.1012.pdf (Pinter and Elhadad 2023)
- https://arxiv.org/pdf/2407.08734 (Miller et al. 2024)
- https://rome.baulab.info/

Interview

- https://www.youtube.com/watch?v=_NMQyOu2HTo&t=2644s&ab_channel=YannicKilcher