Causal Mediation Analysis

CS 221M Week 4, Lecture 8

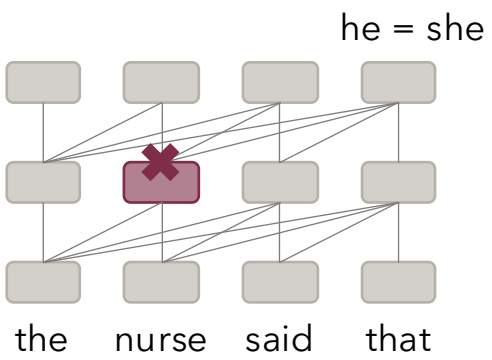
Recap: Causal Interventions

interventions that **causally** change the behavior

(as opposed to probing)

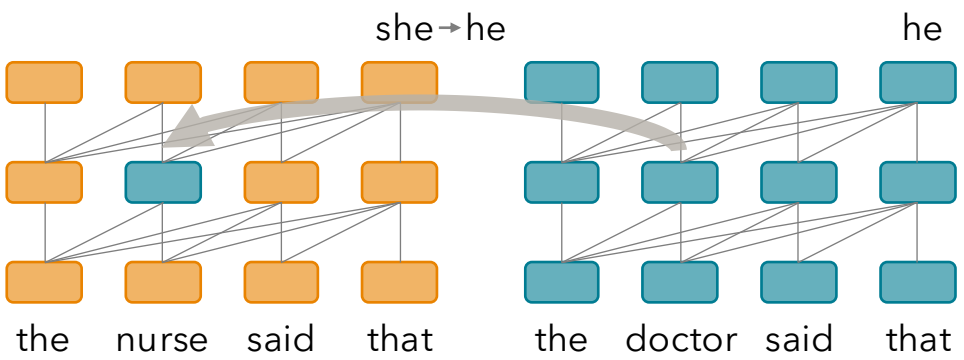
ablation

removing information



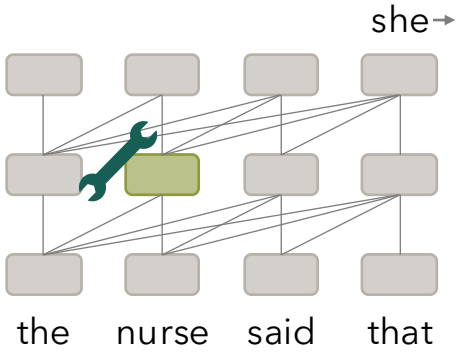
interchange intervention

swapping information



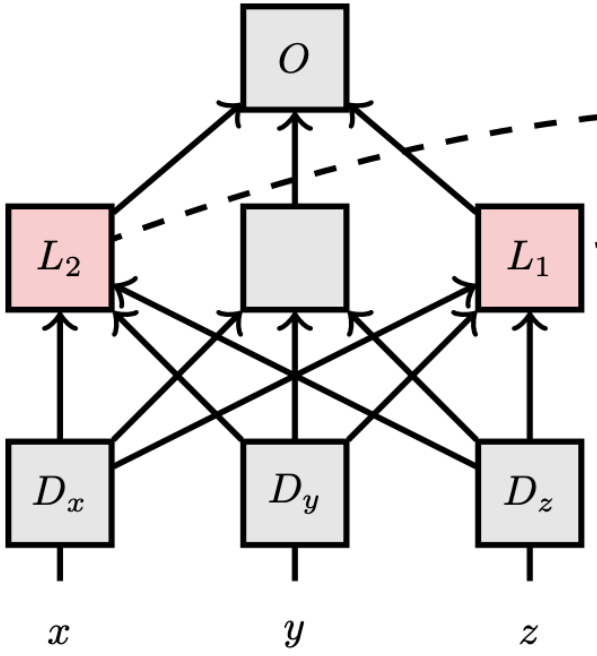
steering

controlling behavior

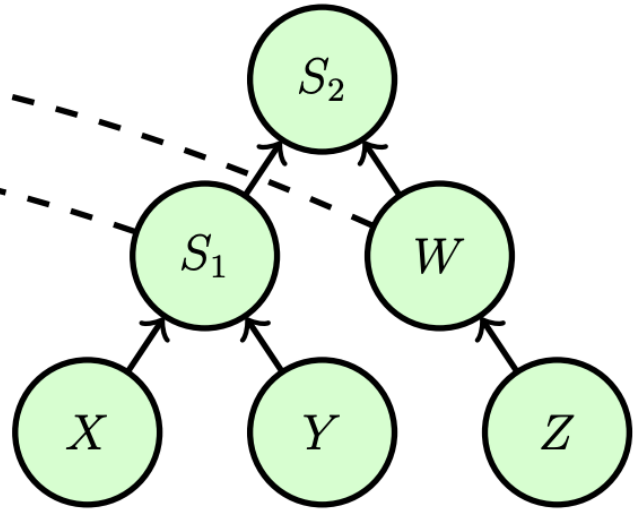


Recap: Causal Abstraction

Neural Network



Symbolic Causal Model



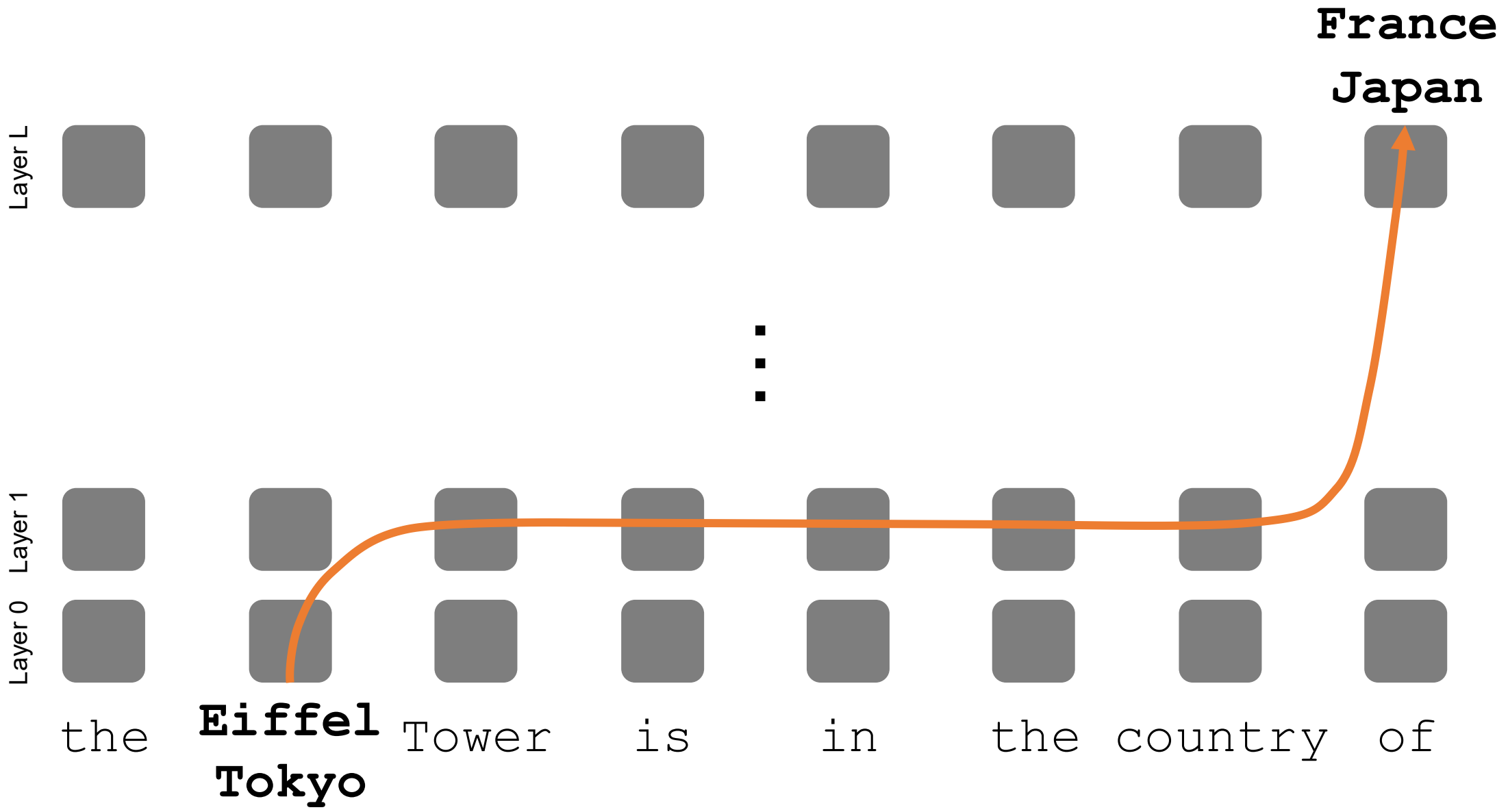
$$S_1 = X + Y$$
$$S_2 = S_1 + W$$

Geiger et al., 2021

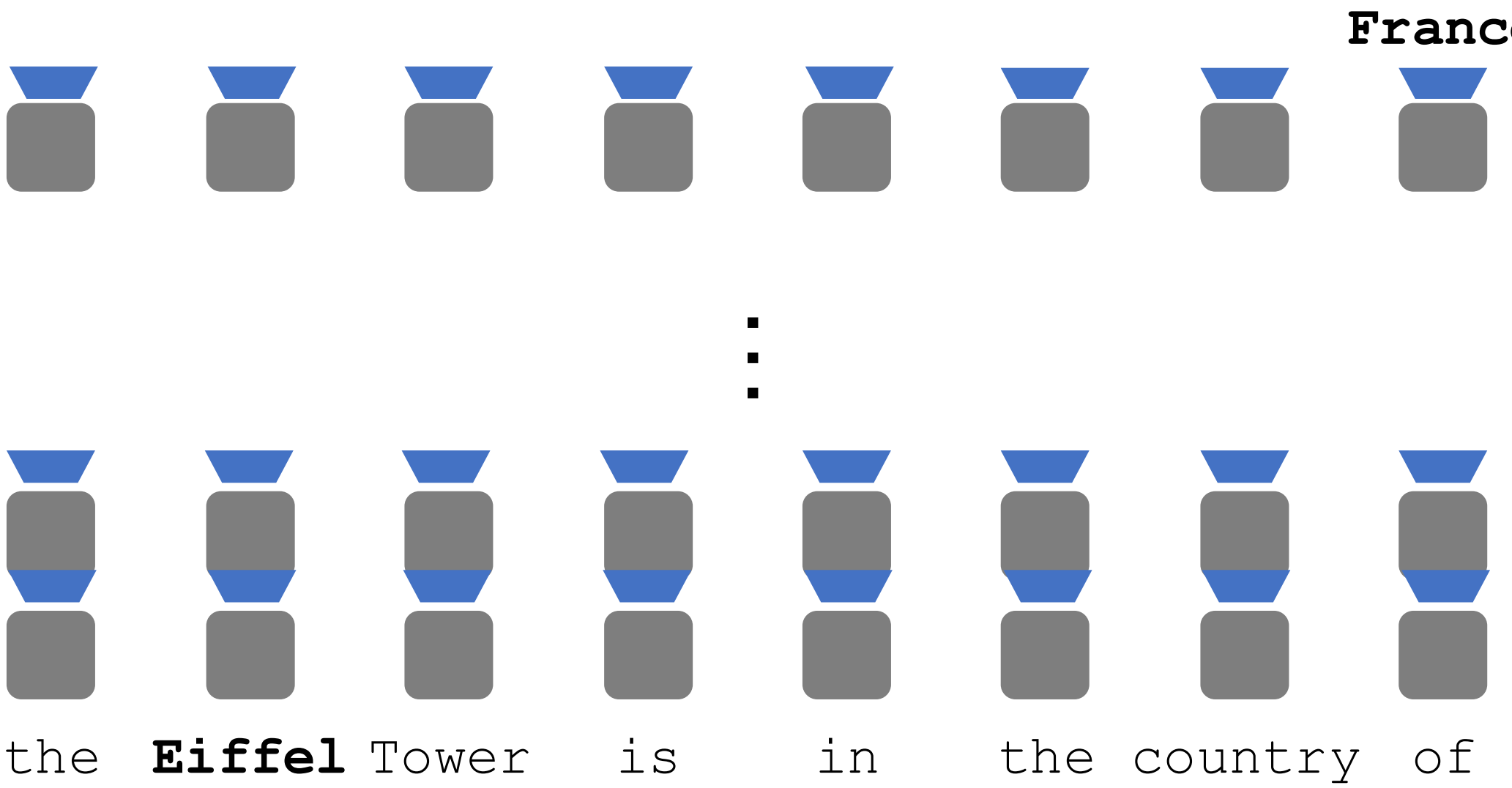
Agenda

- introduce causal mediation analysis
- walk through examples
- build intuitions for types of circuit we find in language models
  - information flow, entity binding, compositions of interventions, etc.

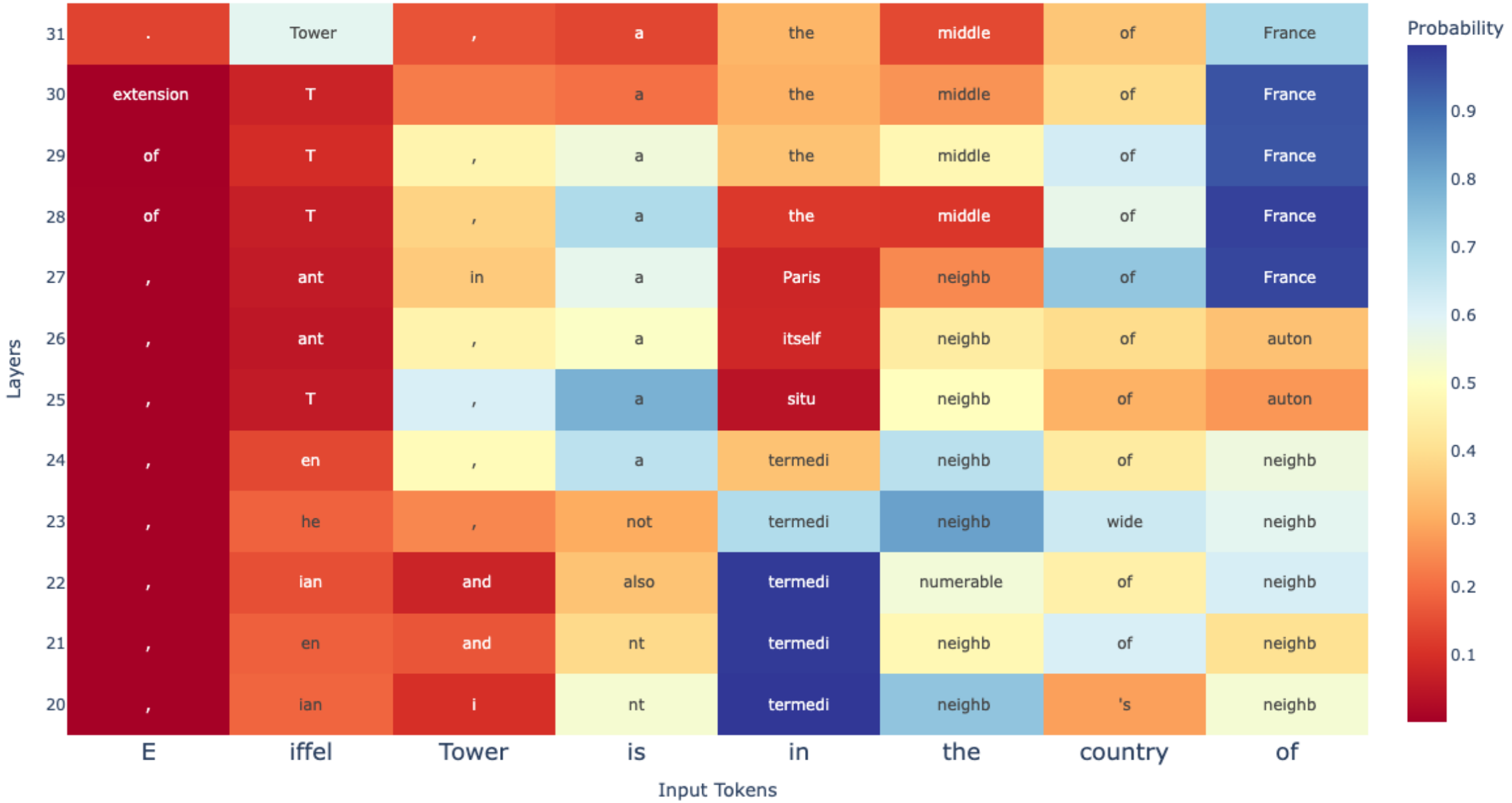
Case Study 1


France
Japan
Layer L
Layer 1
Layer 0
the
Eiffel
Tokyo
Tower
is
in
the
country
of

Revisiting Logit Lens

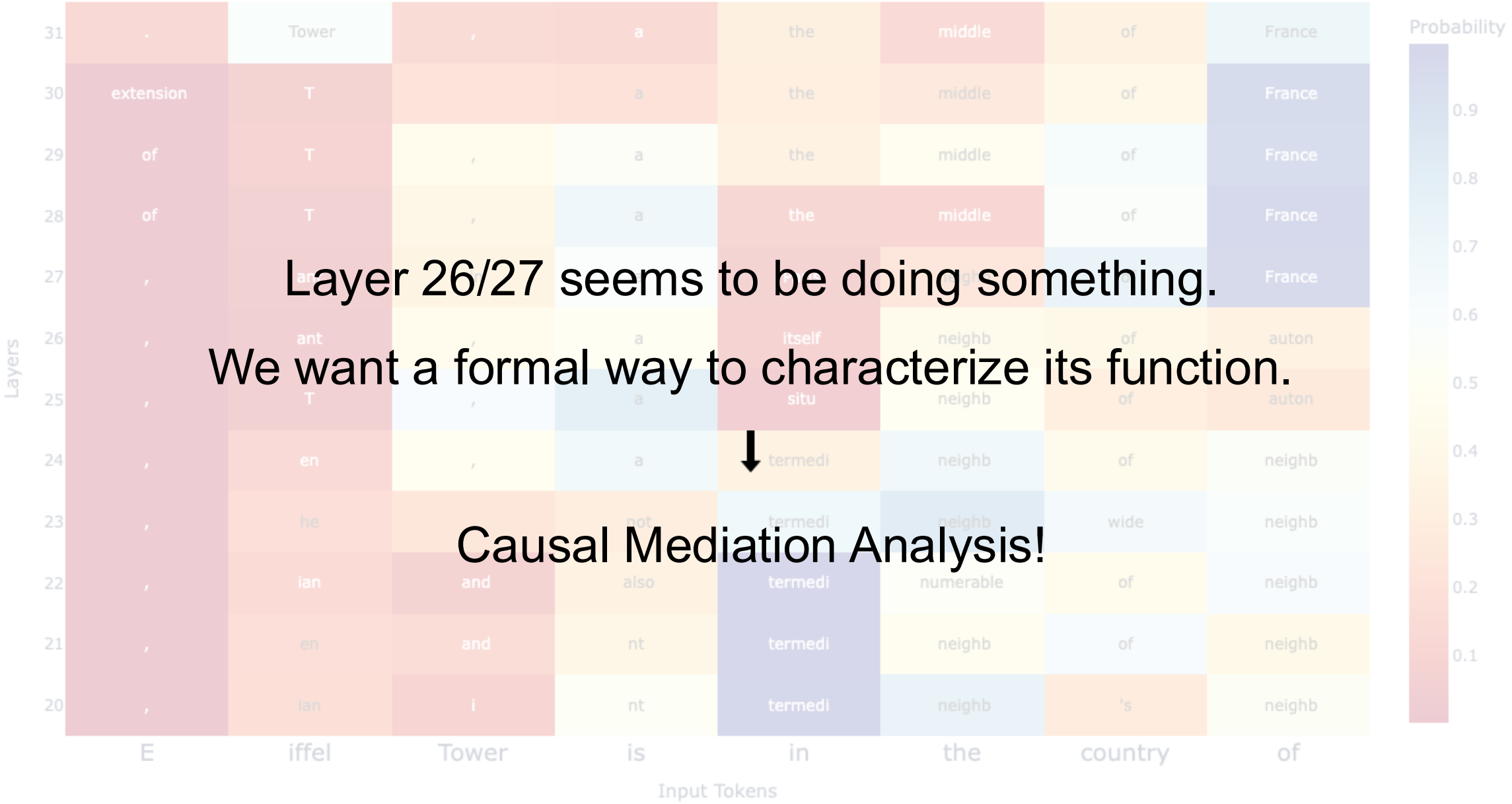


Revisiting Logit Lens

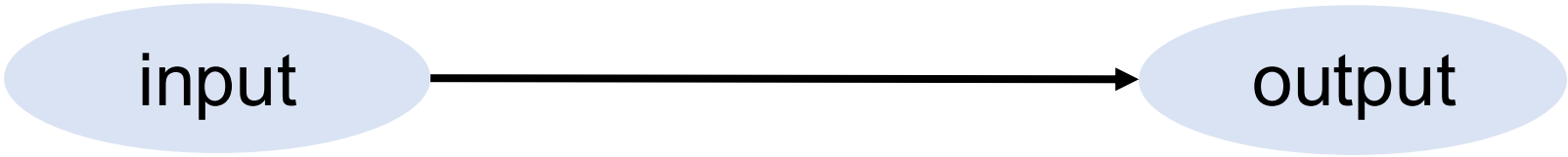


Revisiting Logit Lens

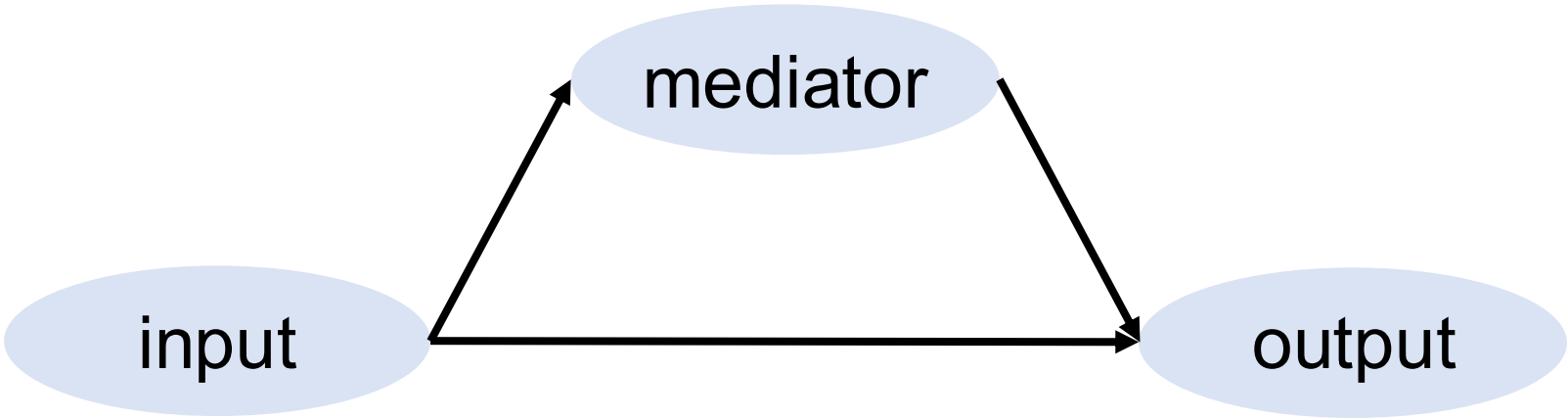
Revisiting Logit Lens



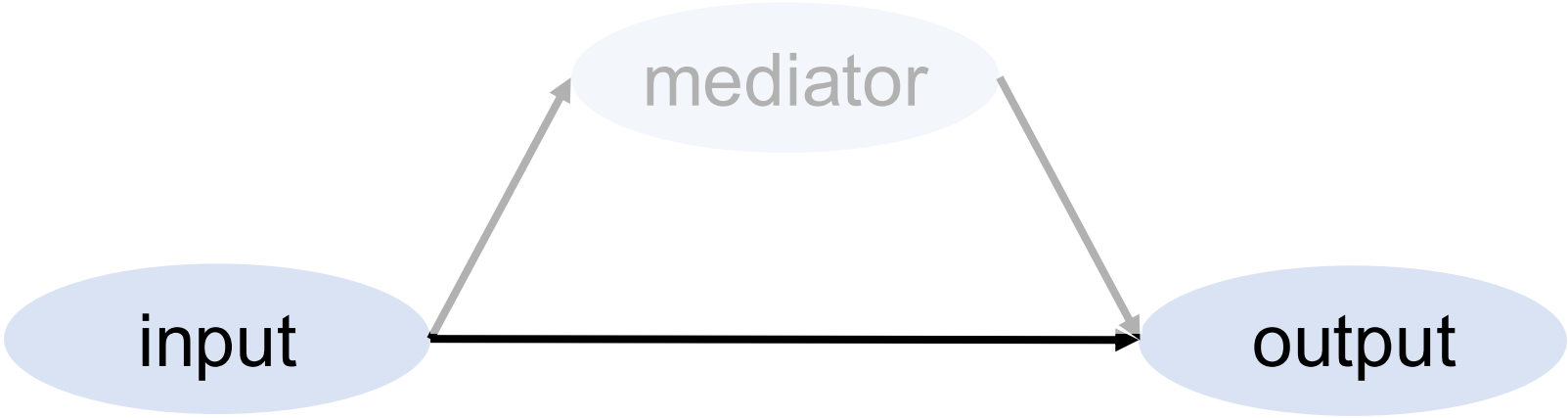
Causal Mediation Analysis



Causal Mediation Analysis



Causal Mediation Analysis



direct effect

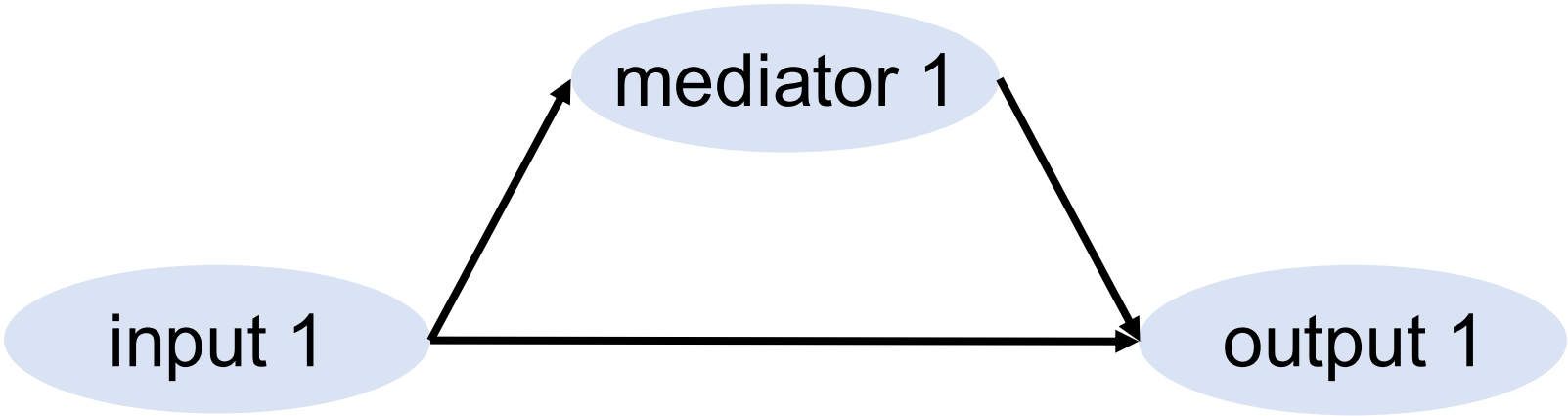
Causal Mediation Analysis

indirect effect

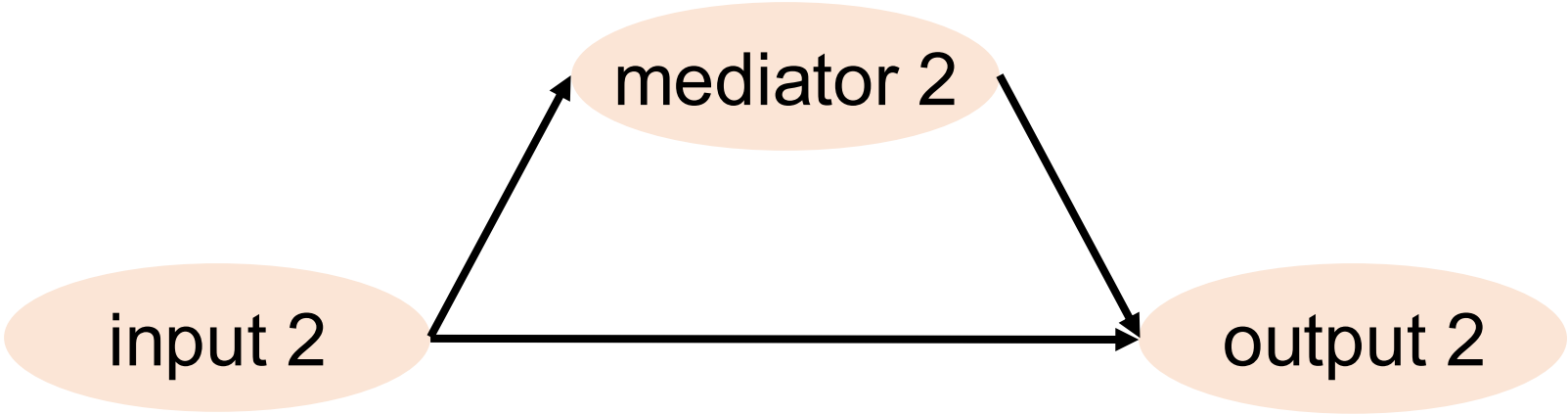
How do we quantify this effect?

Causal Mediation Analysis

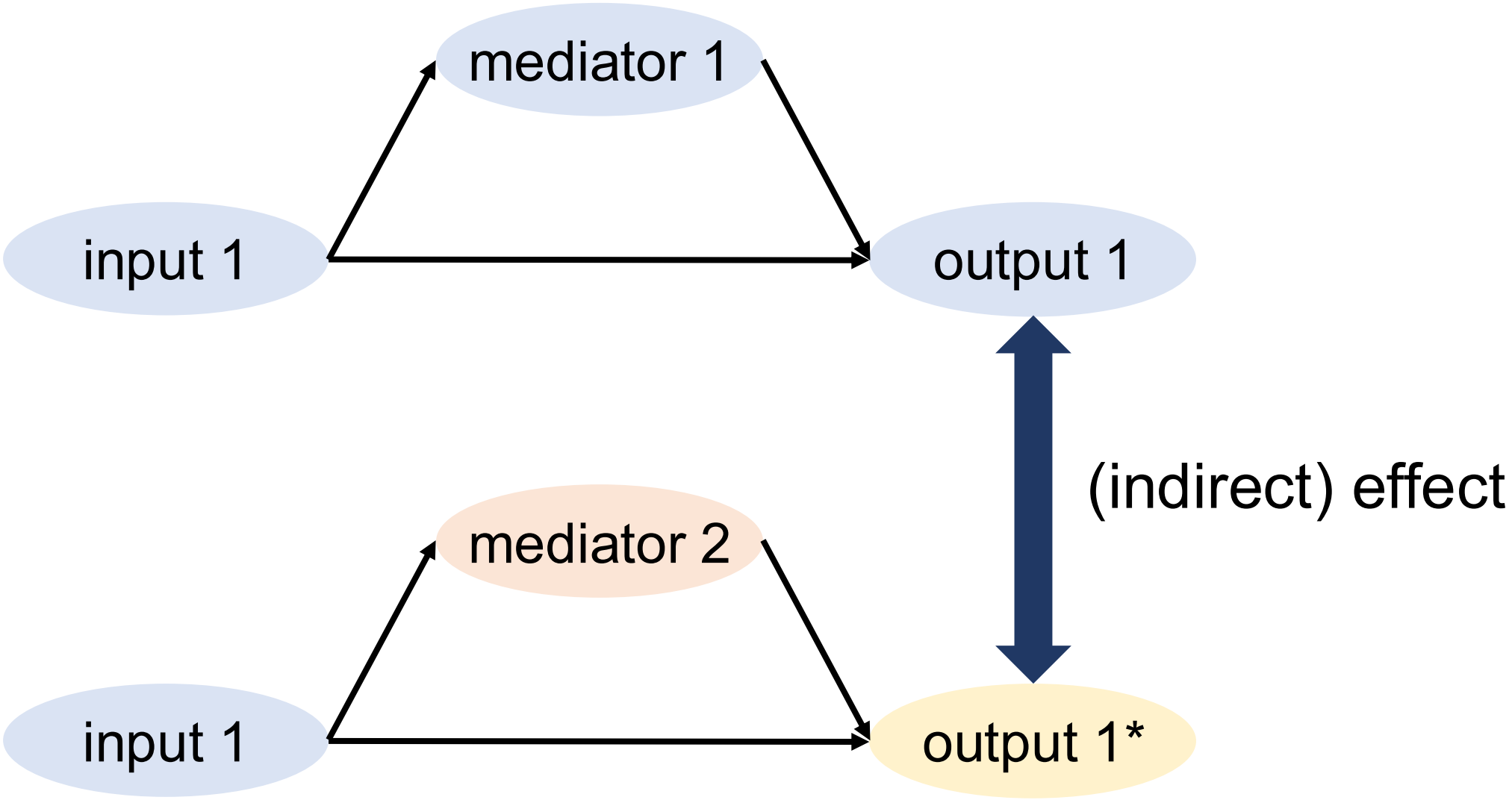
Original 1



Original 2



Causal Mediation Analysis



Going back to our example

(1) The **Eiffel** Tower is in the country of **France**

(2) The **Tokyo** Tower is in the country of **Japan**

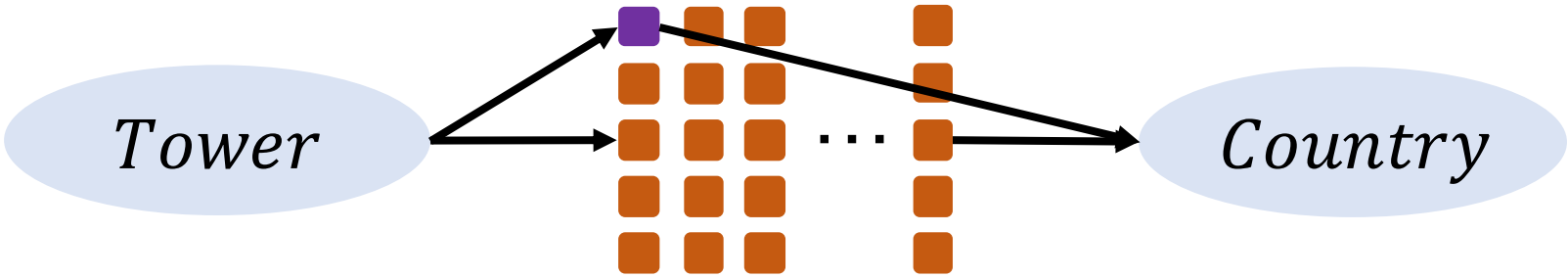
If $Tower$ = “Eiffel”, then $Country$ = “France”.

If $Tower$ = “Tokyo”, then $Country$ = “Japan”.

Going back to our example

(1) The **Eiffel** Tower is in the country of **France**

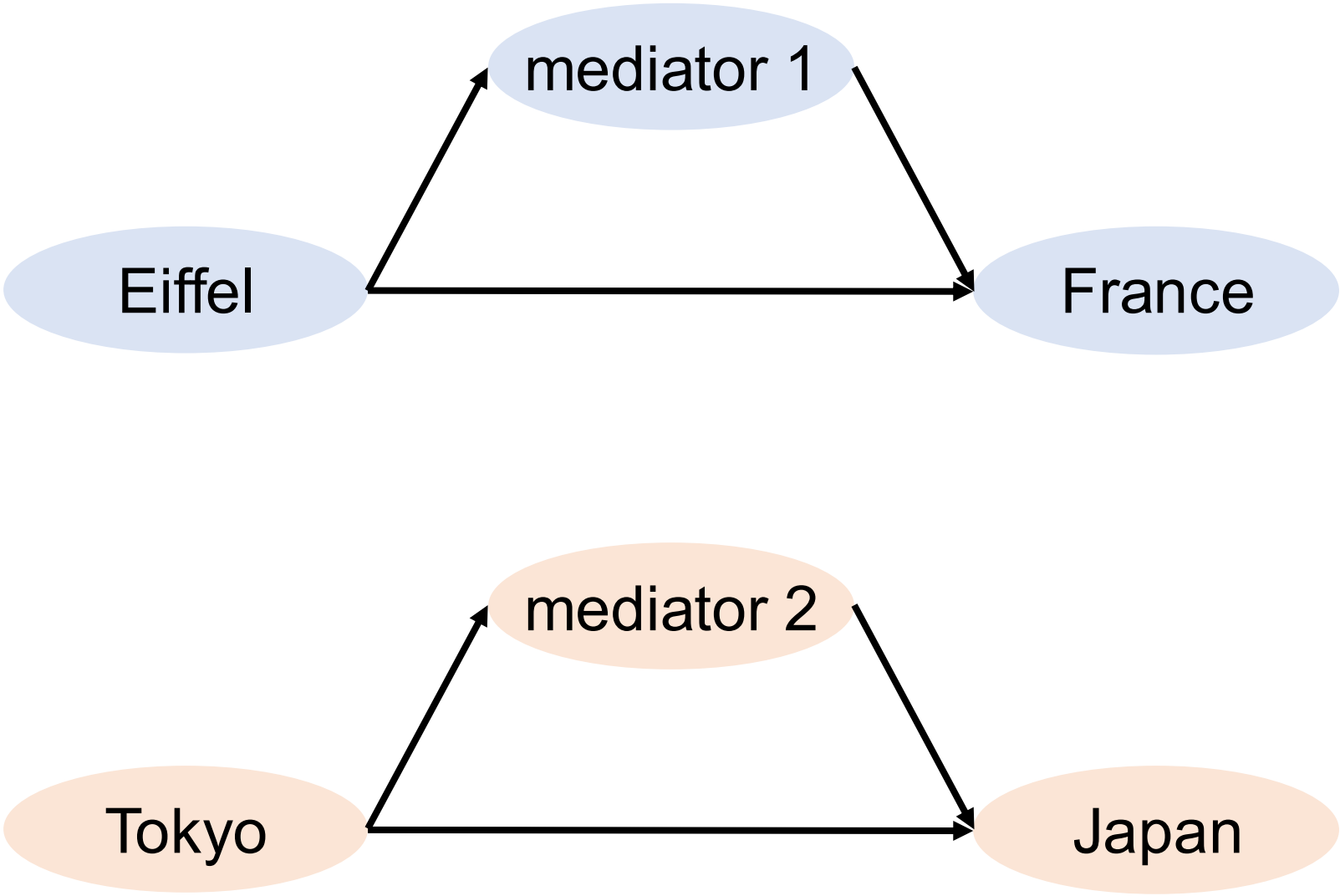
(2) The **Tokyo** Tower is in the country of **Japan**



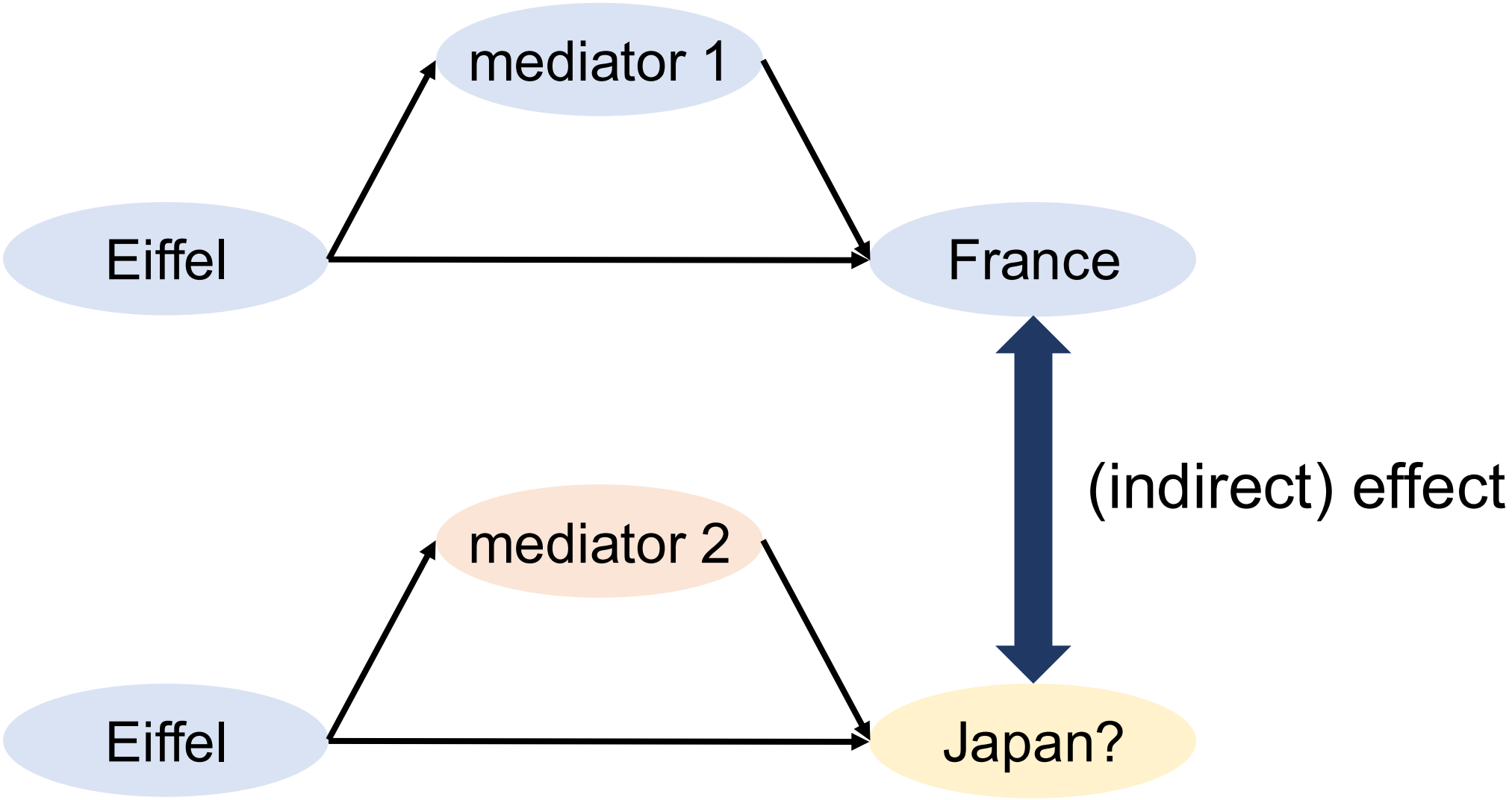
If $Tower$ = “Eiffel”, then $Country$ = “France”.

If $Tower$ = “Tokyo”, then $Country$ = “Japan”.

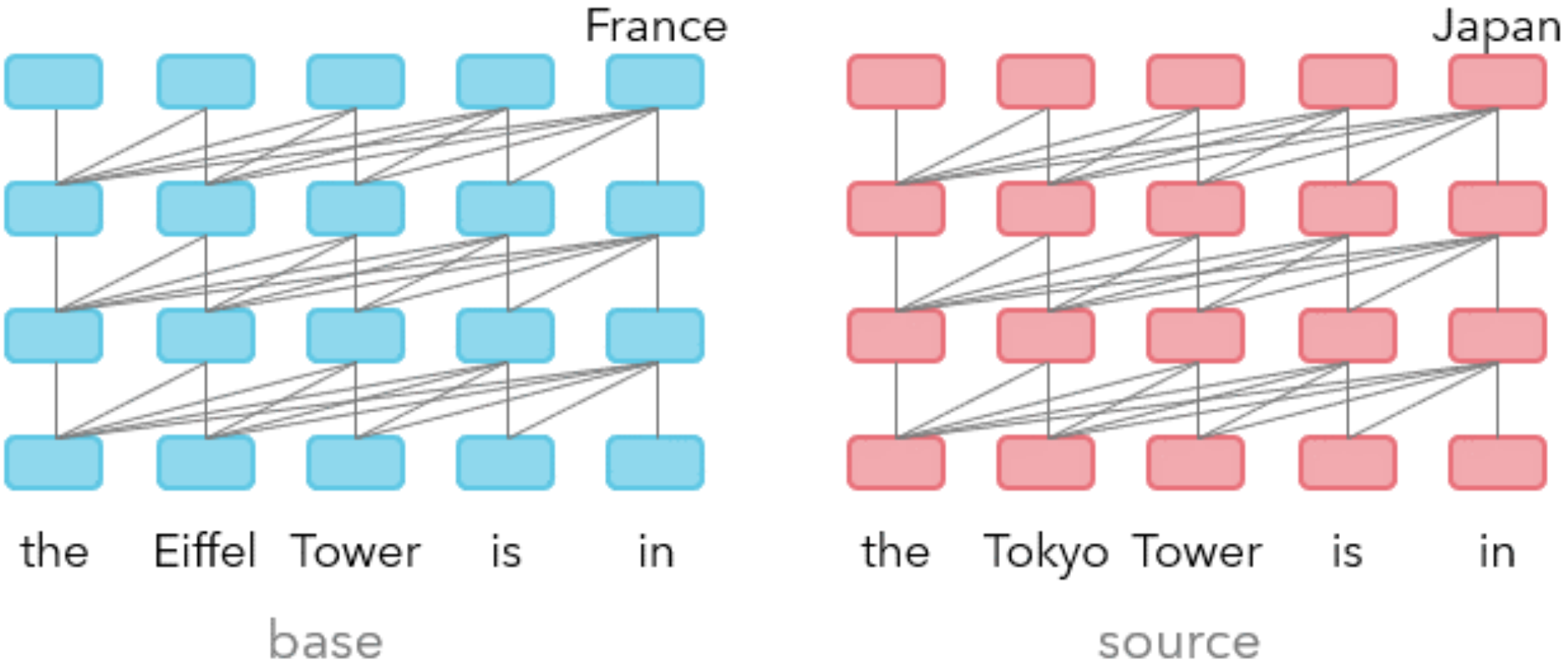
Going back to our example



Causal Mediation Analysis



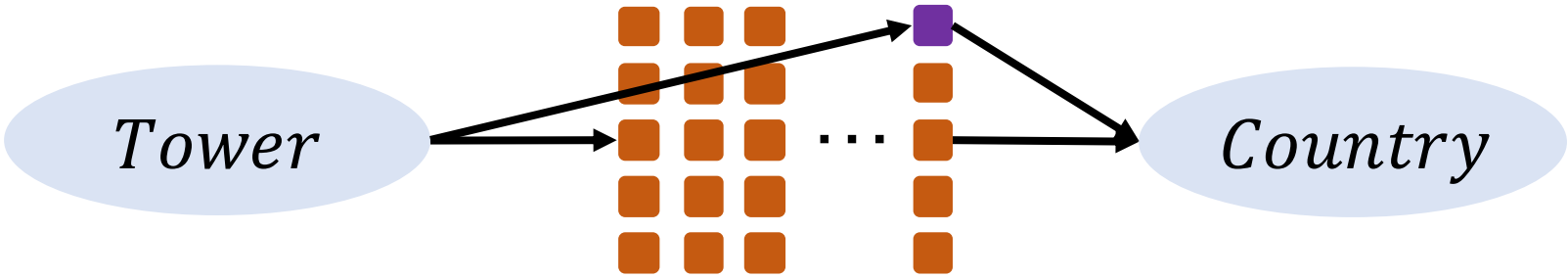
Going back to our example



Going back to our example

(1) The **Eiffel** Tower is in the country of **France**

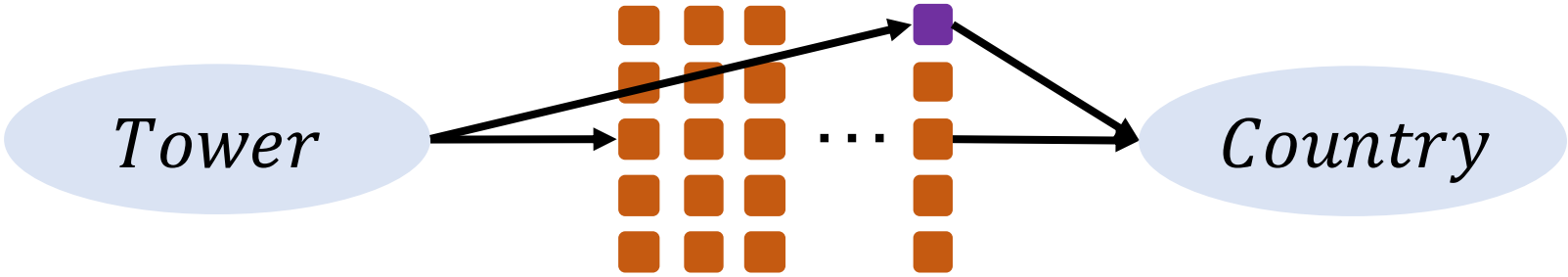
(2) The **Tokyo** Tower is in the country of **Japan**



Going back to our example

(1) The **Eiffel** Tower is in the country of **France**

(2) The **Tokyo** Tower is in the country of **Japan**



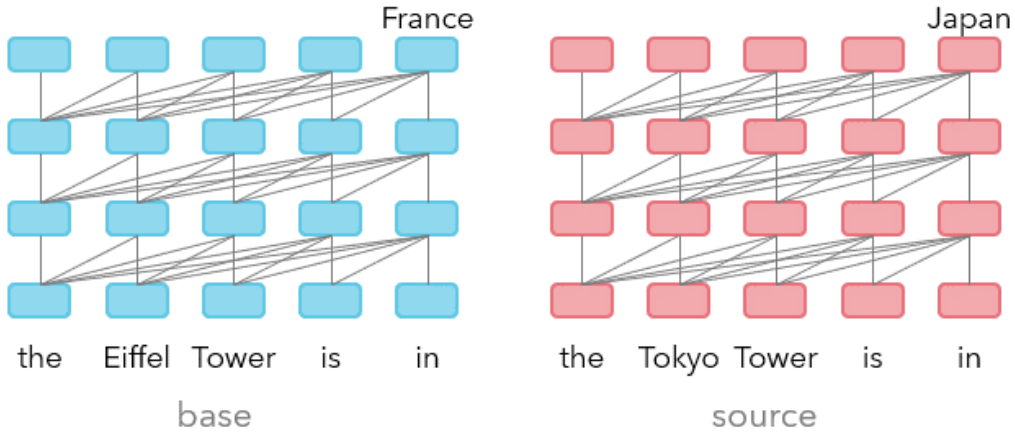
1. Choose a model component as a mediator of interest.
2. Calculate its indirect effect.
3. Repeat for all model components.

Calculating effects

LogOdds(France, Japan | Eiffel Tower)

:= log$P$(France | Eiffel Tower)

- log$P$(Japan | Eiffel Tower)

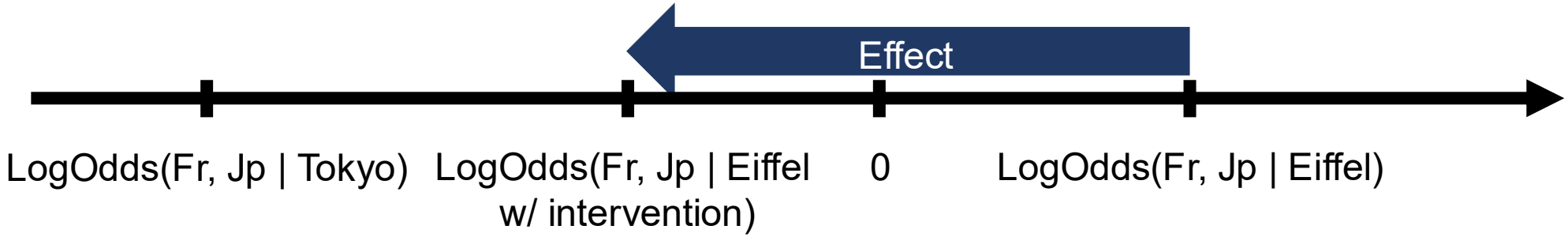


LogOdds(France, Japan | Eiffel Tower) > 0

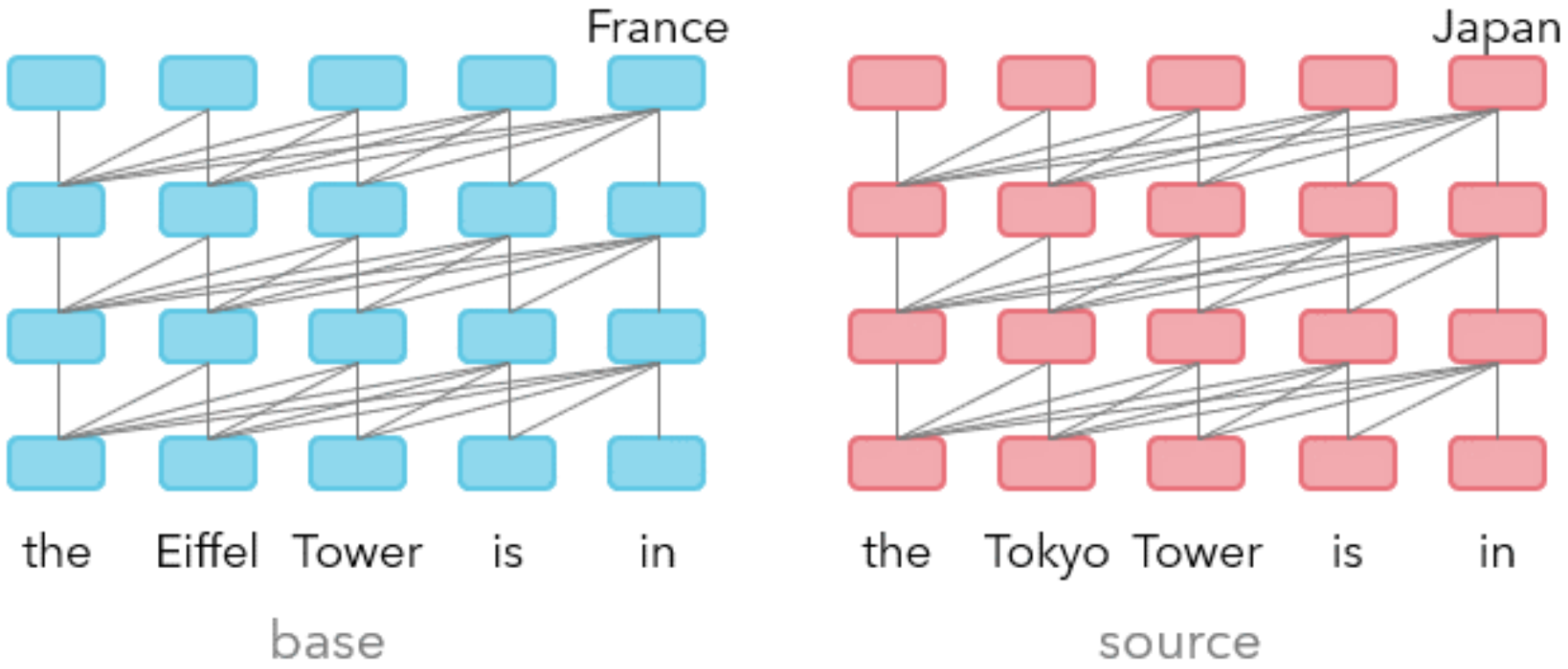
LogOdds(France, Japan | Tokyo Tower) < 0

Effect := LogOdds(France, Japan | Eiffel) –

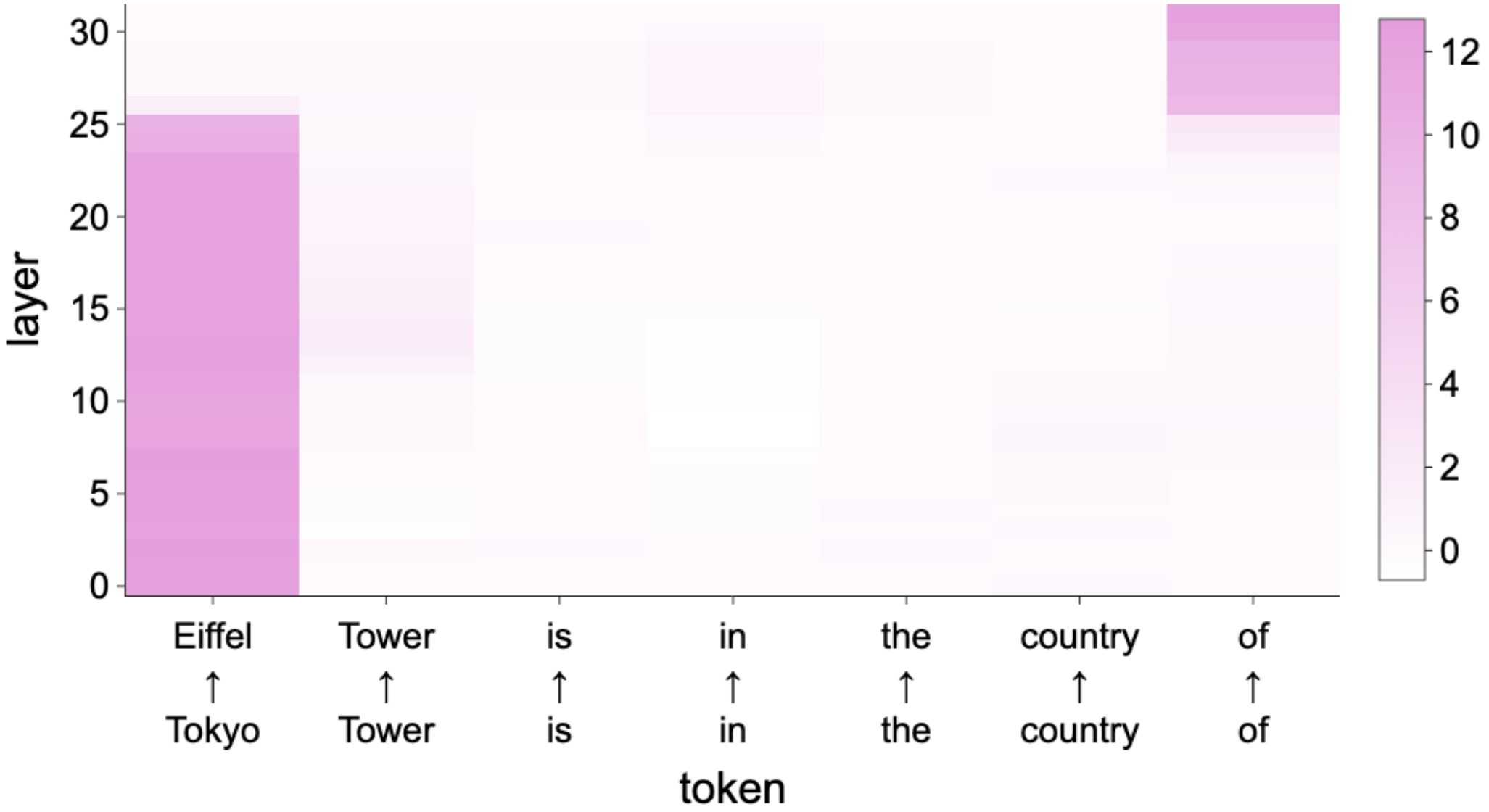
LogOdds(France, Japan | Eiffel w/ intervention)



Summary


France
Japan
the Eiffel Tower is in
the Tokyo Tower is in
base
source
Effect
LogOdds(Fr, Jp | Tokyo)
LogOdds(Fr, Jp | Eiffel w/ intervention)
0
LogOdds(Fr, Jp | Eiffel)

Case Study 1 Result



Exercise 1

(1) The **Eiffel** Tower is in the country of **France**

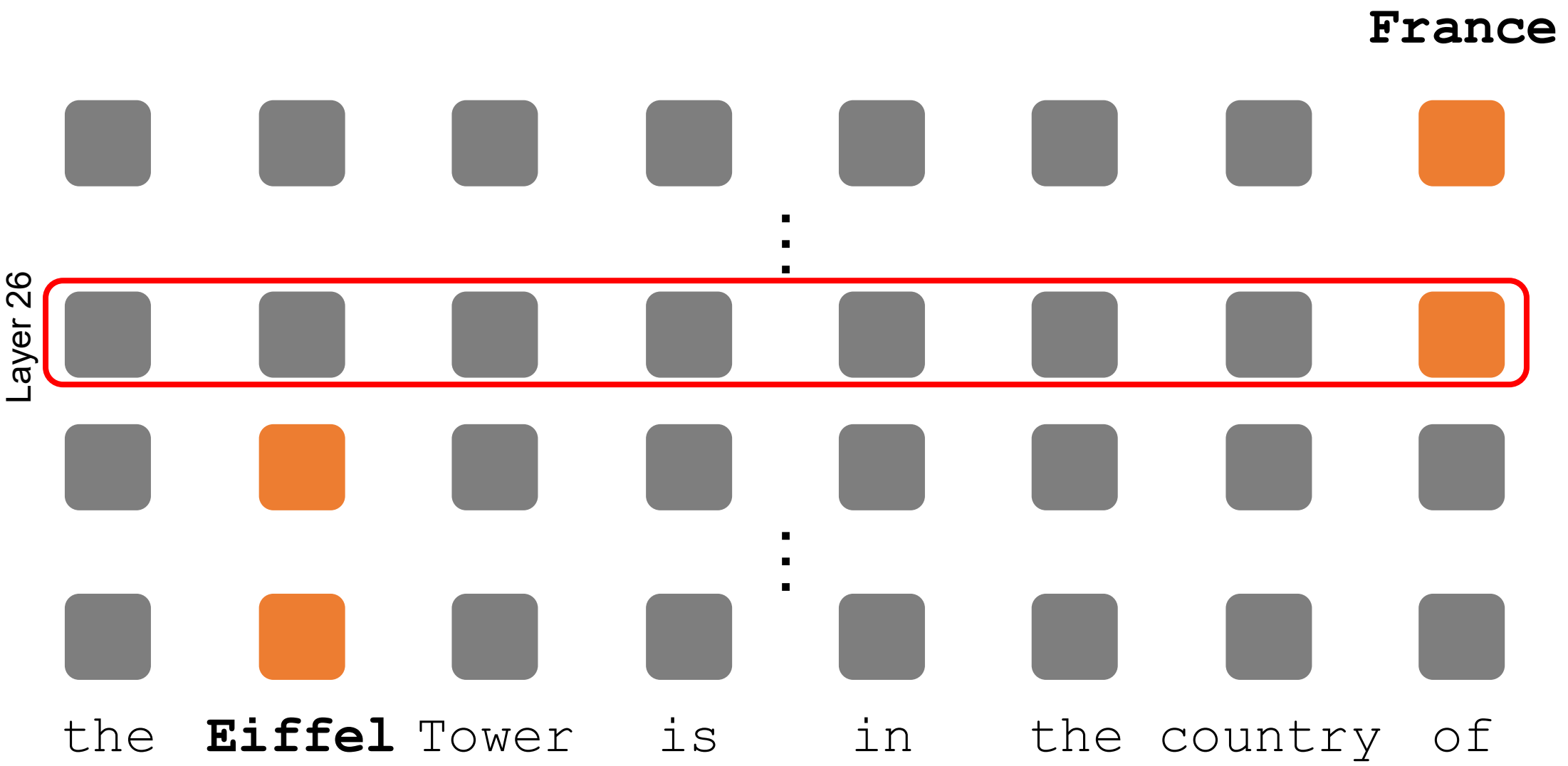
(2) The **Tokyo** Tower is in the country of **Japan**

Exercise 1

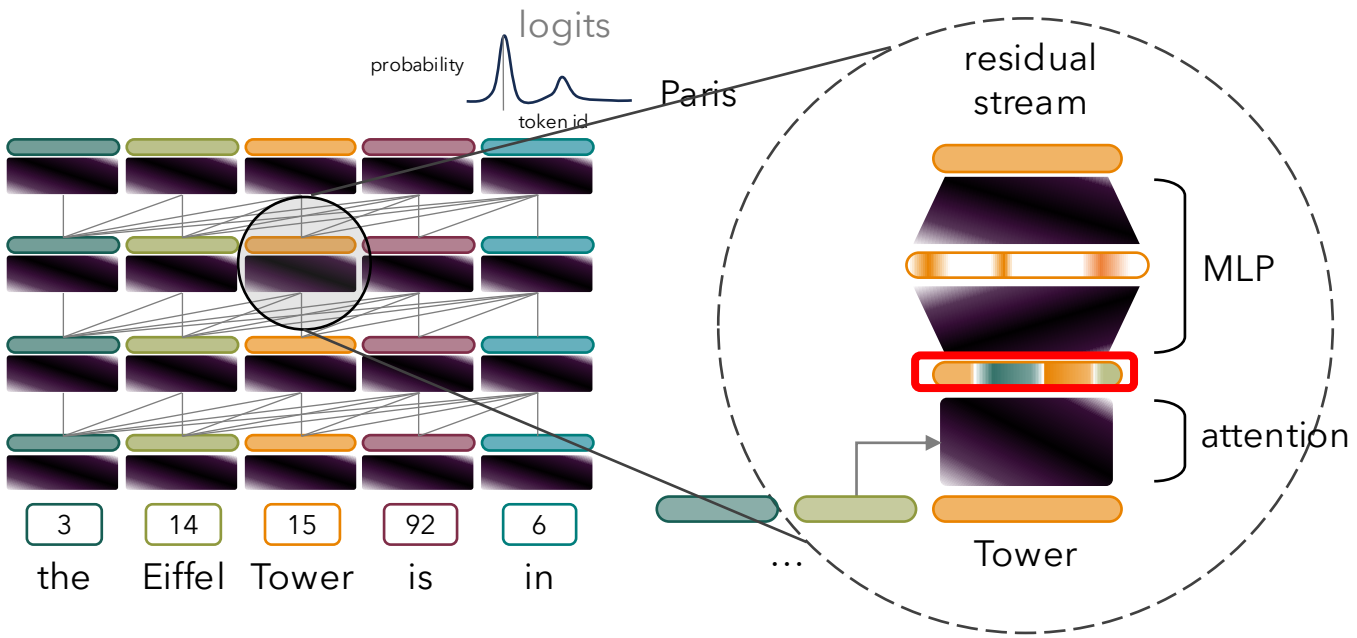
(1) The Eiffel Tower is in the **country** of **France**

city

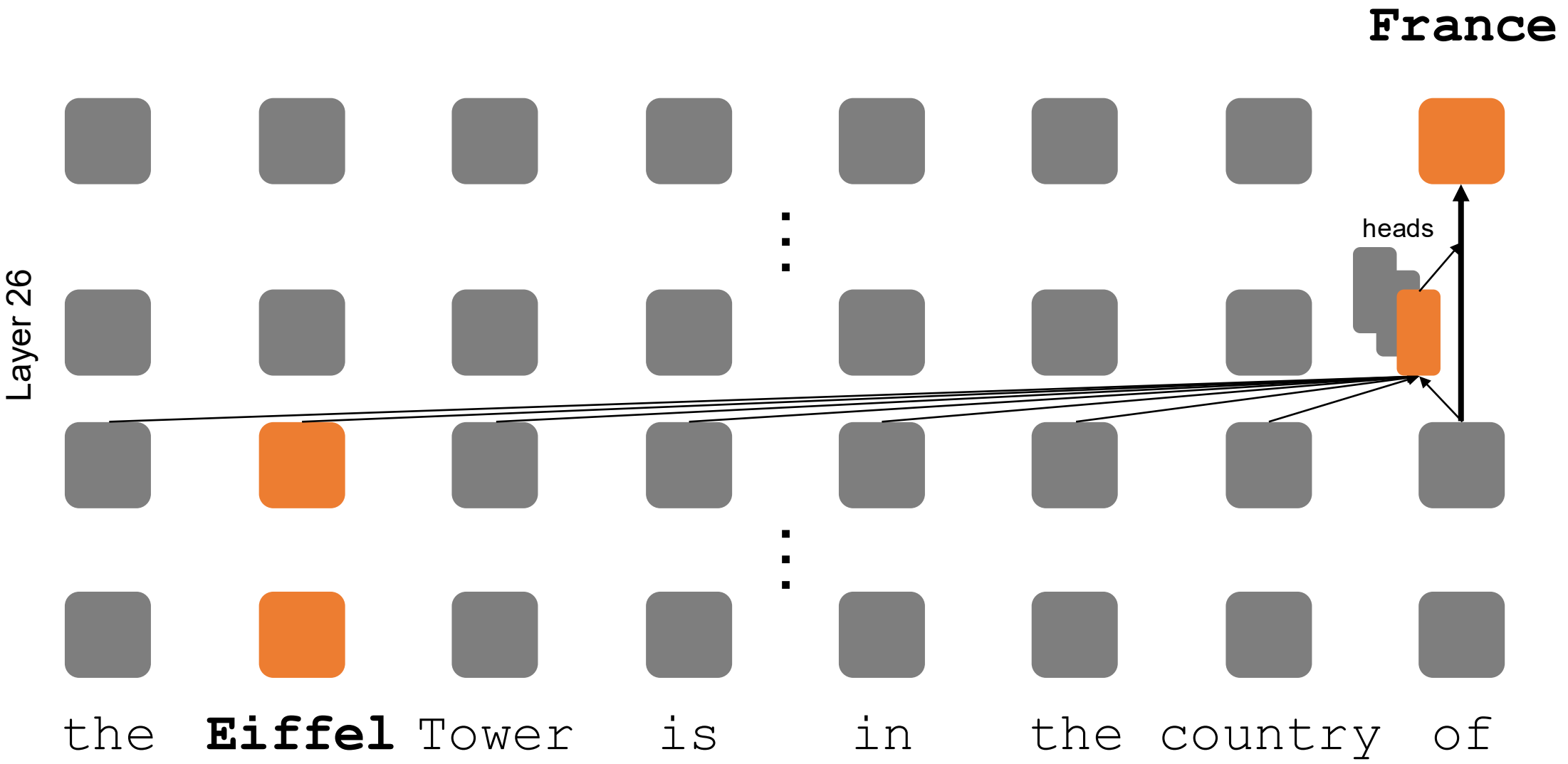
Attention Output



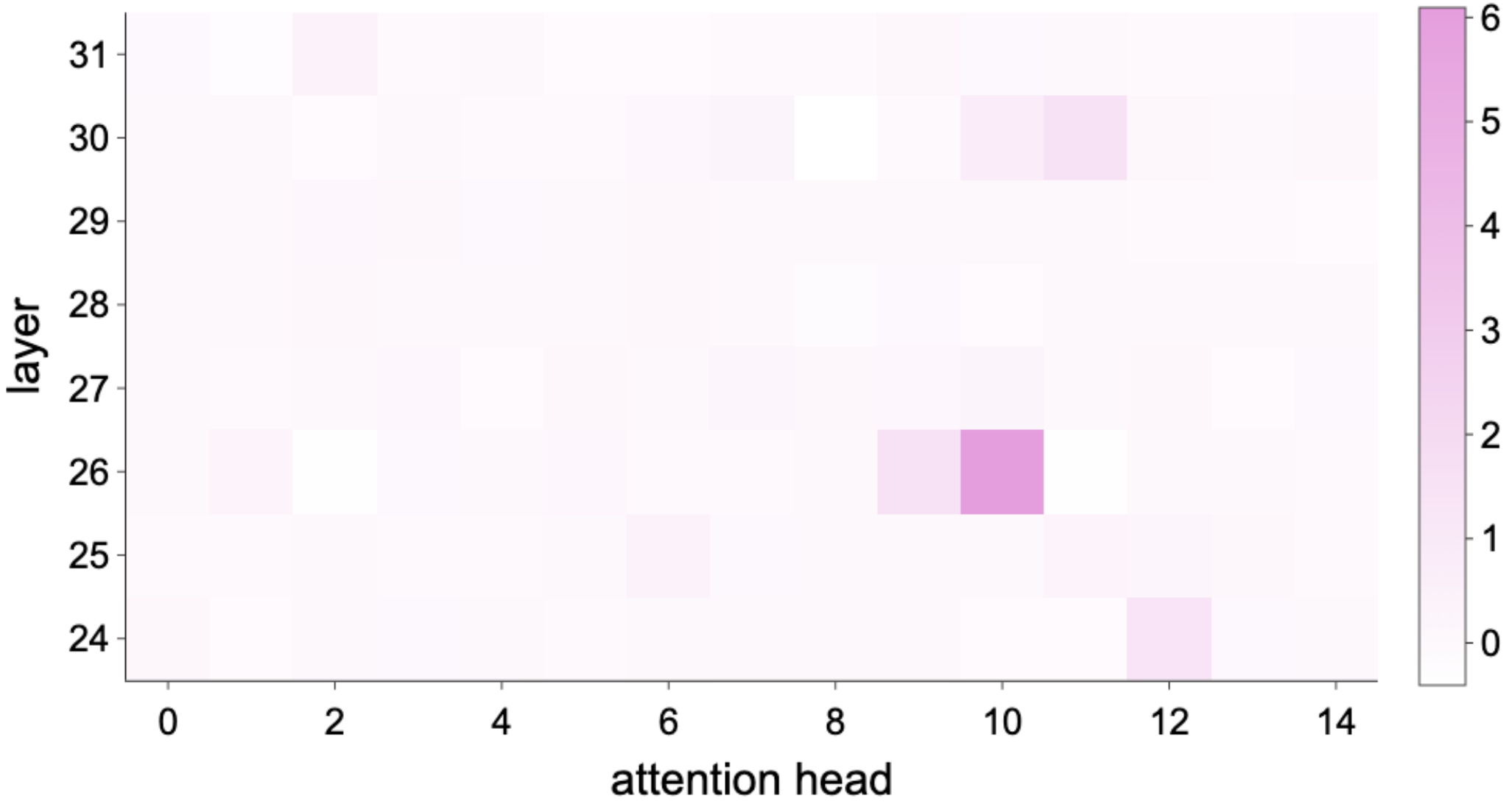
Attention Output



Attention Output



Attention Output Result



Case Study 2: State Tracking

Alice poured coffee into a cup.

Then, she put it in the fridge.

Now, the cup is ___cold___

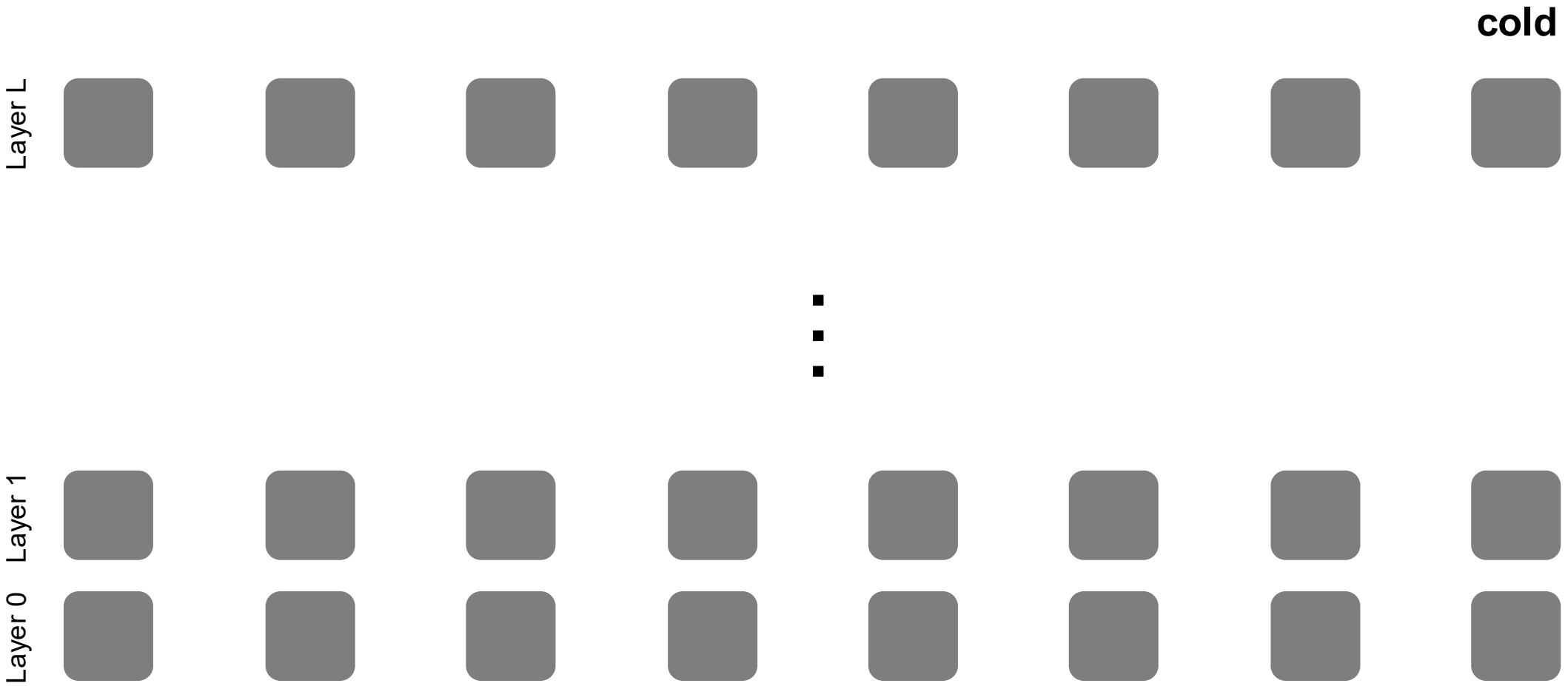
Case Study 2: State Tracking

Alice poured coffee into a cup.

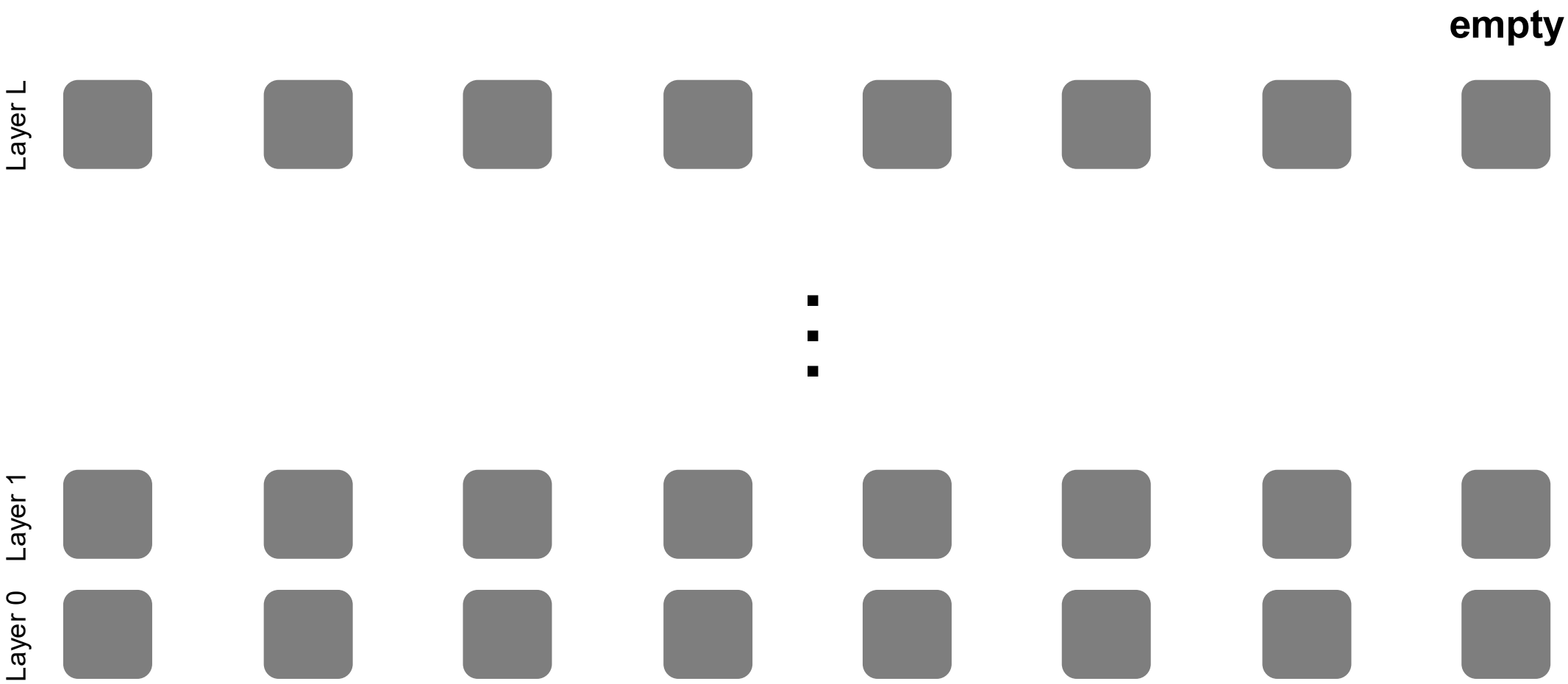
Then, she drank it in one go.

Now, the cup is __empty__

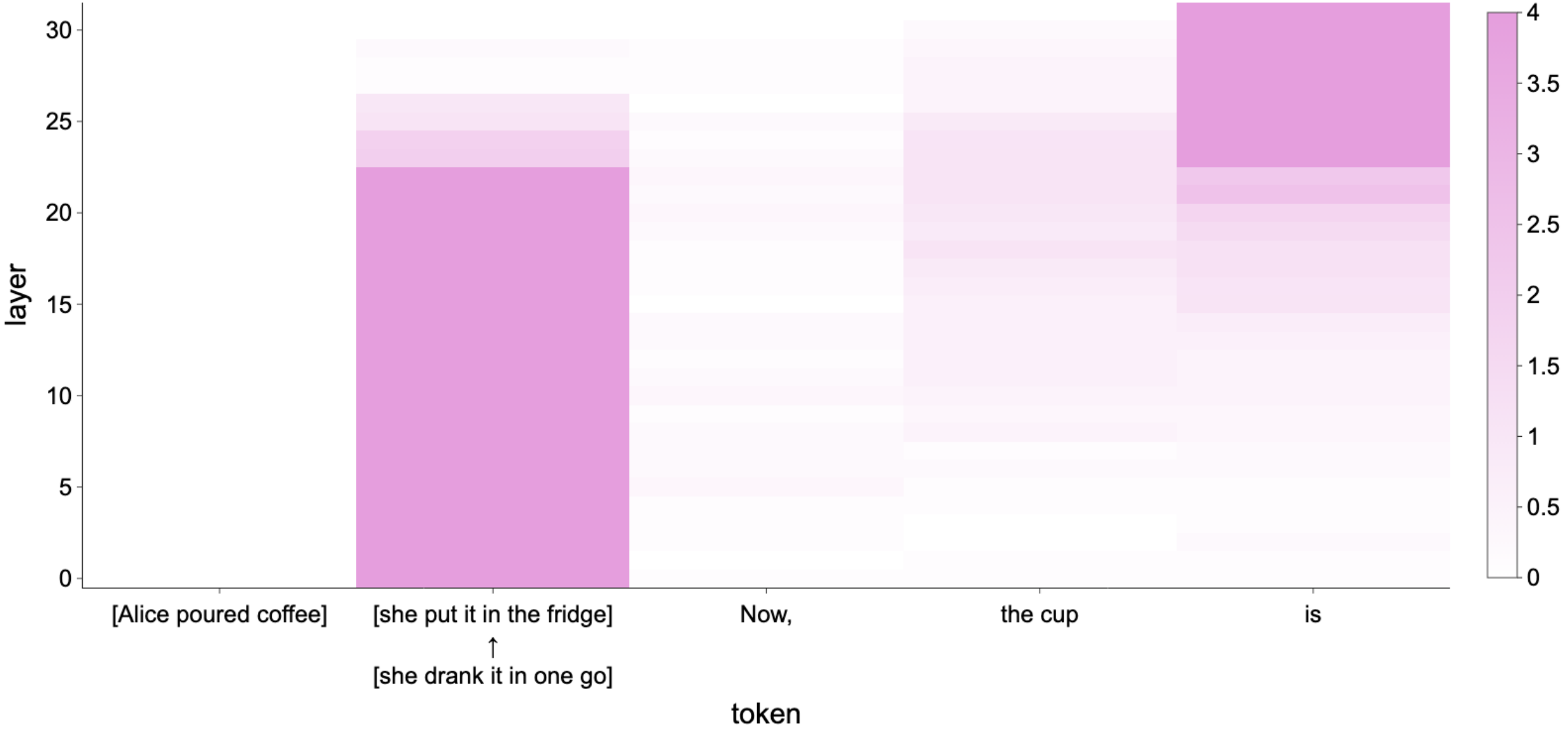
Case Study 2: State Tracking



Case Study 2: State Tracking



Case Study 2 Result



Case Study 3: Situation Model

Alice returned the book to Sarah because she **finished** it.

Alice returned the book to Sarah because she **needed** it.

Exercise 2

Alice returned the book to Sarah because she finished it.
Who does ”she” refer to?
Answer: **Alice**

Alice returned the book to Sarah because she needed it.
Who does ”she” refer to?
Answer: **Sarah**

Exercise 2

Bec returned the book to **Cynthia** because she finished it.
Who does ”she” refer to?
Answer: **Alice** → **Bec**

Alice returned the book to Sarah because she needed it.
Who does ”she” refer to?
Answer: **Sarah**

Case Study 4: Composition of Interventions

(1) The **Eiffel** Tower is in the country of **France**

(2) The **Tokyo** Tower is in the country of **Japan**

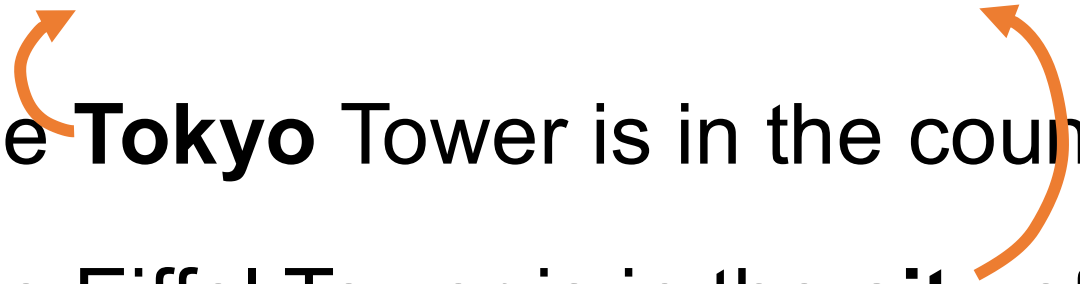
(3) The Eiffel Tower is in the **city** of **Paris**

Case Study 4: Composition of Interventions

(base) The **Eiffel** Tower is in the country of **France**

(source 1) The **Tokyo** Tower is in the country of **Japan**

(source 2) The Eiffel Tower is in the **city** of **Paris**



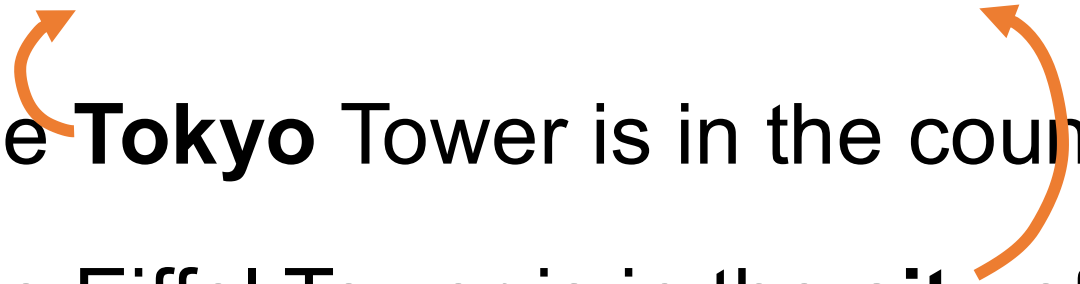
(base') The **Eiffel** Tower is in the country of ____?____

Case Study 4: Composition of Interventions

(base) The **Eiffel** Tower is in the country of **France**

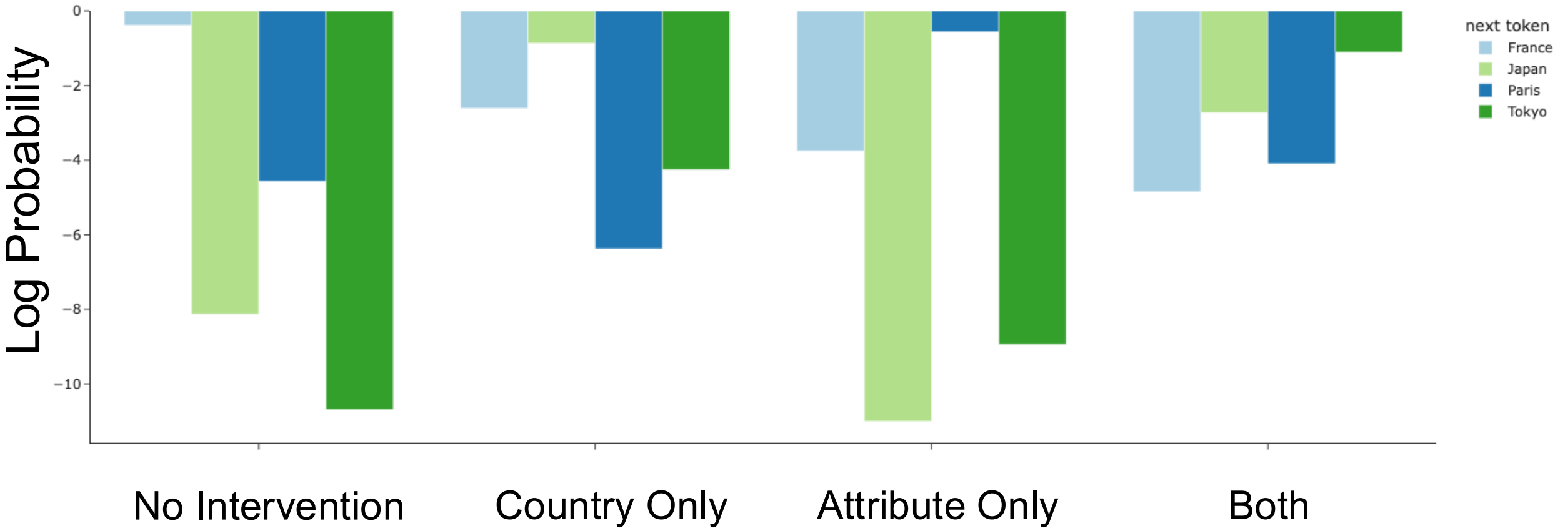
(source 1) The **Tokyo** Tower is in the country of **Japan**

(source 2) The Eiffel Tower is in the **city** of **Paris**



(base') The **Eiffel** Tower is in the country of **Tokyo**

Case Study 4: Composition of Interventions



Summary

- Causal Mediation Analysis allows us to track information flow
- We can study
  - state tracking
  - entity binding
  - etc.

Limitations?

- We need appropriate minimal pairs for each aspect of behavior
- Only a sufficiency condition
- Can be computationally expensive

To be continued…