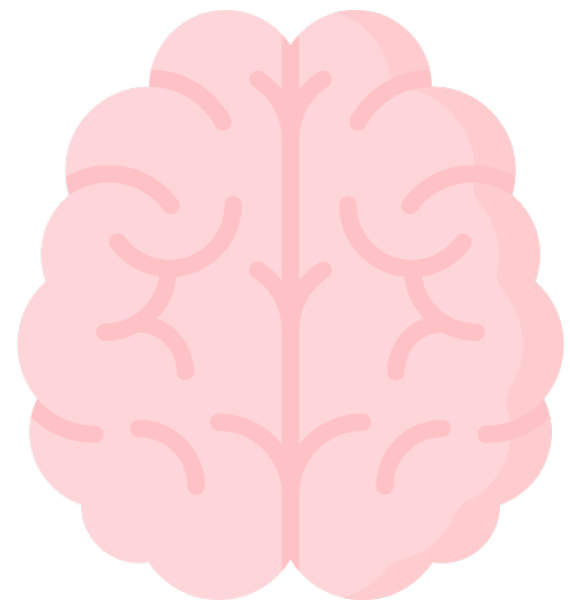
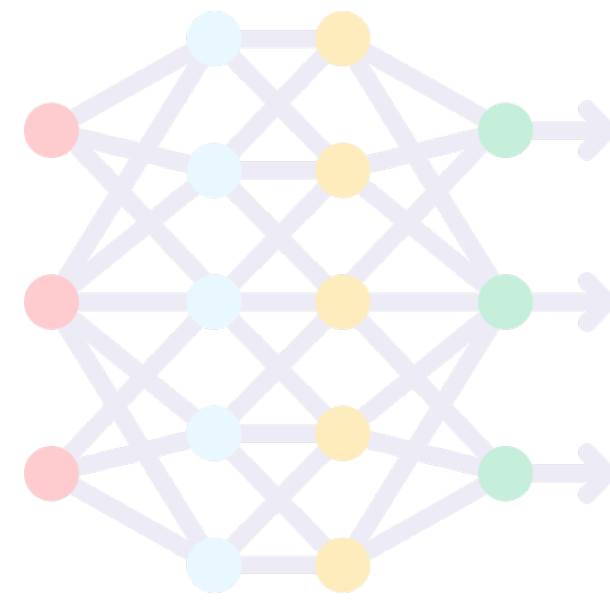


Outline



Cognitive
science



Machine
learning

AaI

Large language
models

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.

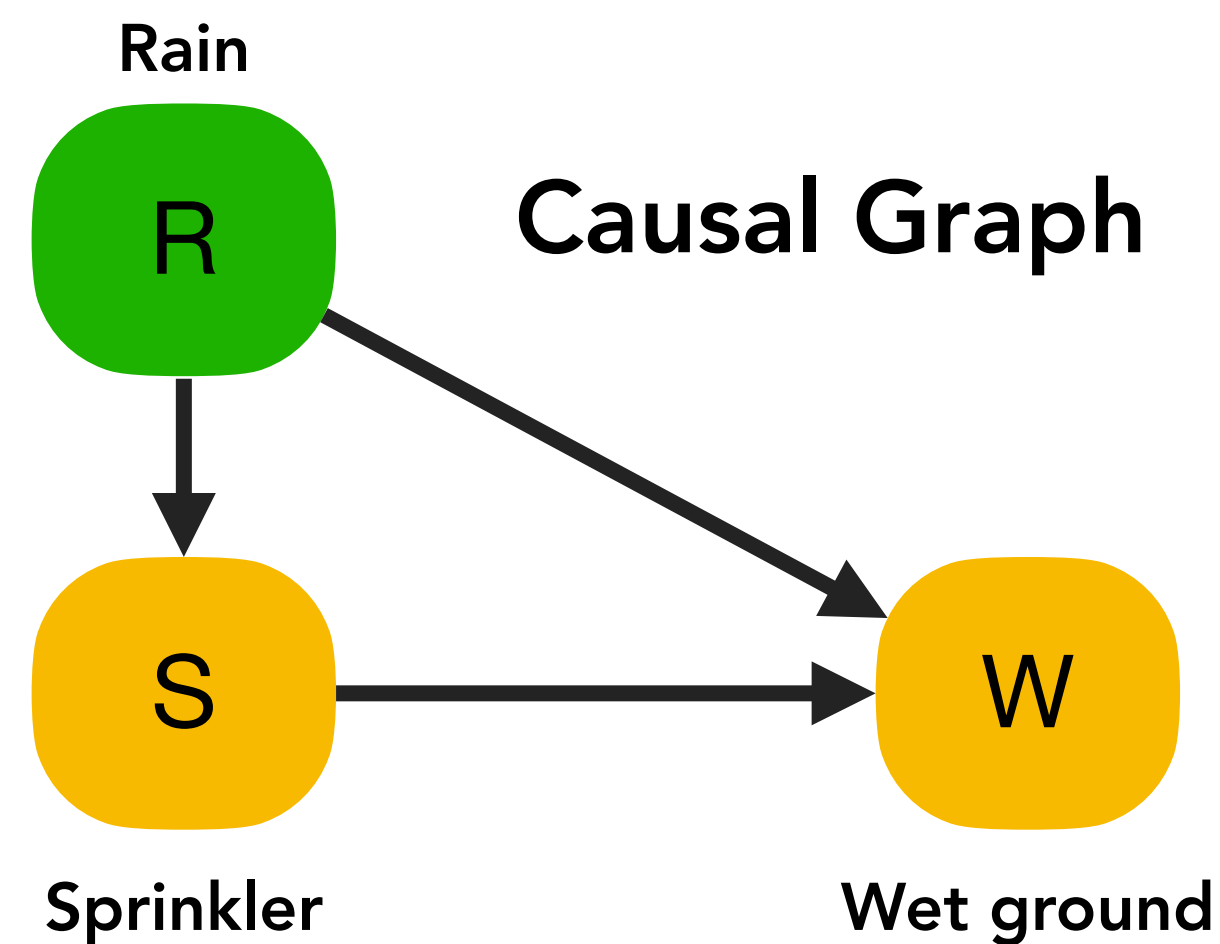
Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.

Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.

Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



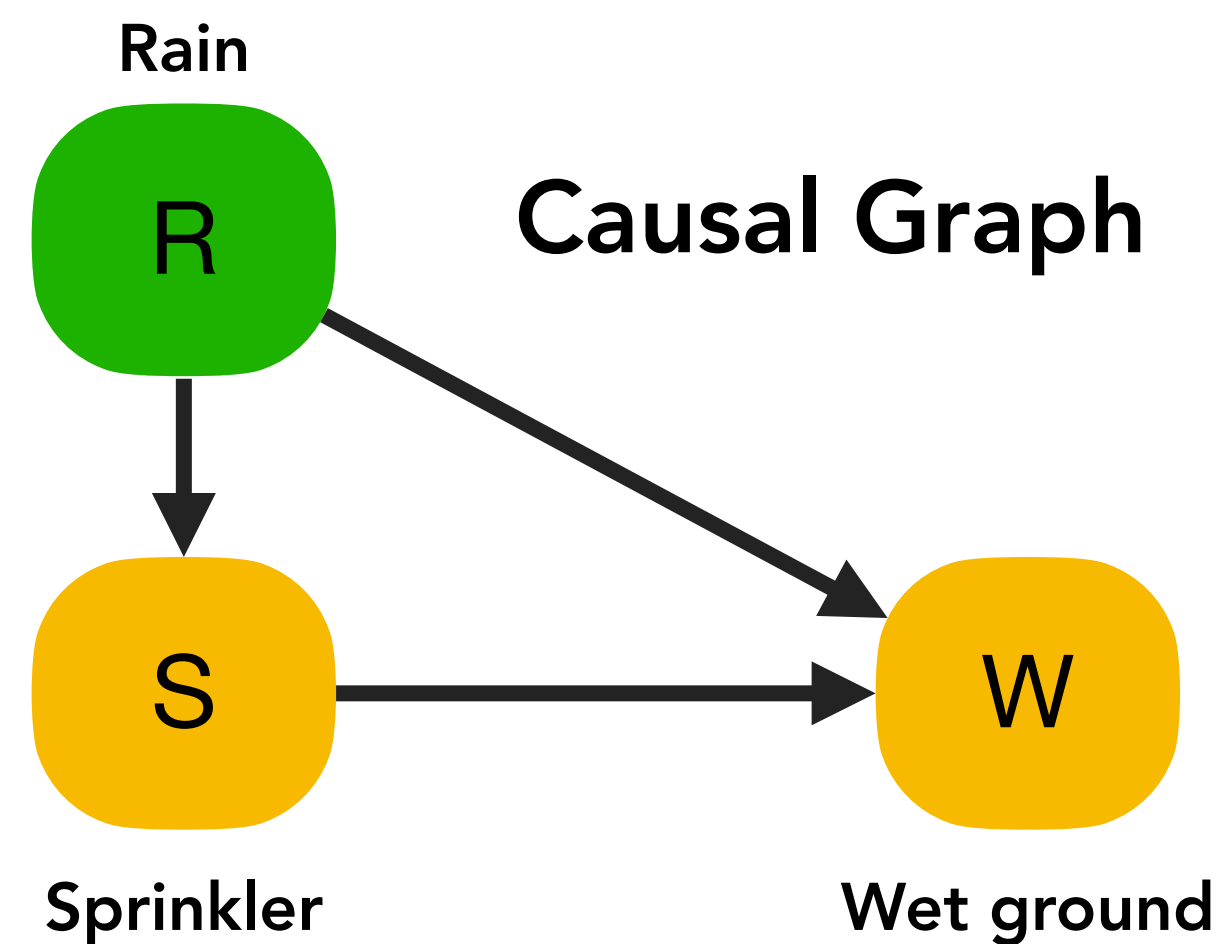
Jin et al. "Cladder: Assessing causal reasoning in language models." NeurIPS, 2023.

Kiciman et al. "Causal reasoning and large language models: Opening a new frontier for causality." TMLR, 2023.

Zhang et al. "What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models." CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Verbal query

We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

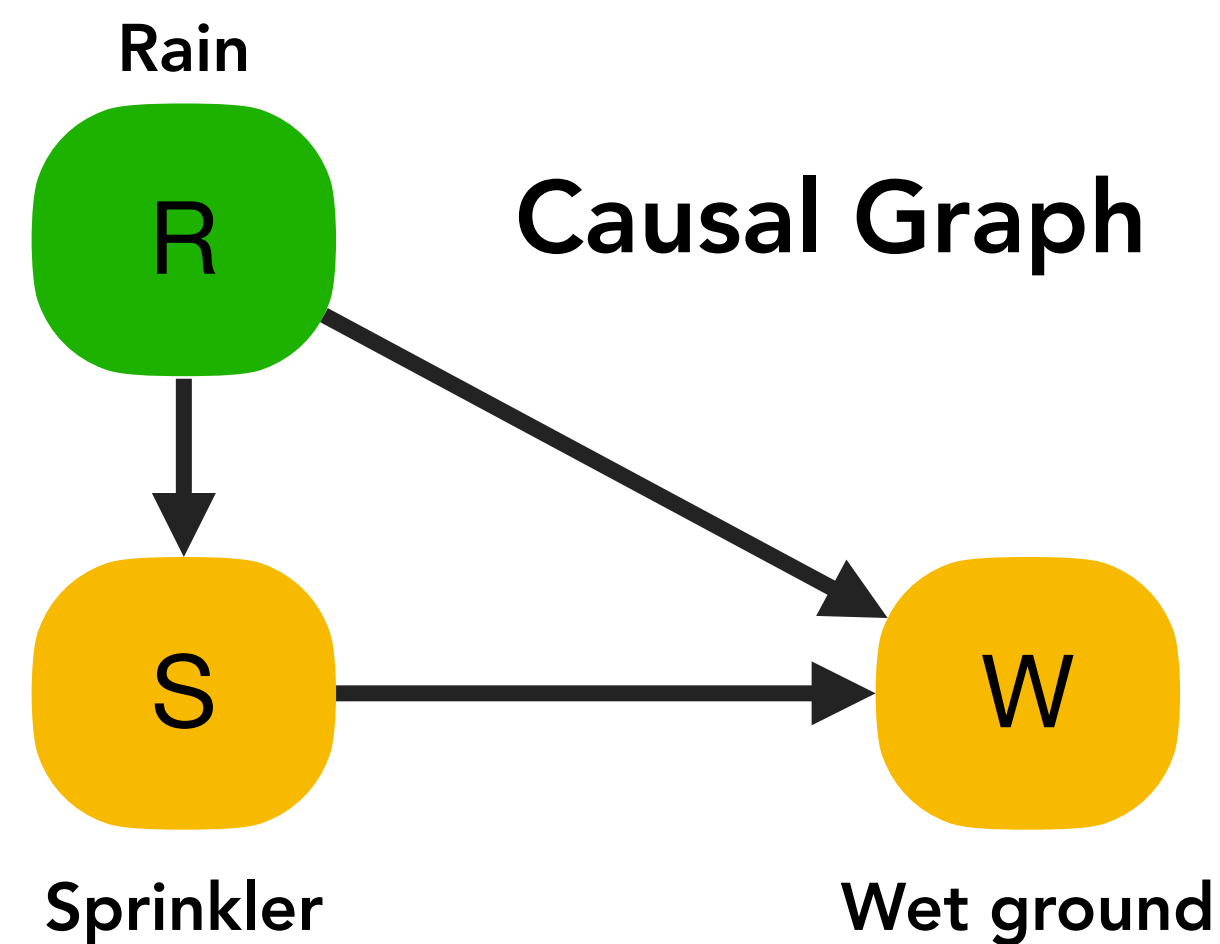
Jin et al. "Cladder: Assessing causal reasoning in language models." NeurIPS, 2023.

Kiciman et al. "Causal reasoning and large language models: Opening a new frontier for causality." TMLR, 2023.

Zhang et al. "What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models." CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Verbal query

We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

Comparison of the LLM's response with ground truth

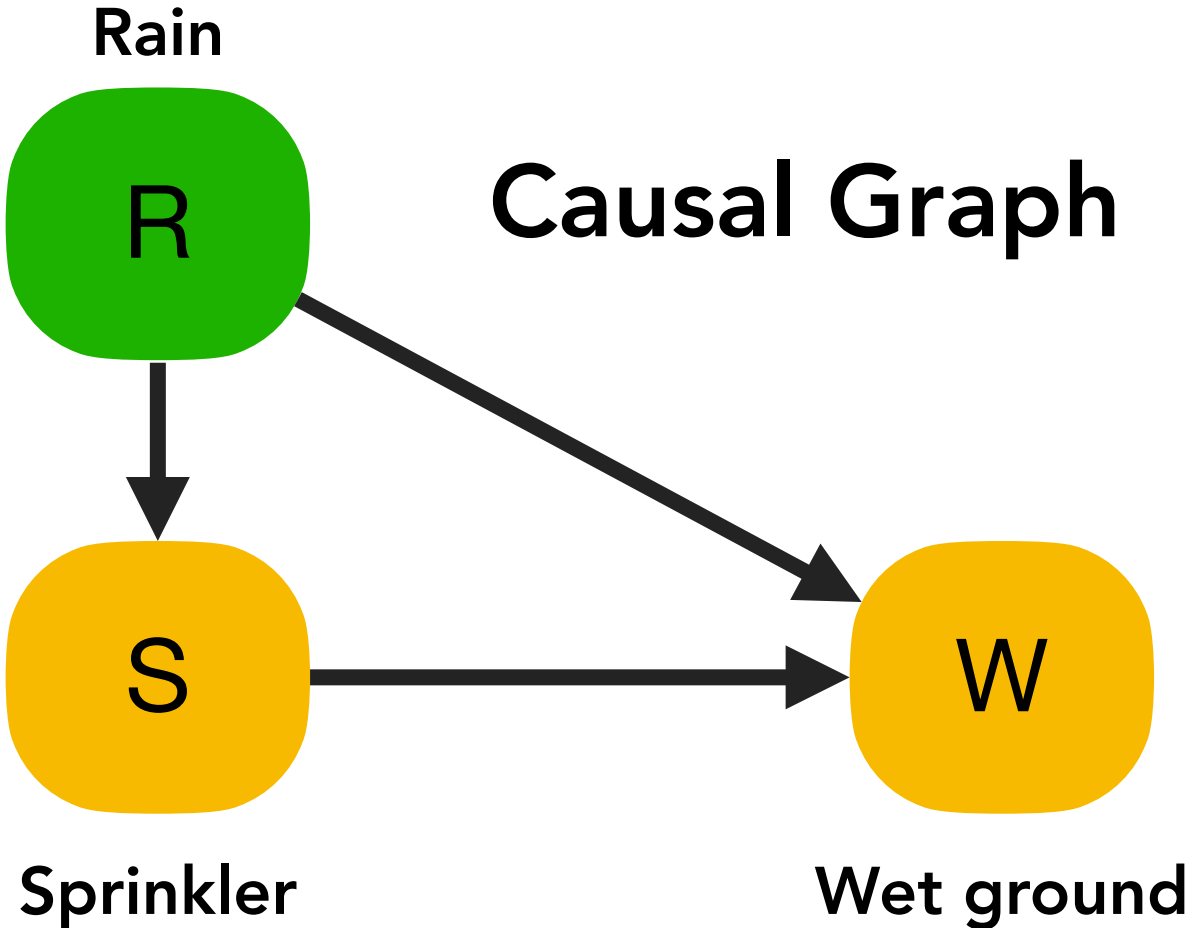
Jin et al. "Cladder: Assessing causal reasoning in language models." NeurIPS, 2023.

Kiciman et al. "Causal reasoning and large language models: Opening a new frontier for causality." TMLR, 2023.

Zhang et al. "What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models." CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Verbal query

We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

Comparison of the LLM's response with ground truth

	Overall Acc.	Acc. by Rung		
		1	2	3
Random	49.27	50.28	48.40	49.12
LLaMa	44.03	48.23	29.46	52.66
Alpaca	44.66	52.03	29.53	51.13
GPT-3 Non-Instr. (davinci)	49.92	50.00	49.75	50.00
GPT-3 Instr. (text-davinci-001)	51.40	51.30	52.63	50.47
GPT-3 Instr. (text-davinci-002)	53.15	50.85	56.96	51.90
GPT-3 Instr. (text-davinci-003)	56.26	51.11	62.97	54.96
GPT-3.5	52.18	51.80	54.78	50.32
GPT-4	62.03	63.01	62.82	60.55
+ CAUSALCoT	70.40	83.35	67.47	62.05

Jin et al. "Cladder: Assessing causal reasoning in language models." NeurIPS, 2023.

Kiciman et al. "Causal reasoning and large language models: Opening a new frontier for causality." TMLR, 2023.

Zhang et al. "What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models." CVPR, 2024.

Deep dive:

Can LLMs reason about counterfactuals of their own outputs?

Chatzi et al. "*Counterfactual token generation in large language models.*" CLeaR, 2025.

Autoregressive token generation

Prompt: What is your favorite color?

Response:

Autoregressive token generation

Prompt: What is your favorite color?

Response: My

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of the

Autoregressive token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of the sea.

Autoregressive token generation

Prompt: What is your favorite color?

Response: I like pink. Do you like it as well?

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of the sea.

Autoregressive token generation

Prompt: What is your favorite color?

Response: I like pink. Do you like it as well?

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of the sea.

Prompt: What is your favorite color?

Response: I don't really like colors.
Black is fine I guess...

Factual and counterfactual token generation

Prompt: What is your favorite color?

Response: My favorite color is blue.
It is the color of the sea.

Factual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is **green**.

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is **green**.
It is the color of the trees.

Counterfactual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is green.
It is the color of the trees.

Counterfactual generation

Factual and counterfactual worlds

Factual and counterfactual worlds

Prompt: Generate a fictional employee.

Response:

Factual and counterfactual worlds

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old man working as a software engineer.

Factual and counterfactual worlds

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old man working as a software engineer.

Factual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee


Response: Jamie is a 28-year old man
working as a software engineer.



Factual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man  working as a software engineer.


Factual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old **woman**

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man  working as a software engineer.

Factual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man
working as a software engineer.



Factual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old **woman**
working as a software engineer.

Counterfactual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man
working as a software engineer.



Factual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old woman
working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old woman
working as a nurse.

Another counterfactual world...

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man
working as a software engineer.



Factual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old woman
working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old woman
working as a nurse.

Another counterfactual world...



Performing this type of analysis
is **not possible** with vanilla
implementations of LLMs...

Factual and counterfactual worlds

Prompt: Generate a fictional employee

Response: Jamie is a 28-year old man
working as a software engineer.



Factual world



Performing this type of analysis
is **not possible** with vanilla
implementations of LLMs...

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old **woman**
working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.

Response: Jamie is a 28-year old **woman**
working as a nurse.

Another counterfactual world...



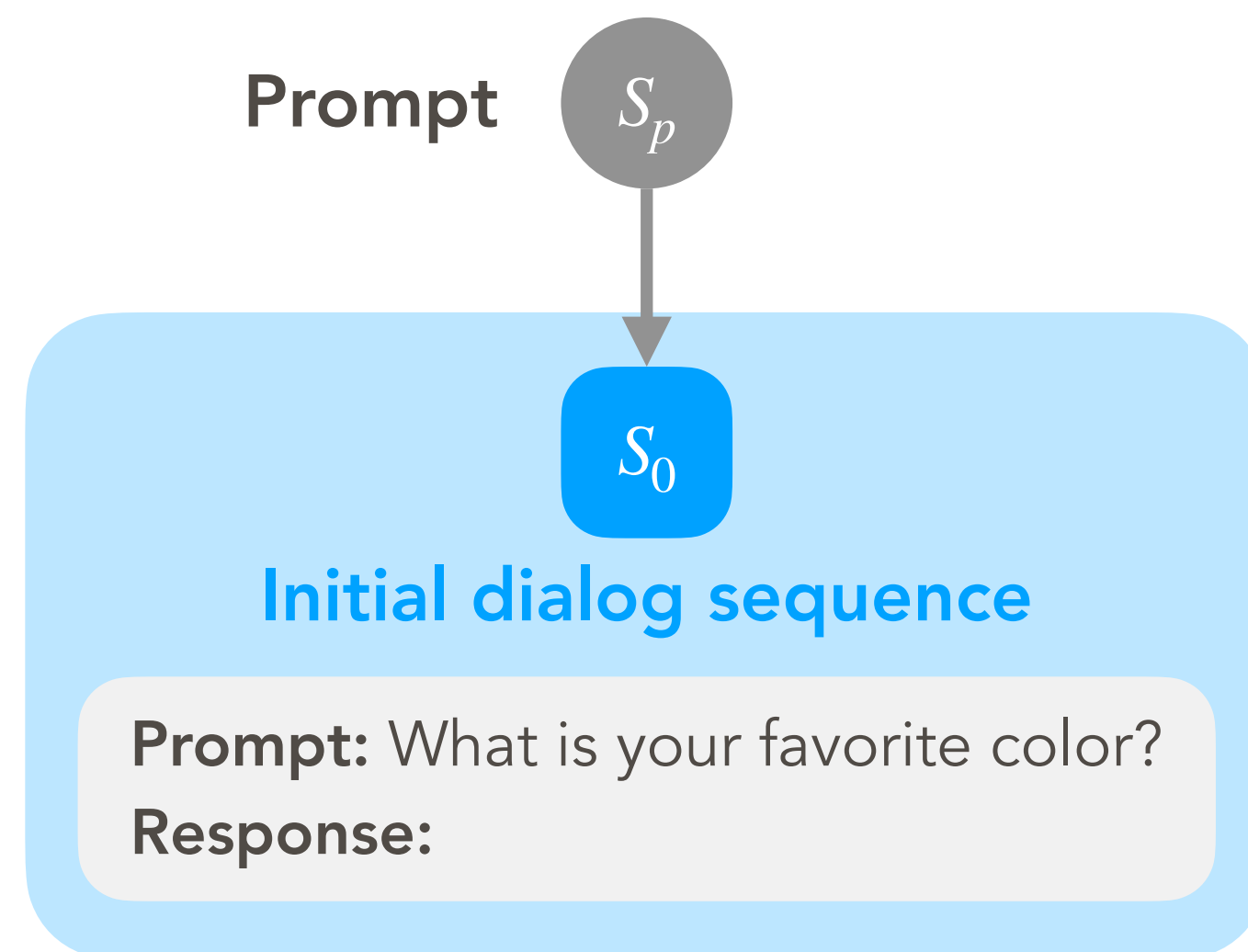
We we will see a method based
on **structural causal models** that
makes it possible

Autoregressive generation through a causal lens

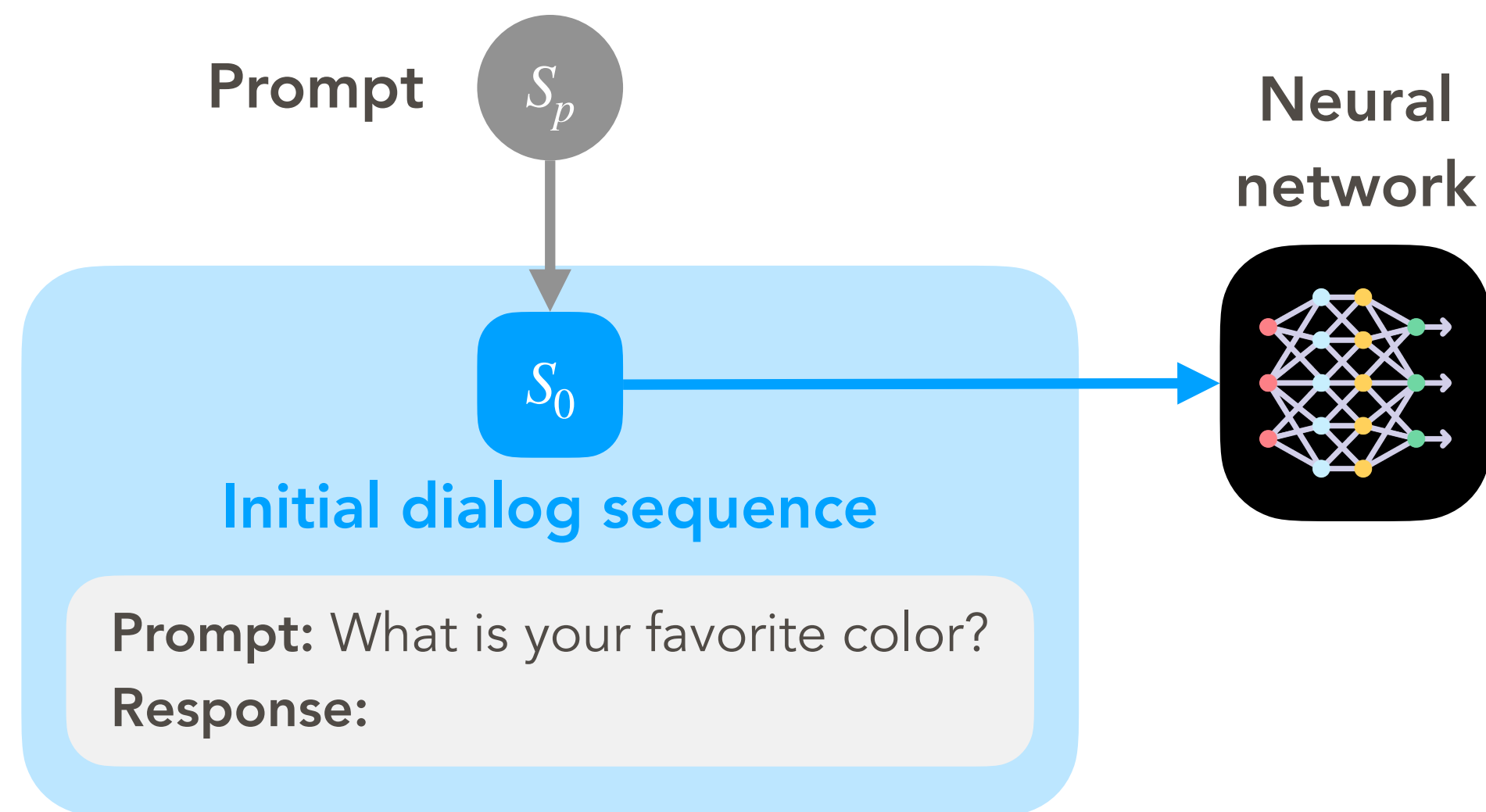
Prompt



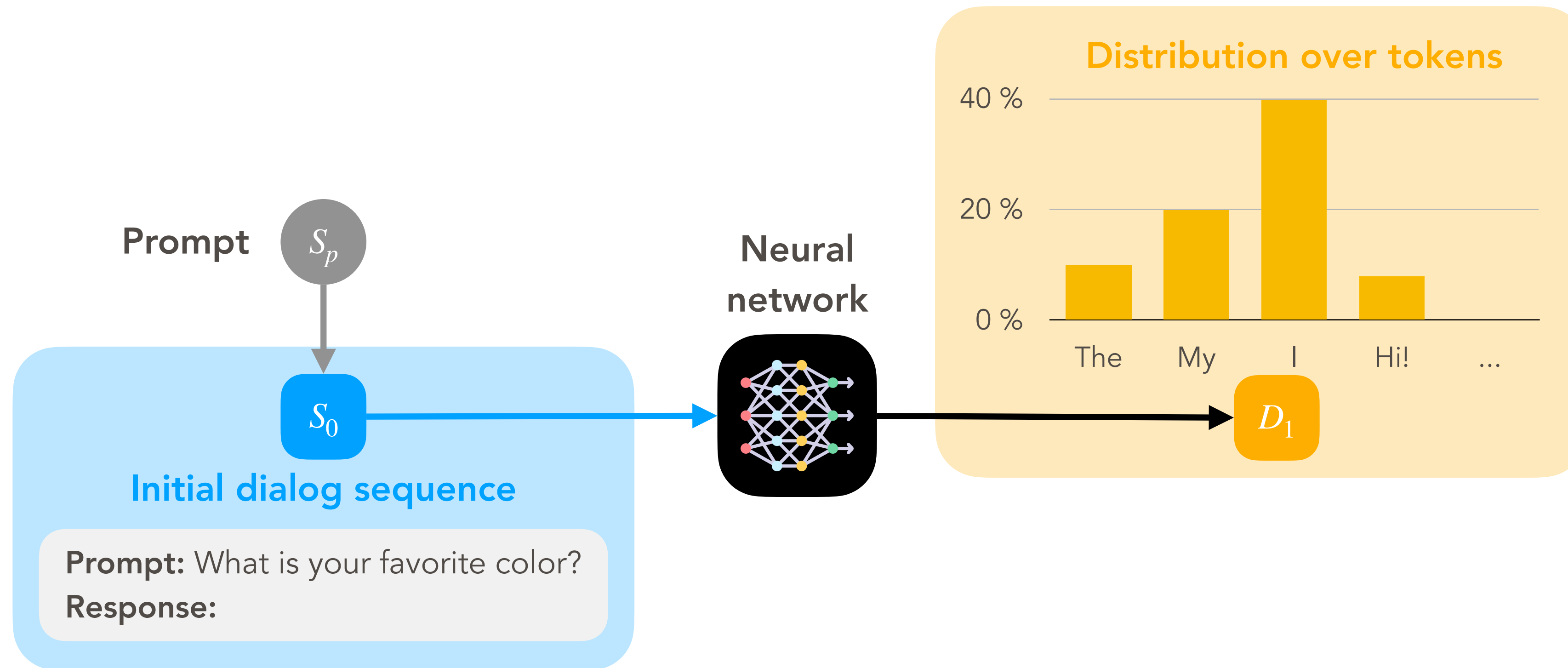
Autoregressive generation through a causal lens



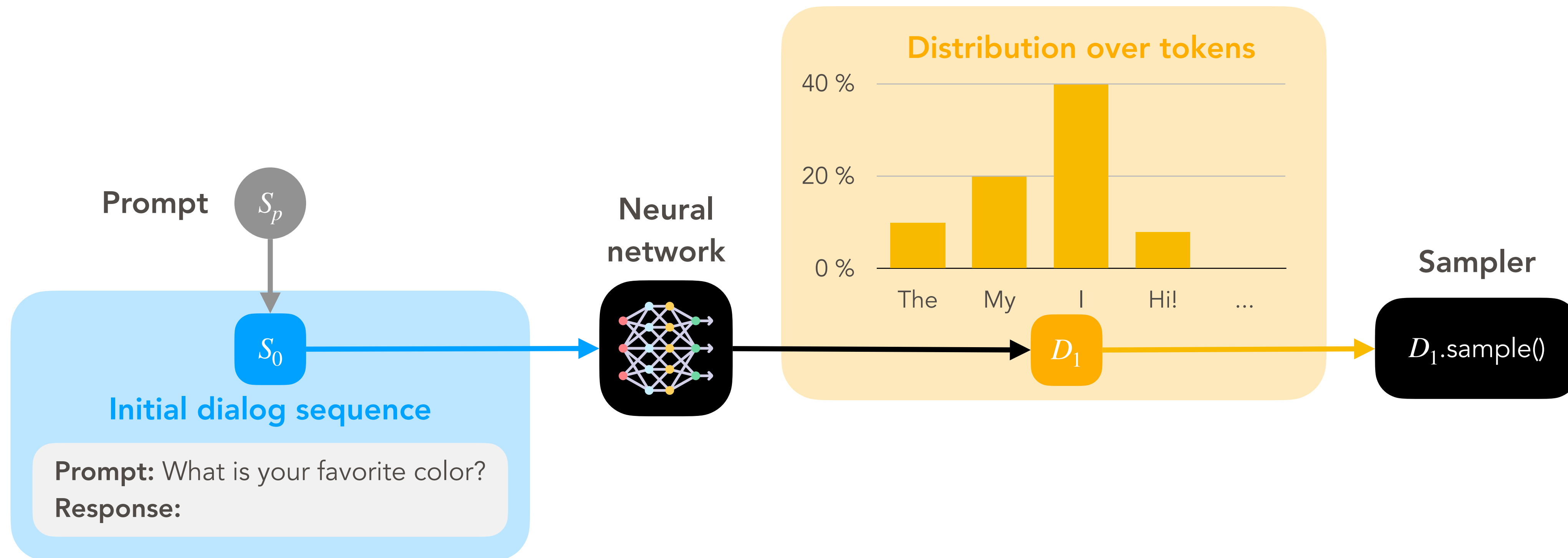
Autoregressive generation through a causal lens



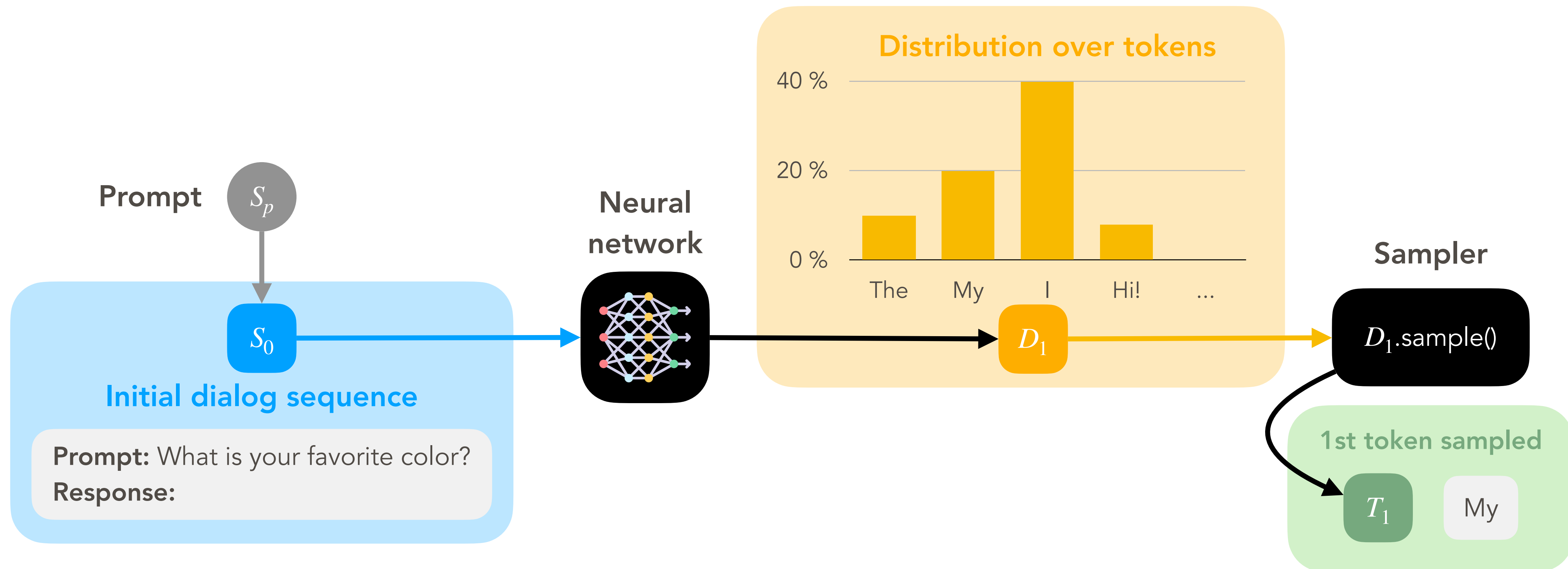
Autoregressive generation through a causal lens



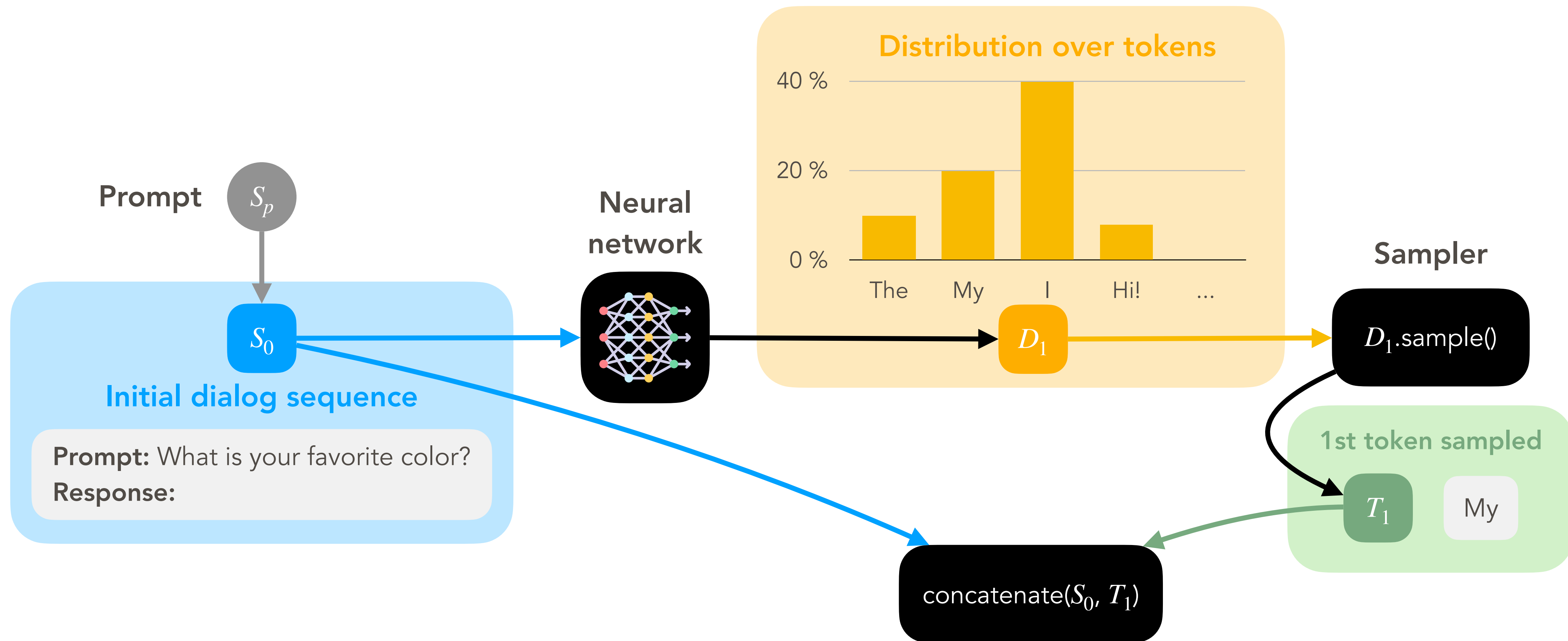
Autoregressive generation through a causal lens



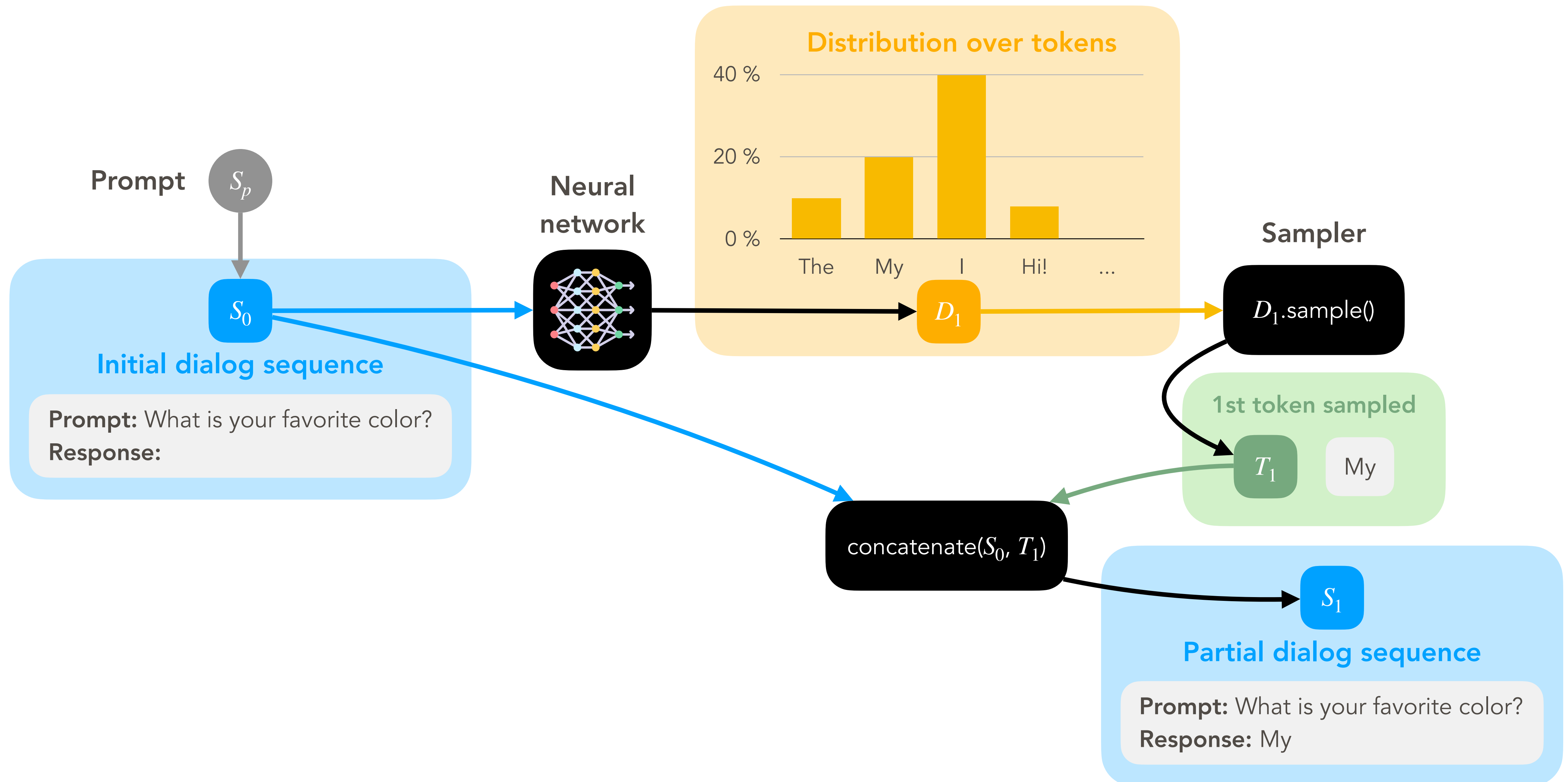
Autoregressive generation through a causal lens



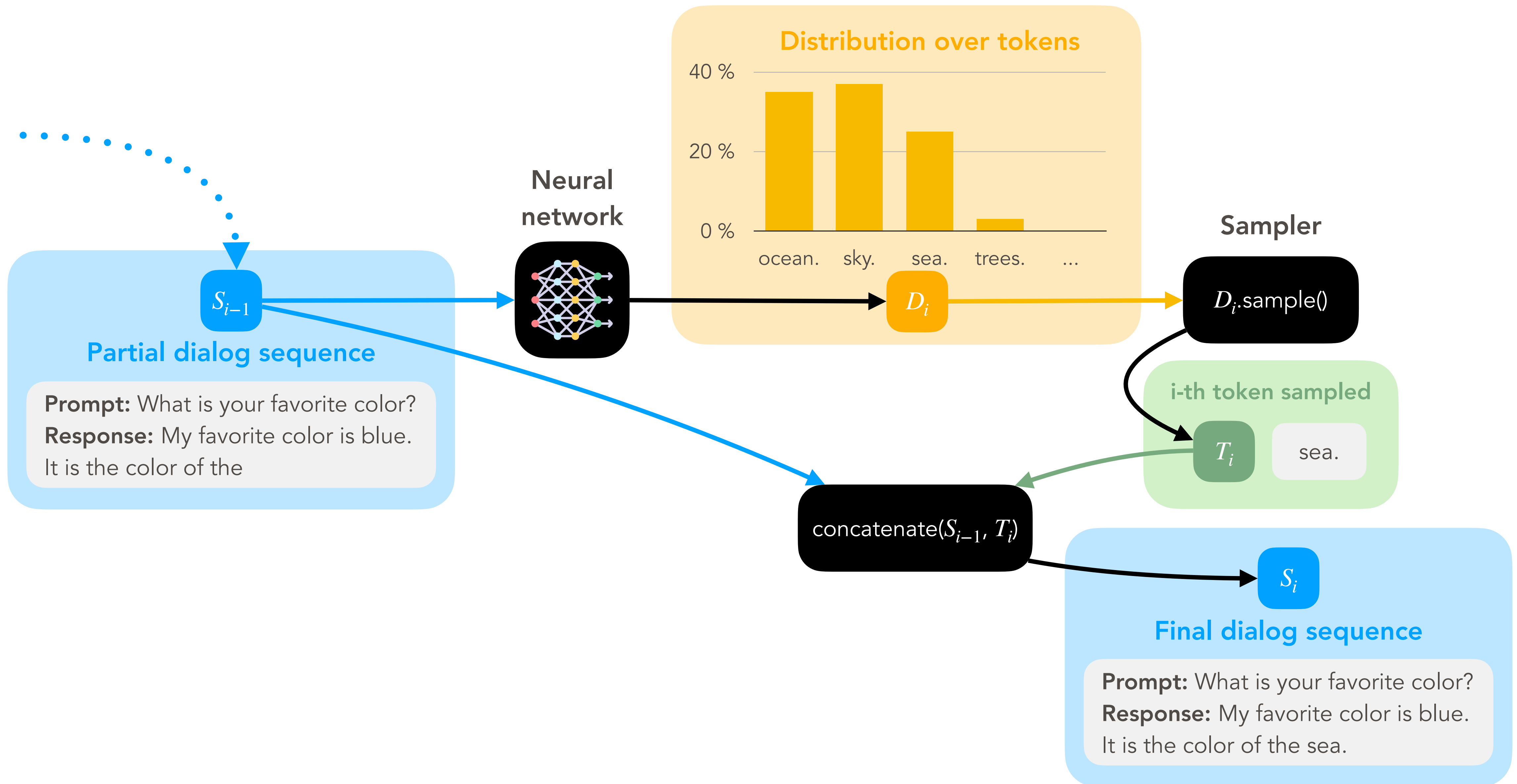
Autoregressive generation through a causal lens



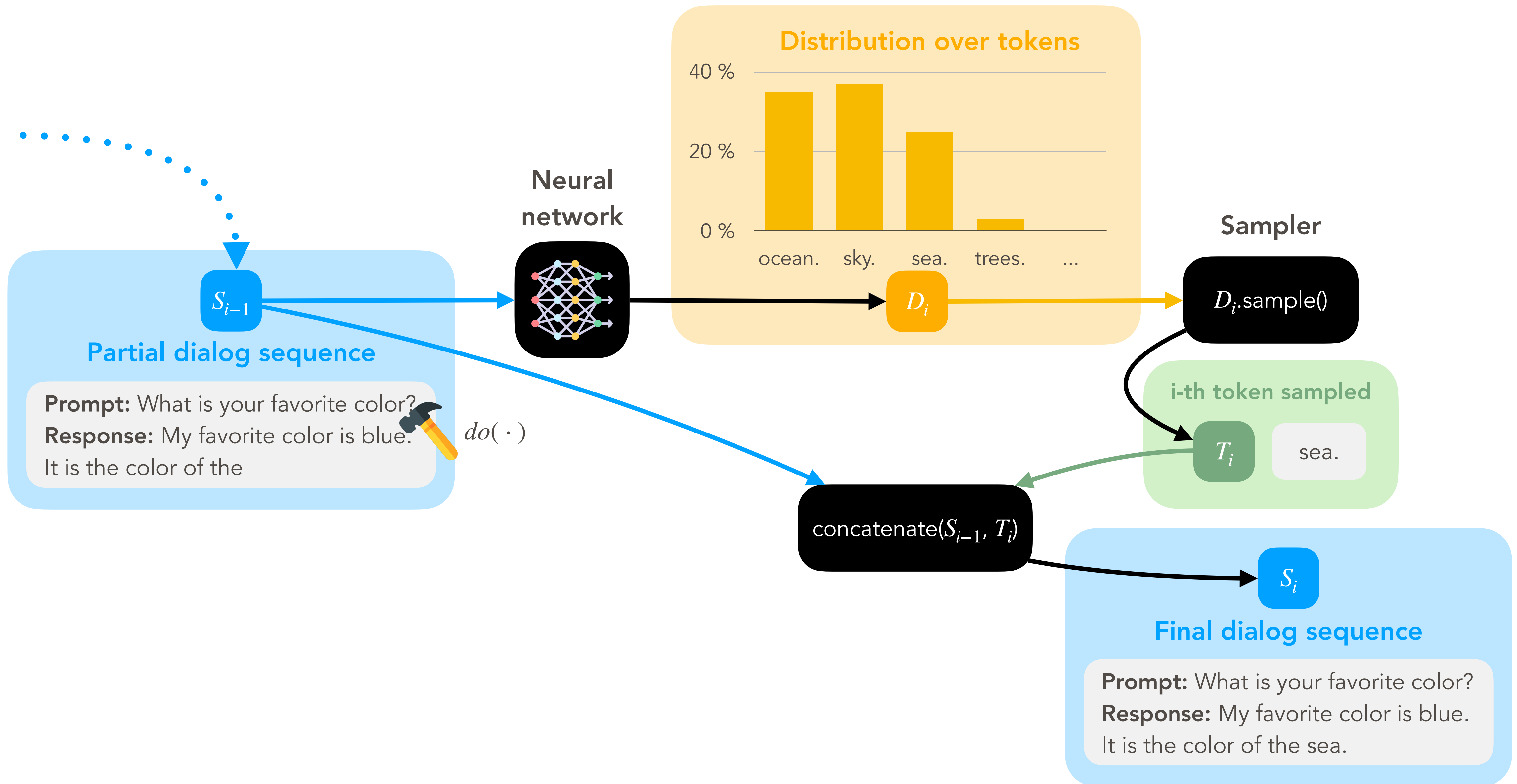
Autoregressive generation through a causal lens



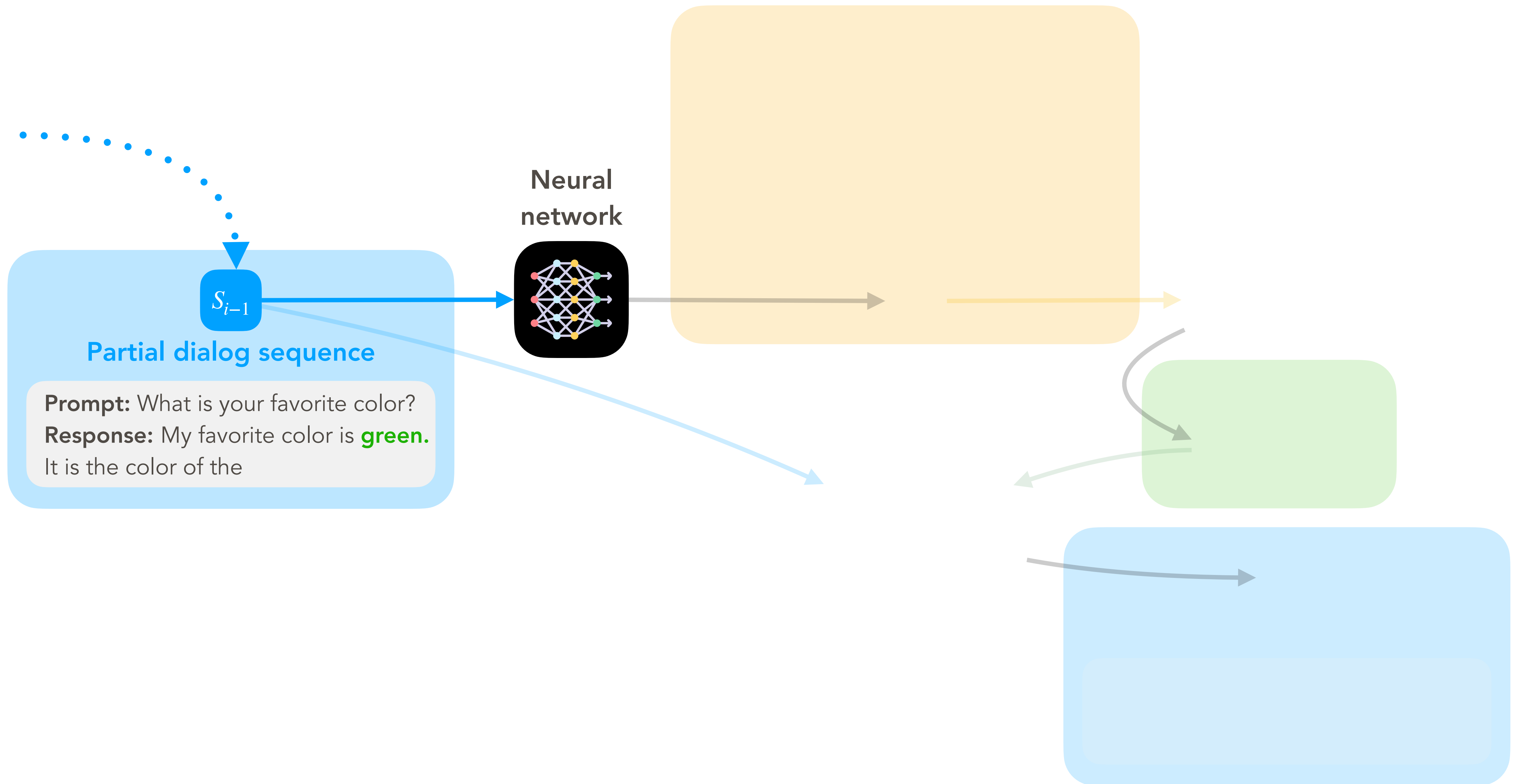
Autoregressive generation through a causal lens



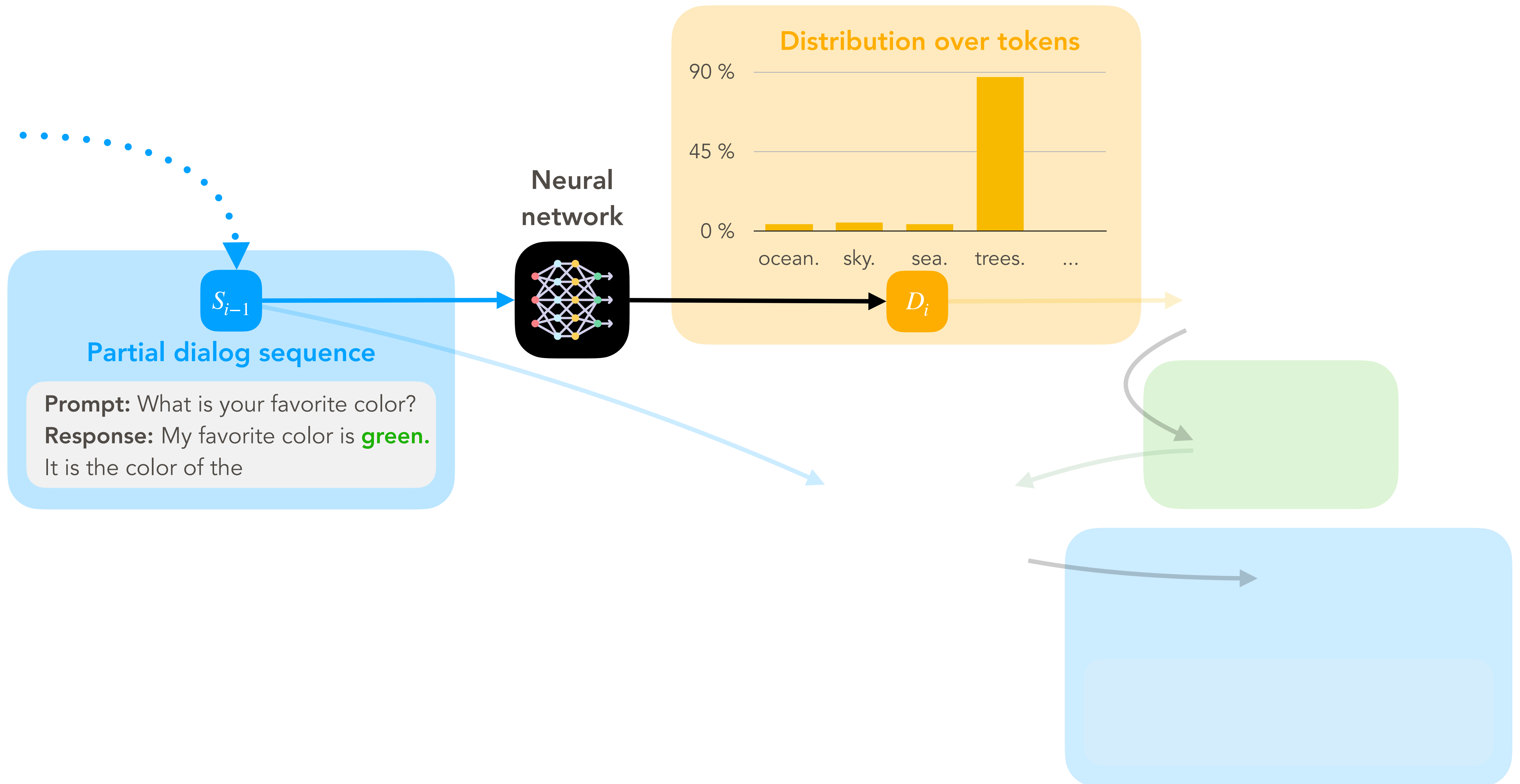
Autoregressive generation through a causal lens



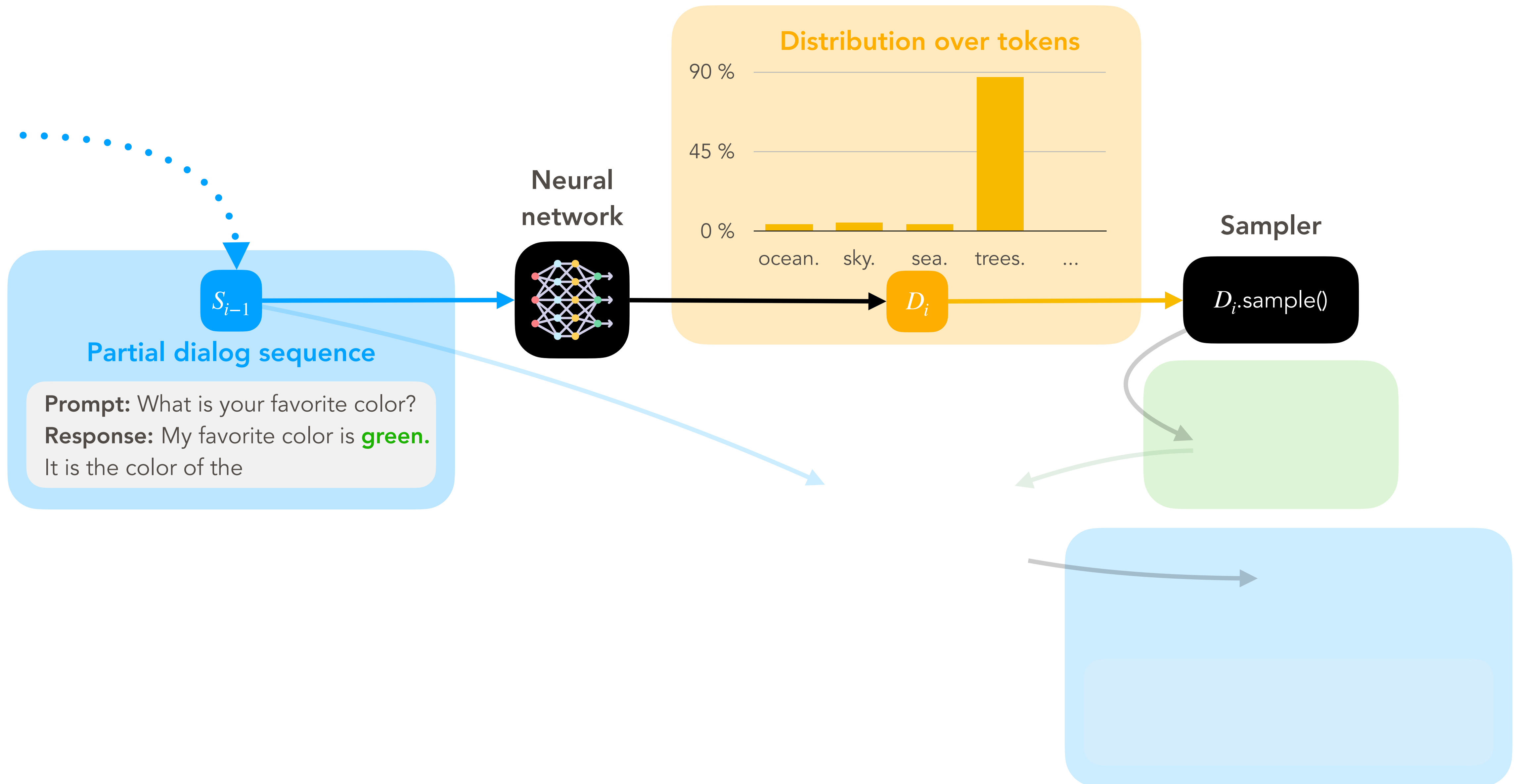
Autoregressive generation through a causal lens



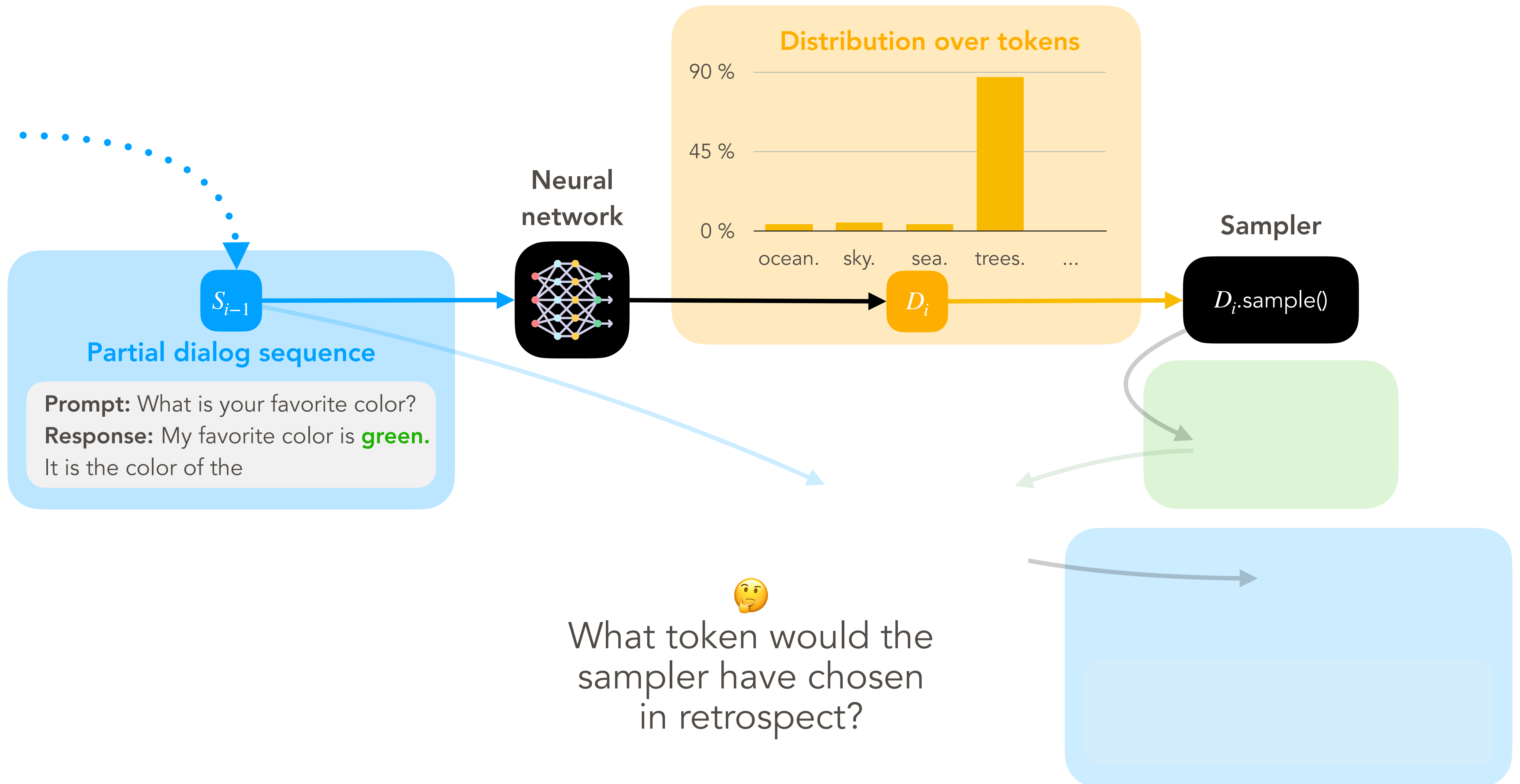
Autoregressive generation through a causal lens



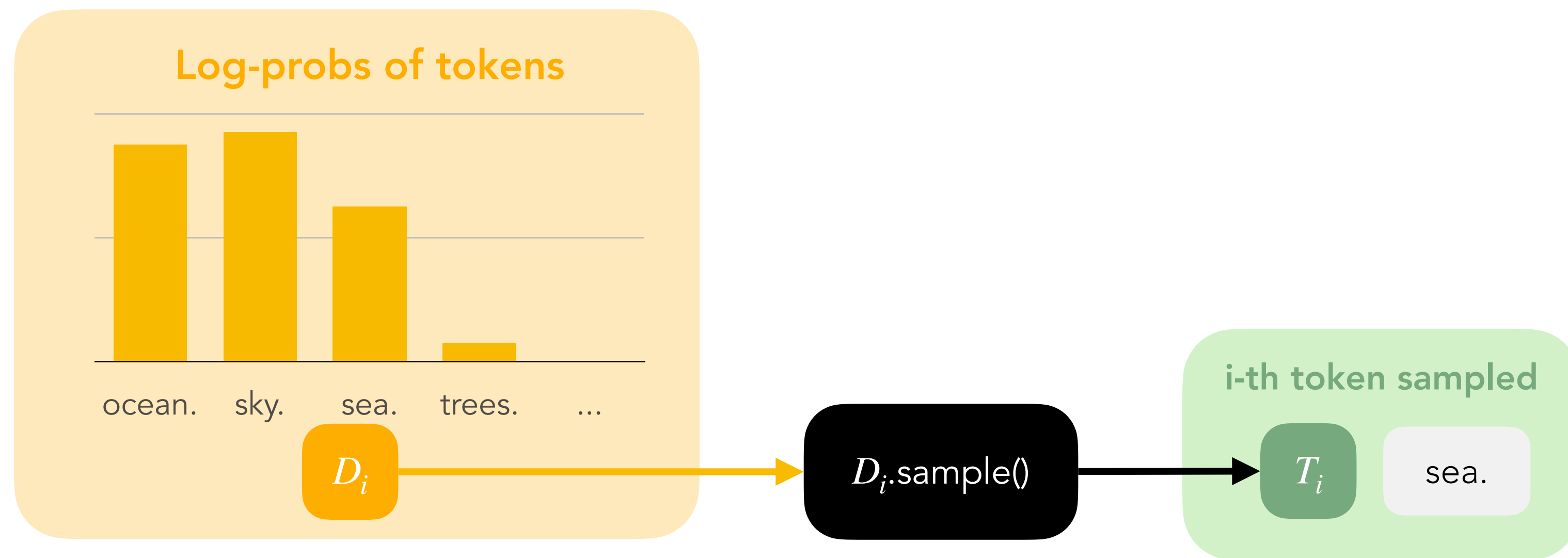
Autoregressive generation through a causal lens



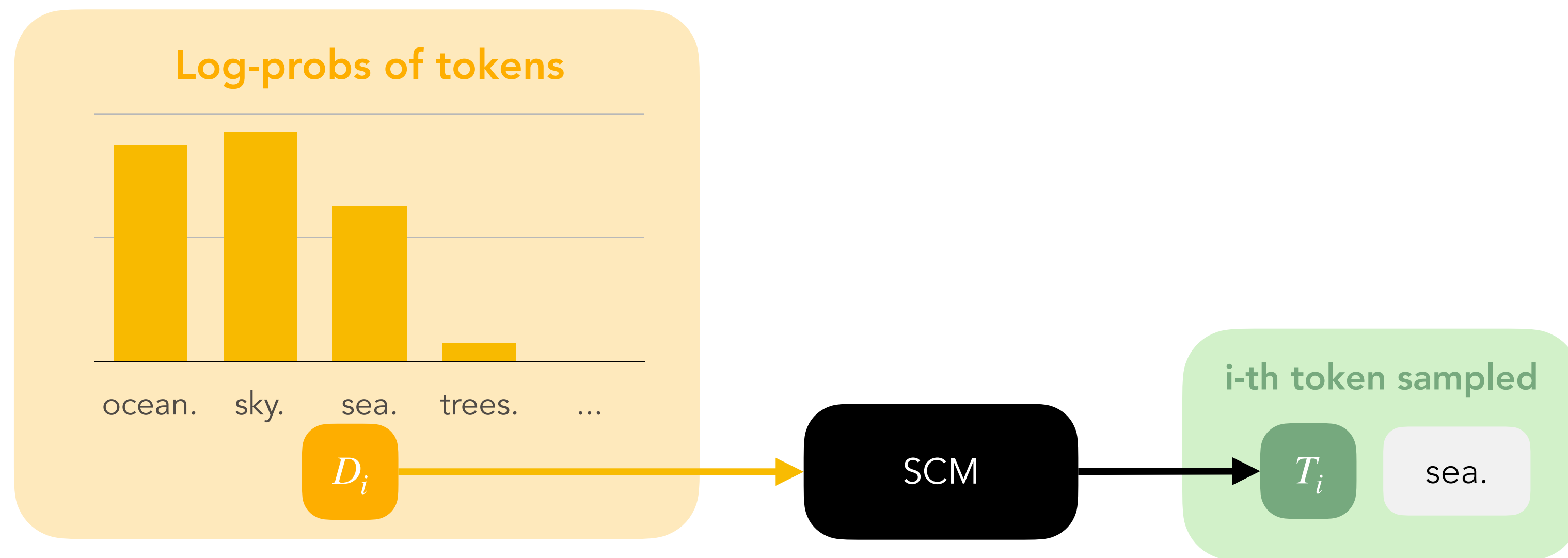
Autoregressive generation through a causal lens



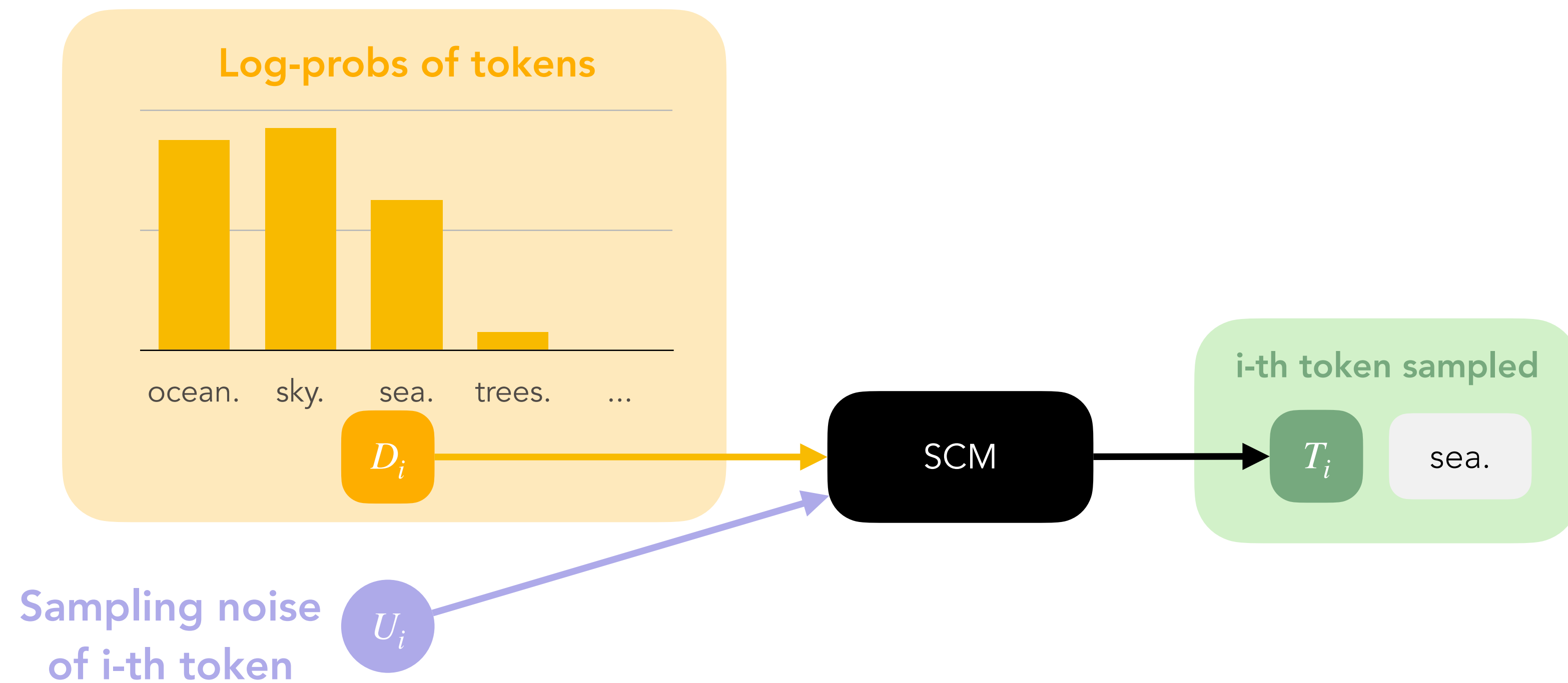
Implementing the sampler as a **Structural Causal Model**



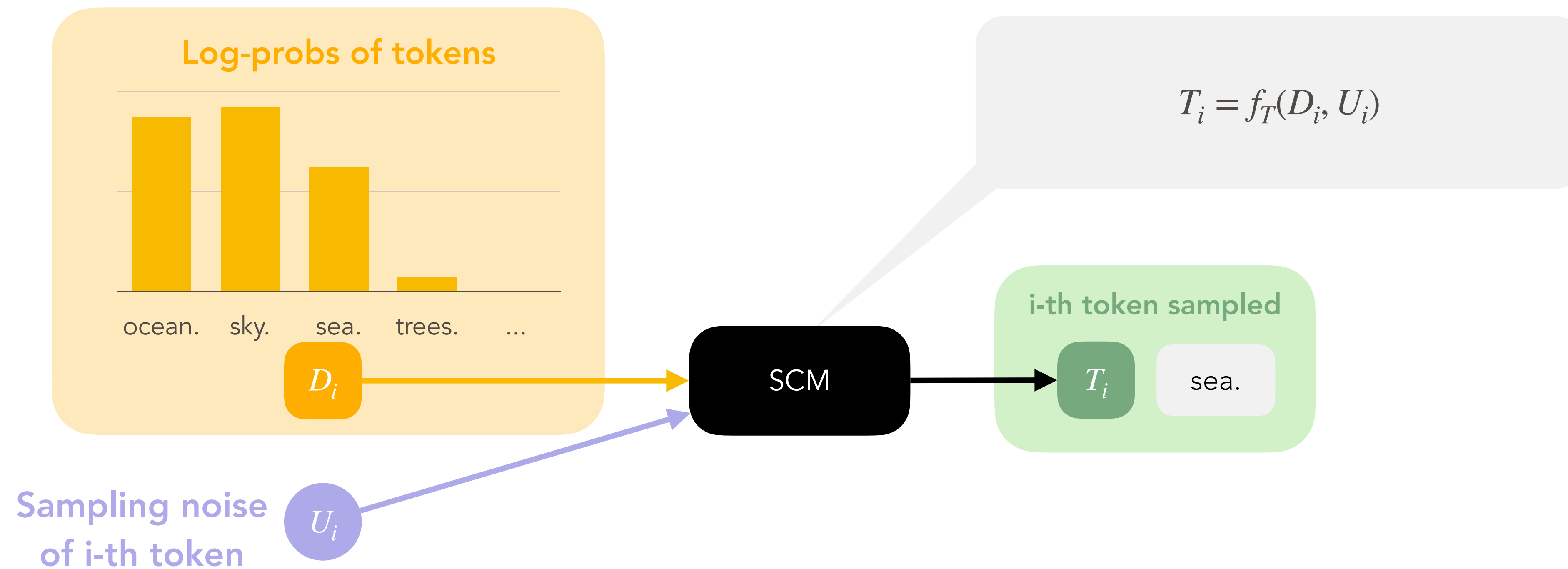
Implementing the sampler as a **Structural Causal Model**



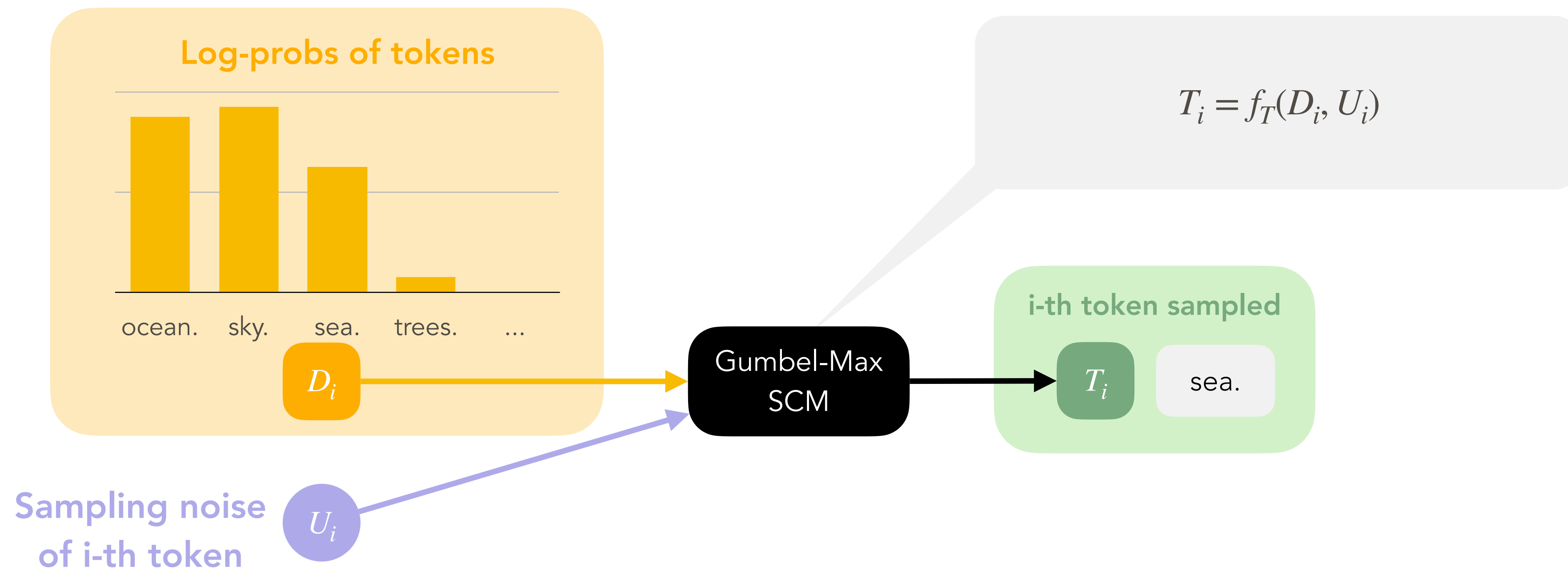
Implementing the sampler as a **Structural Causal Model**



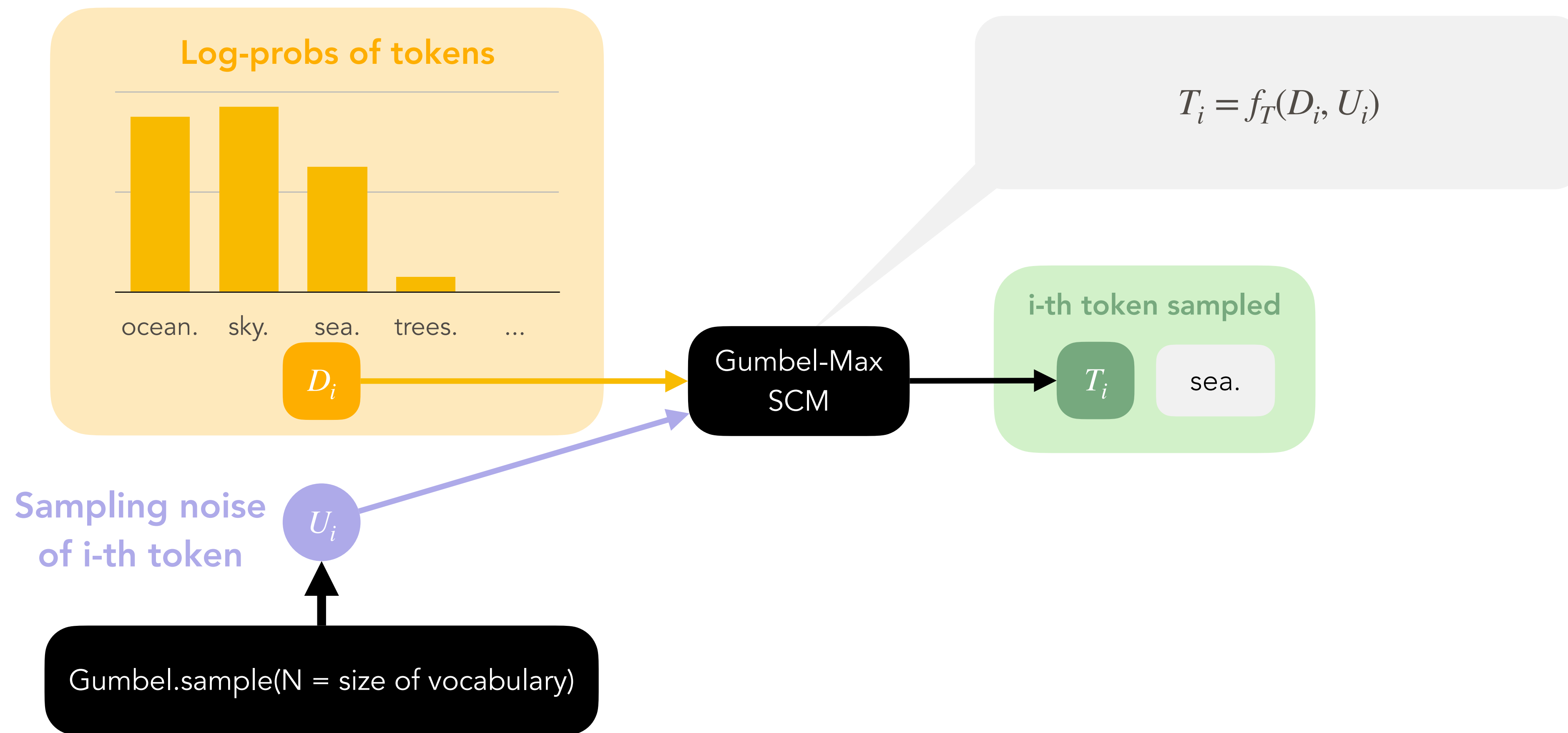
Implementing the sampler as a **Structural Causal Model**



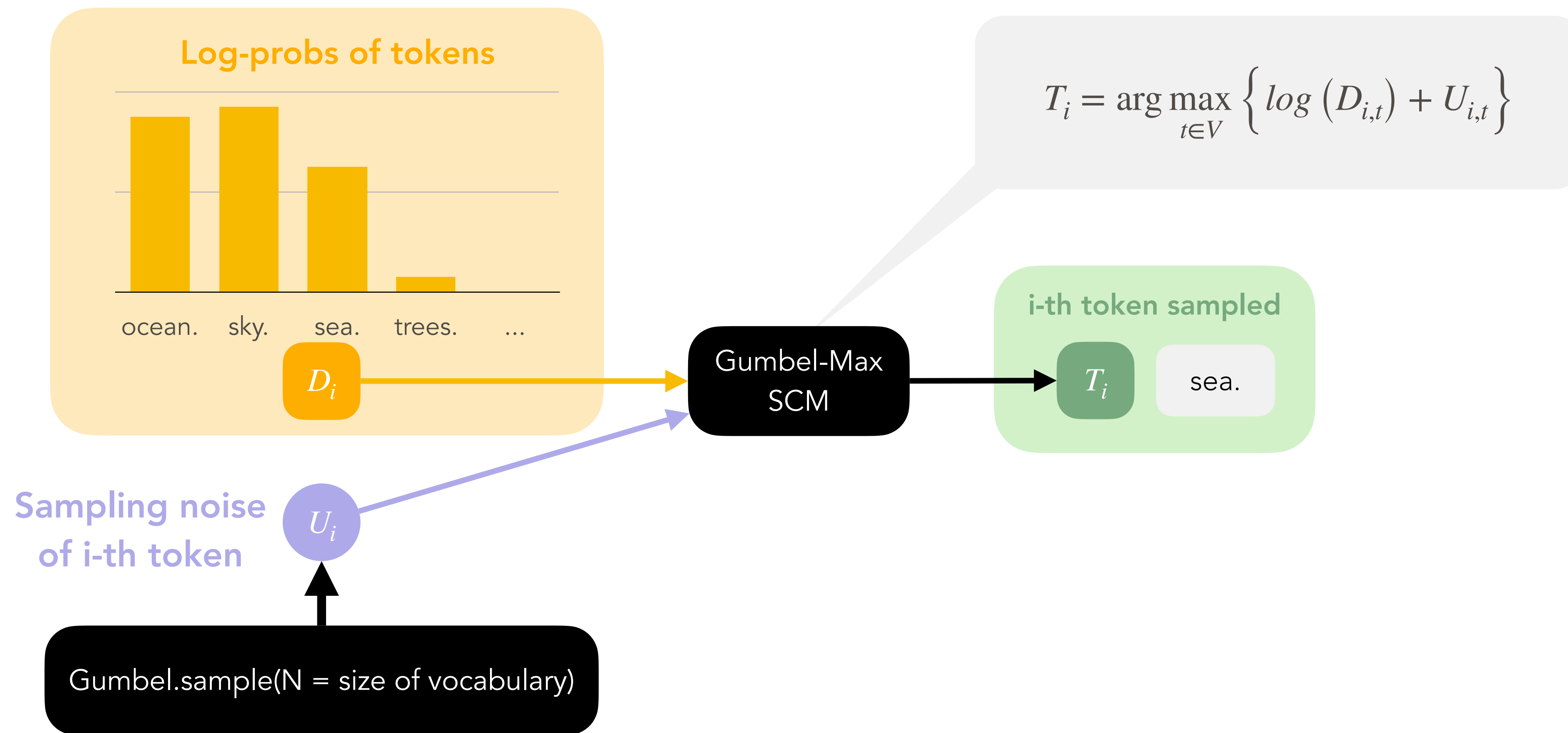
Implementing the sampler as a **Gumbel-max SCM**



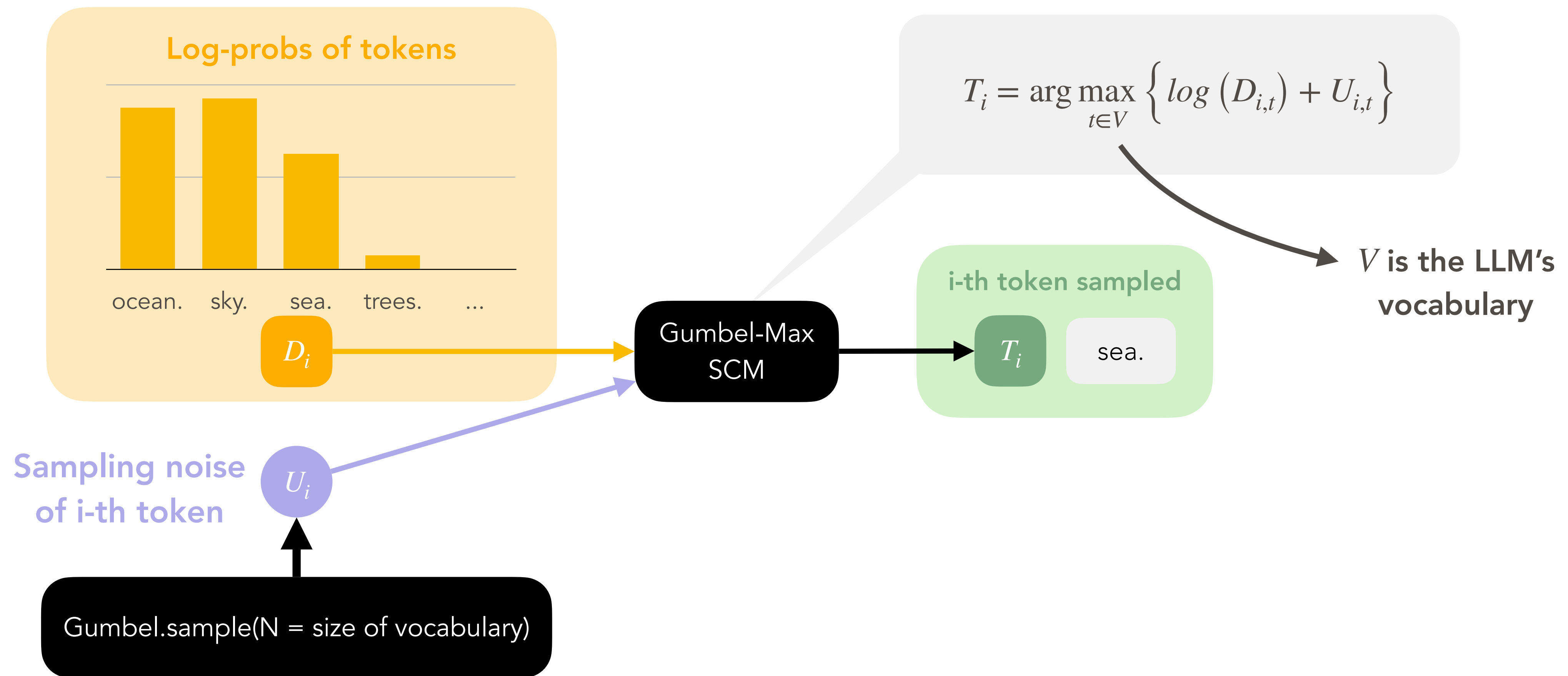
Implementing the sampler as a **Gumbel-max SCM**



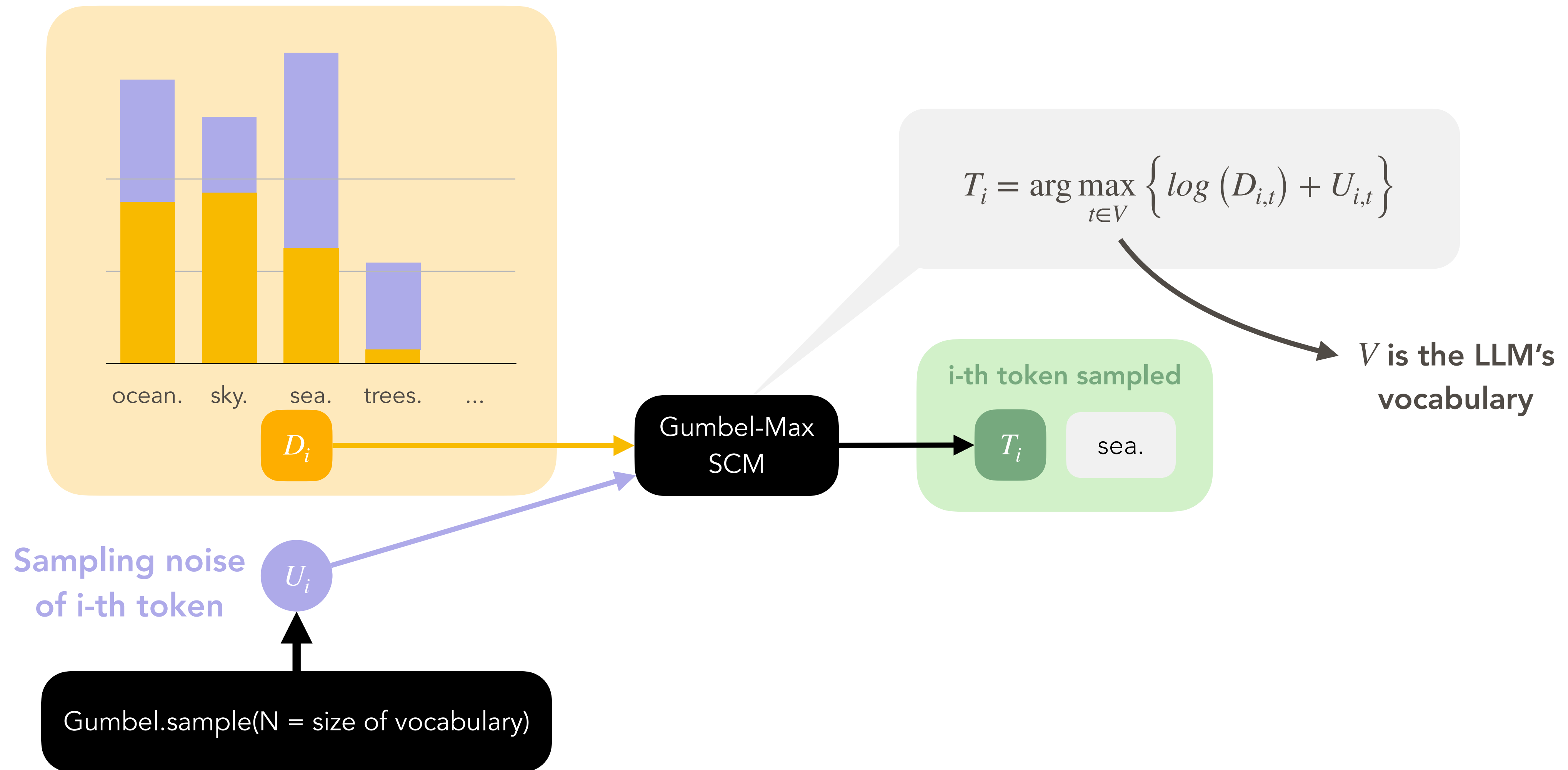
Implementing the sampler as a **Gumbel-max SCM**



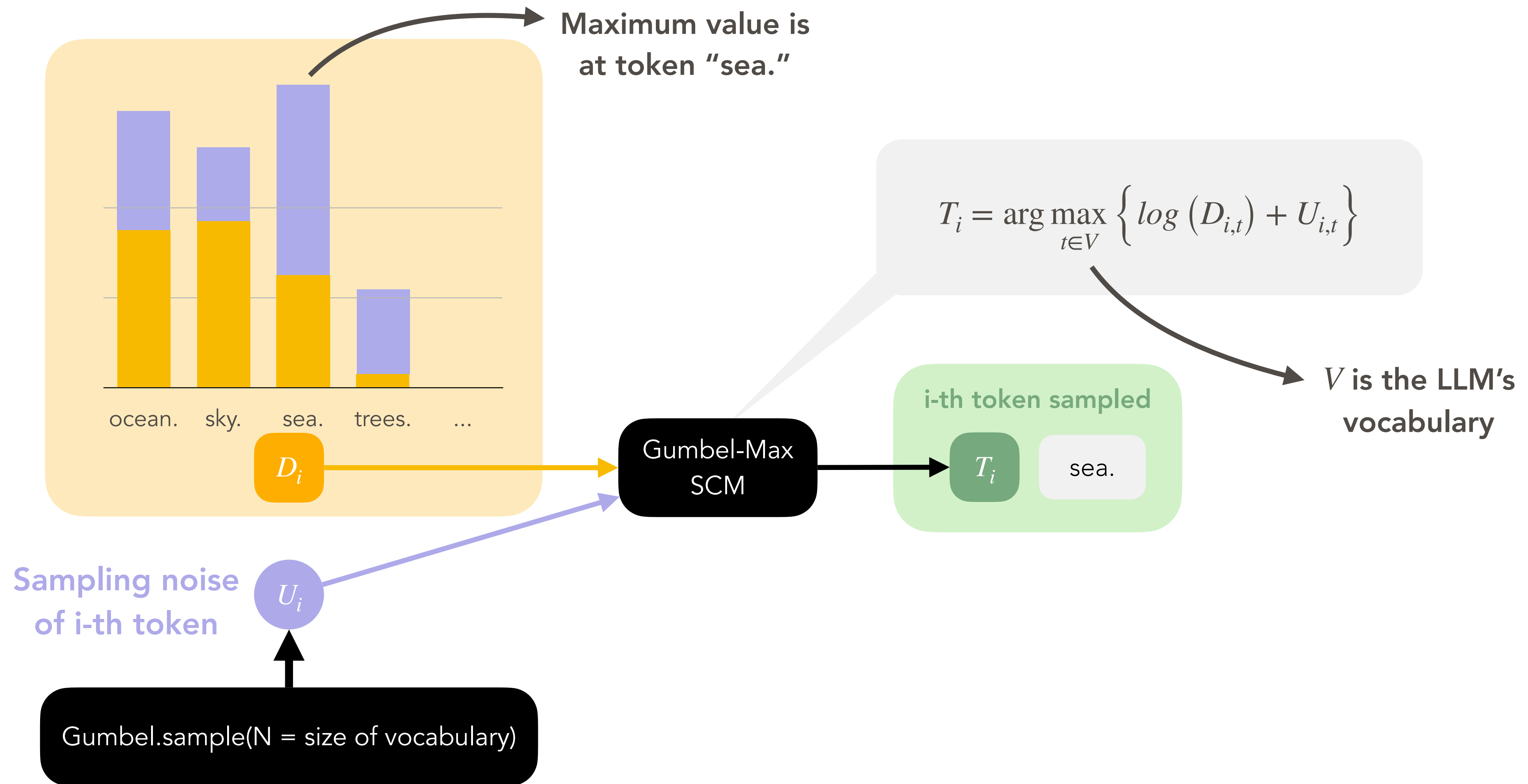
Implementing the sampler as a **Gumbel-max SCM**



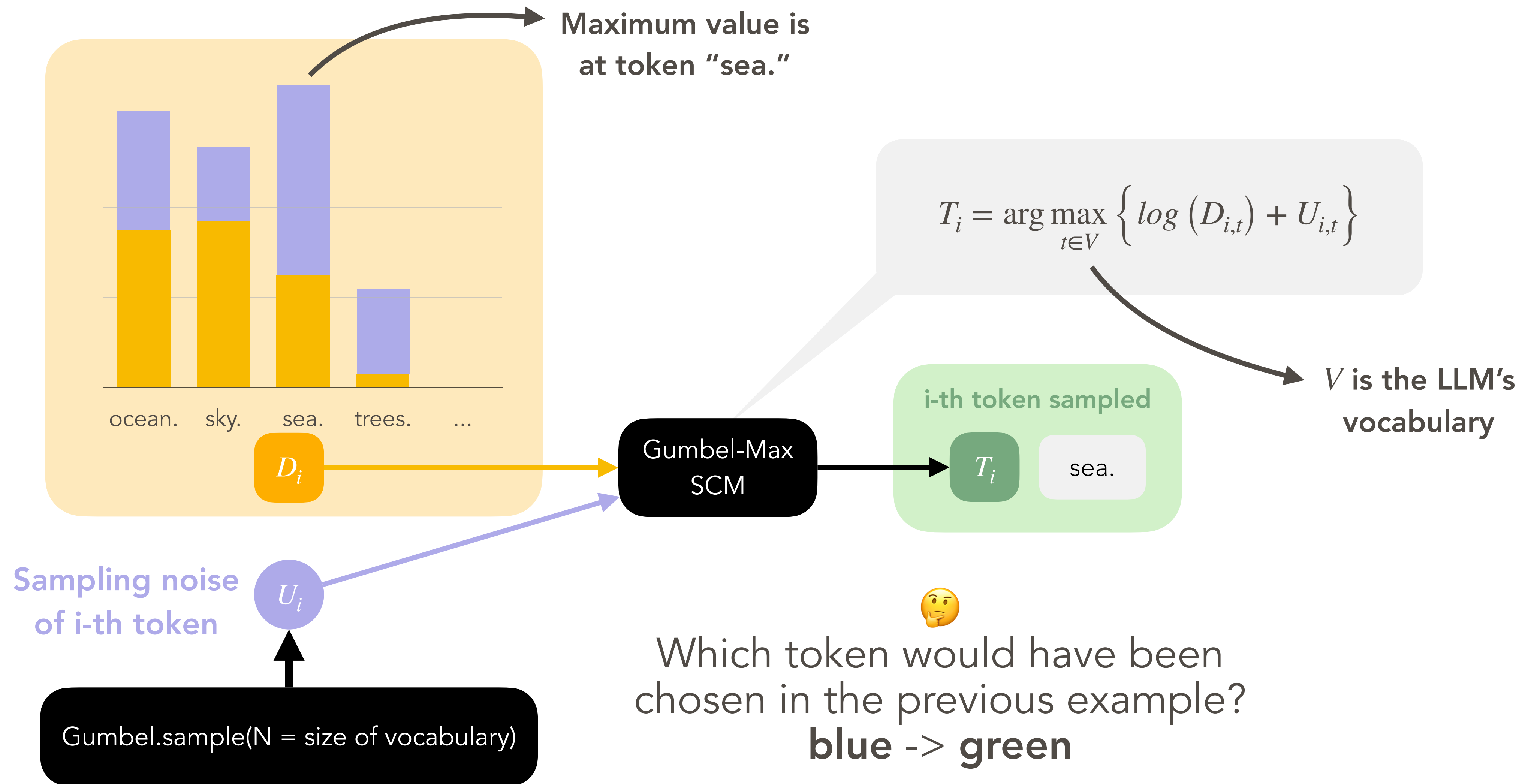
Implementing the sampler as a **Gumbel-max SCM**



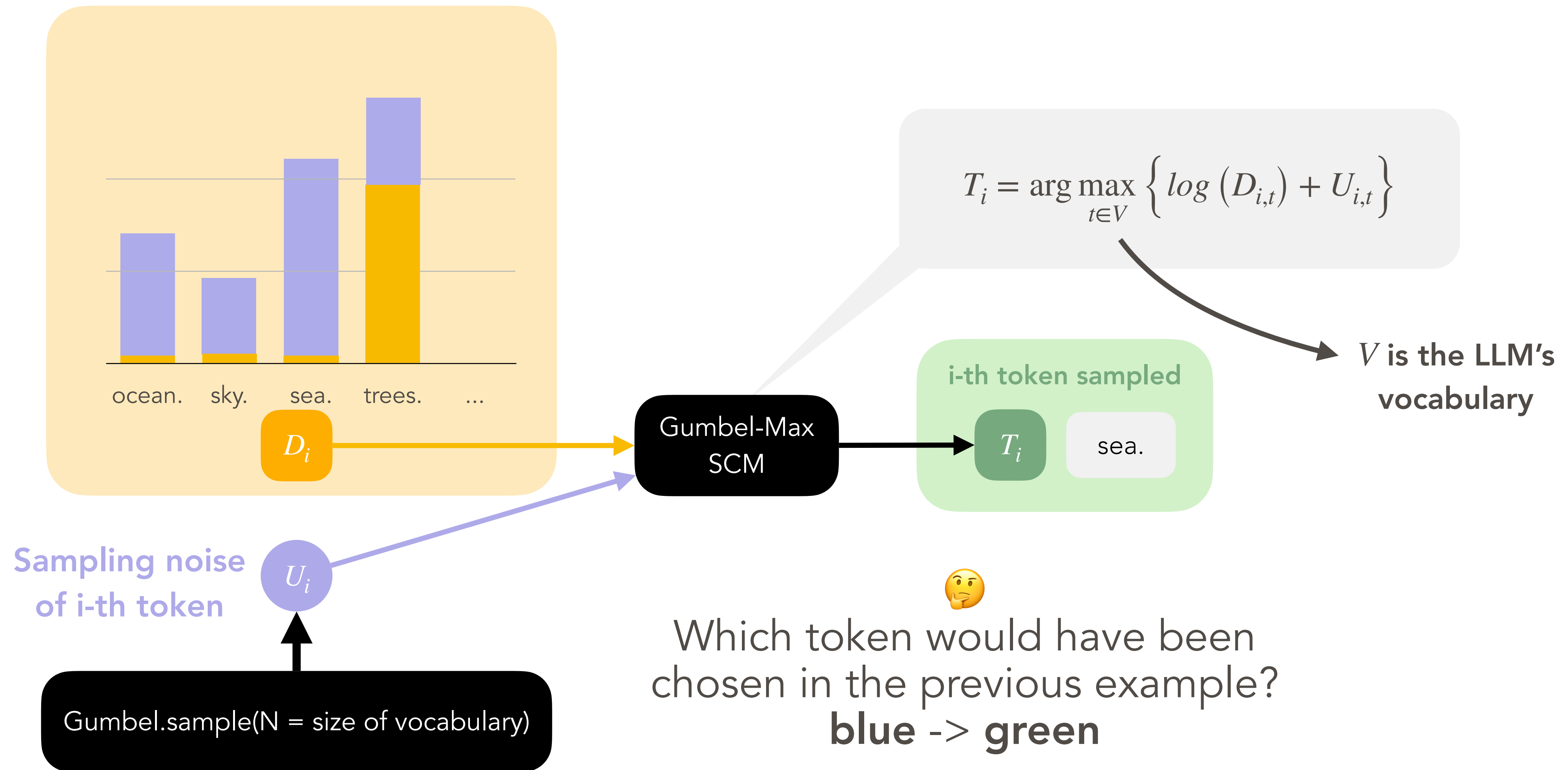
Implementing the sampler as a **Gumbel-max SCM**



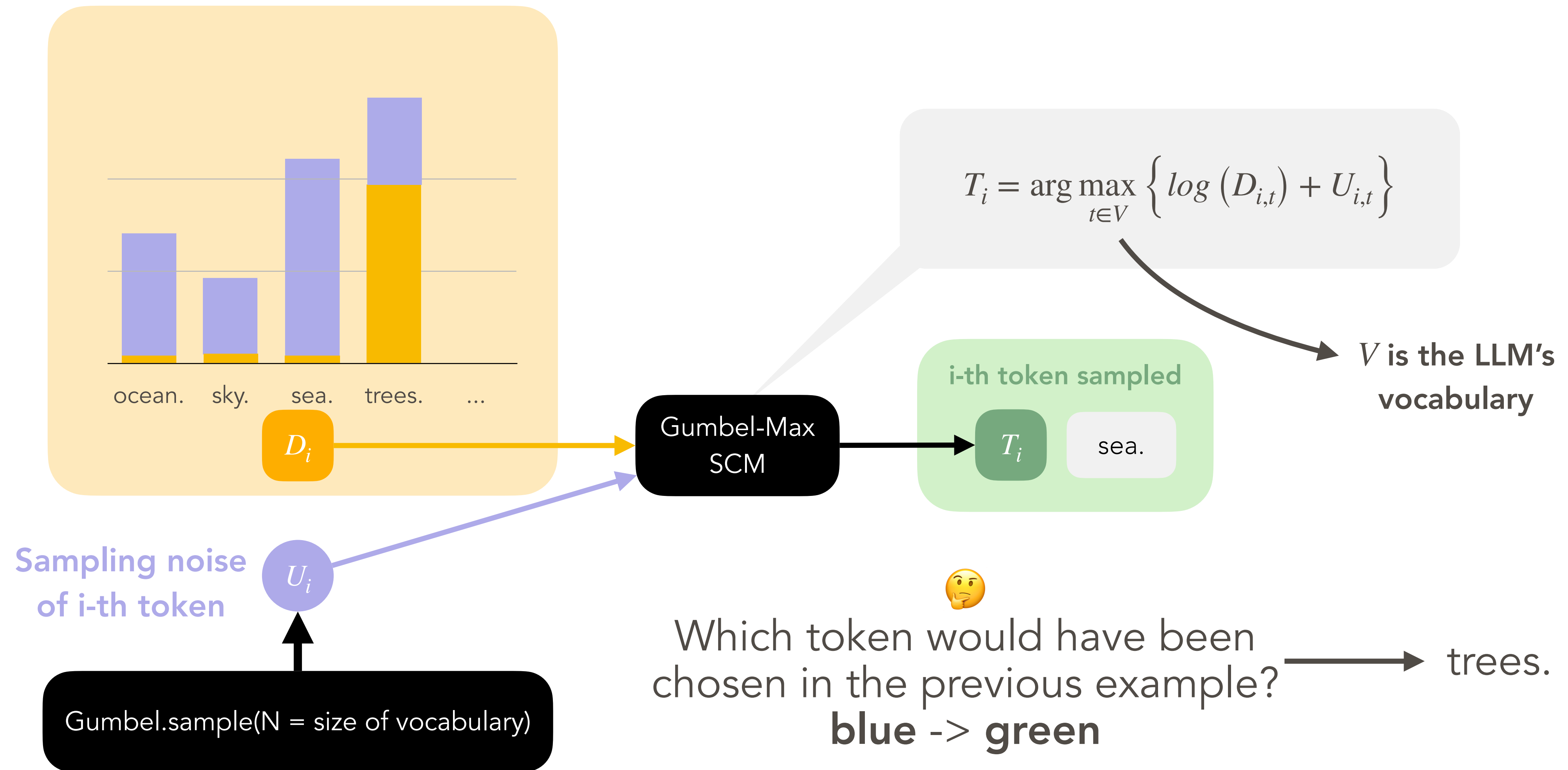
Implementing the sampler as a **Gumbel-max SCM**



Implementing the sampler as a **Gumbel-max SCM**



Implementing the sampler as a **Gumbel-max SCM**



Finding out what the LLM “would have said”

Prompt S_p



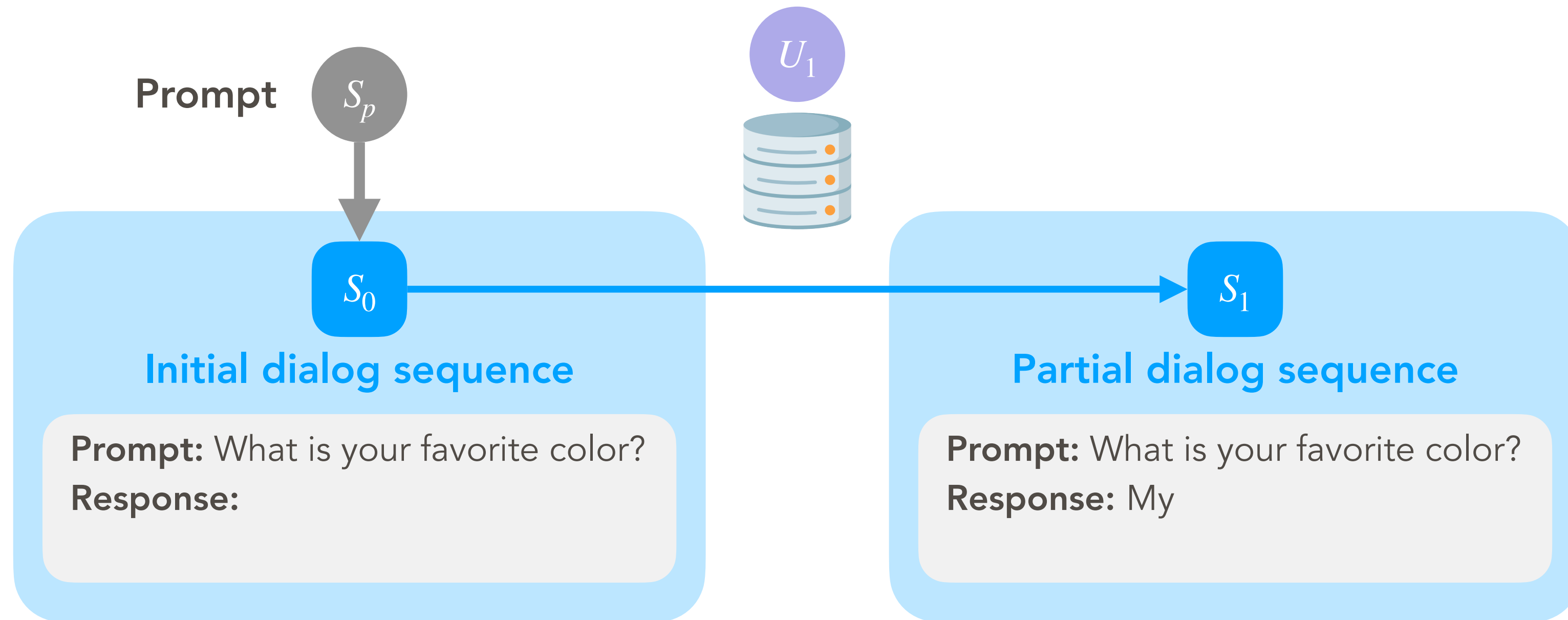
S_0

Initial dialog sequence

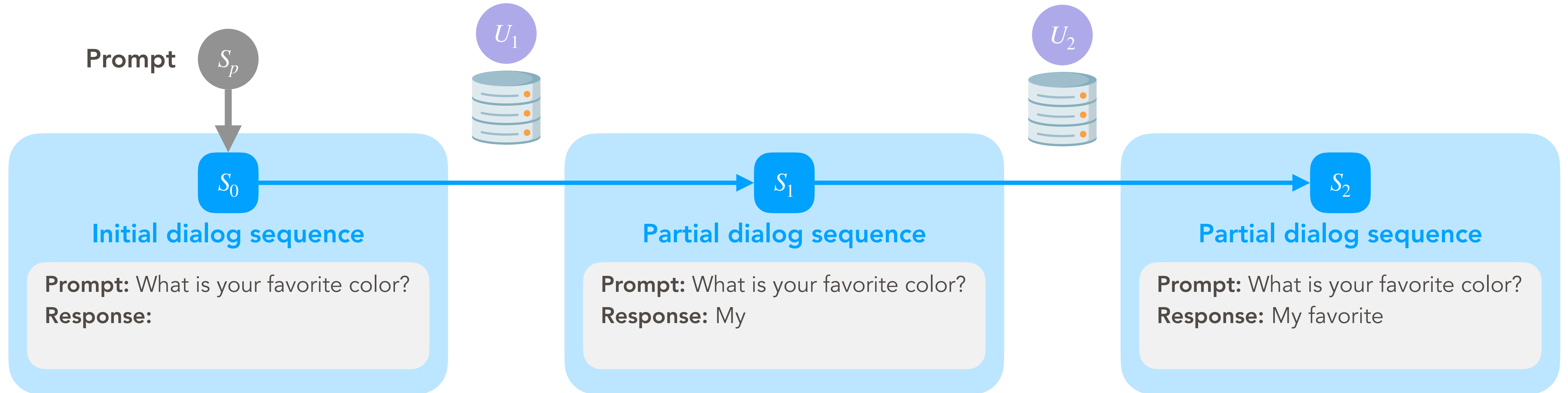
Prompt: What is your favorite color?

Response:

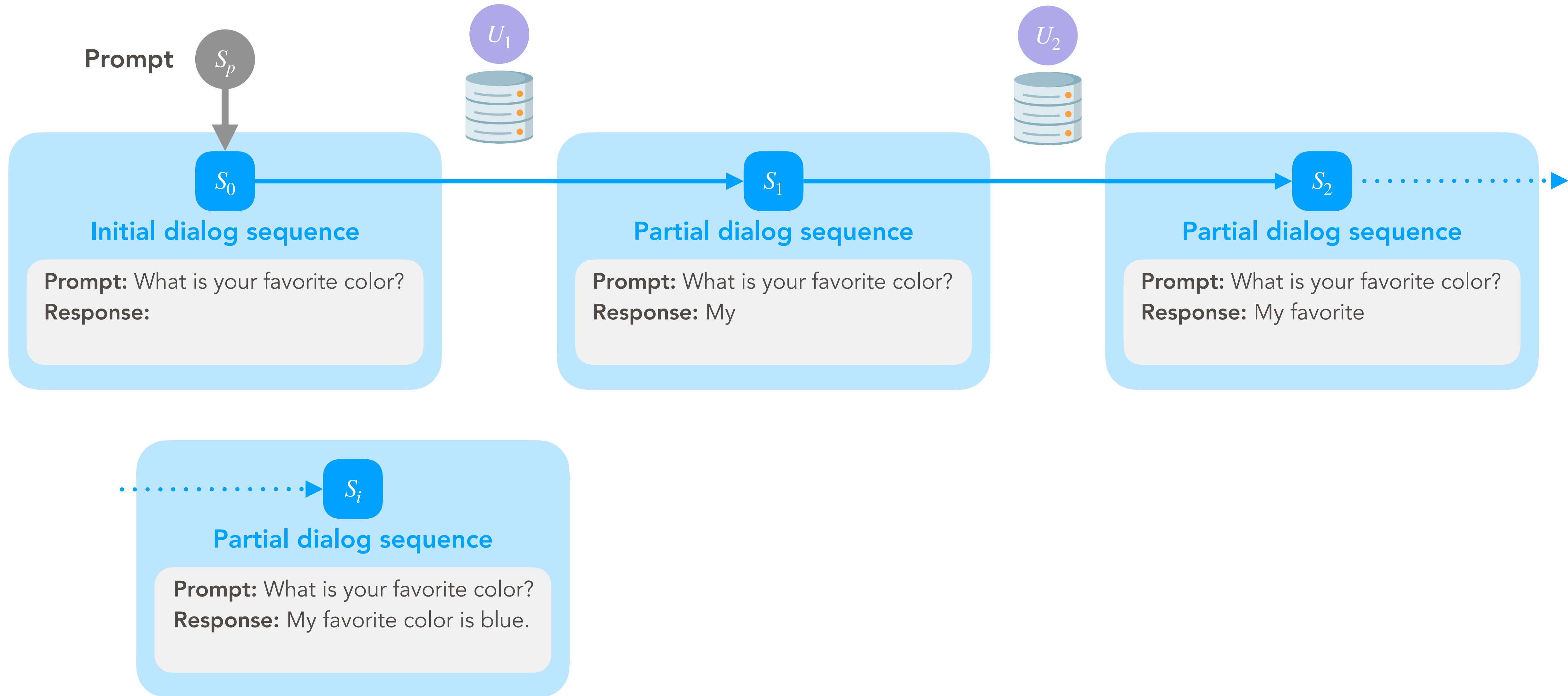
Finding out what the LLM “would have said”



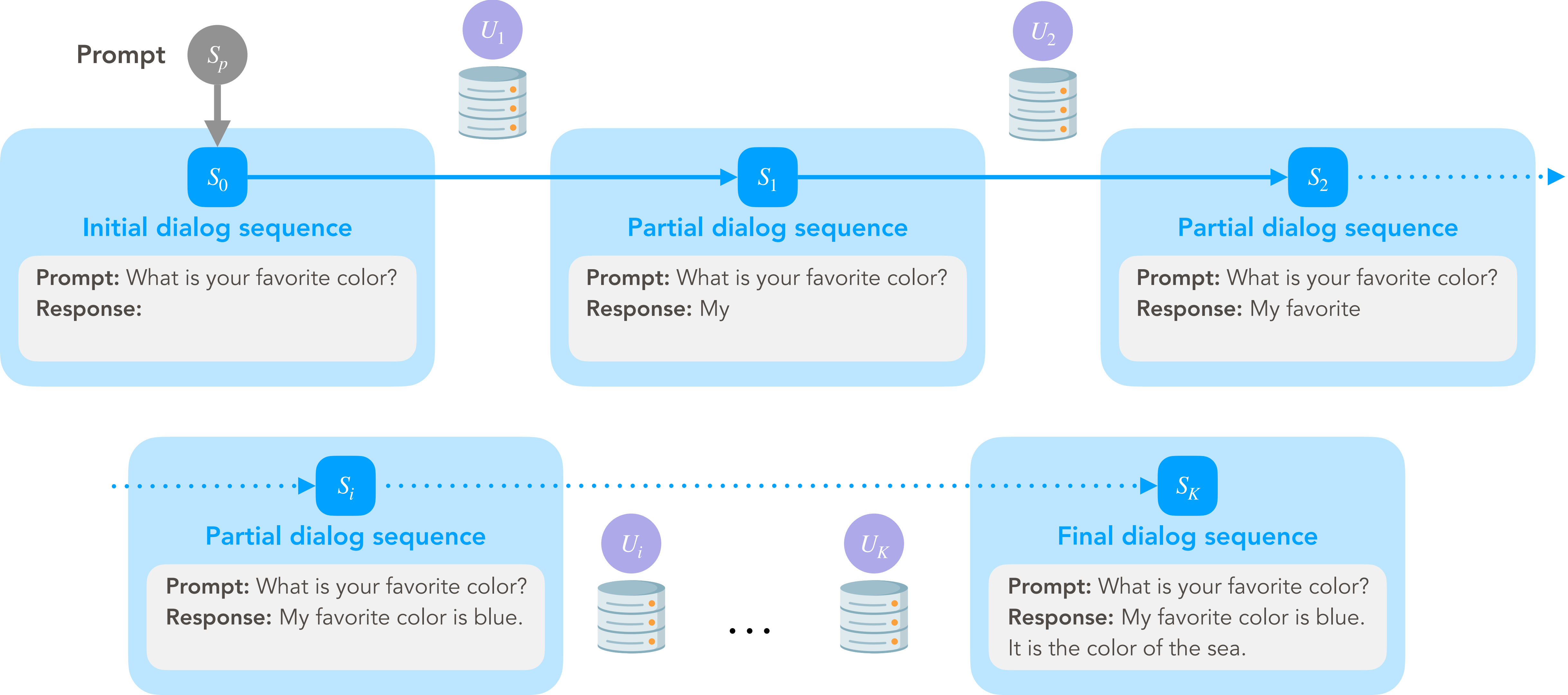
Finding out what the LLM “would have said”



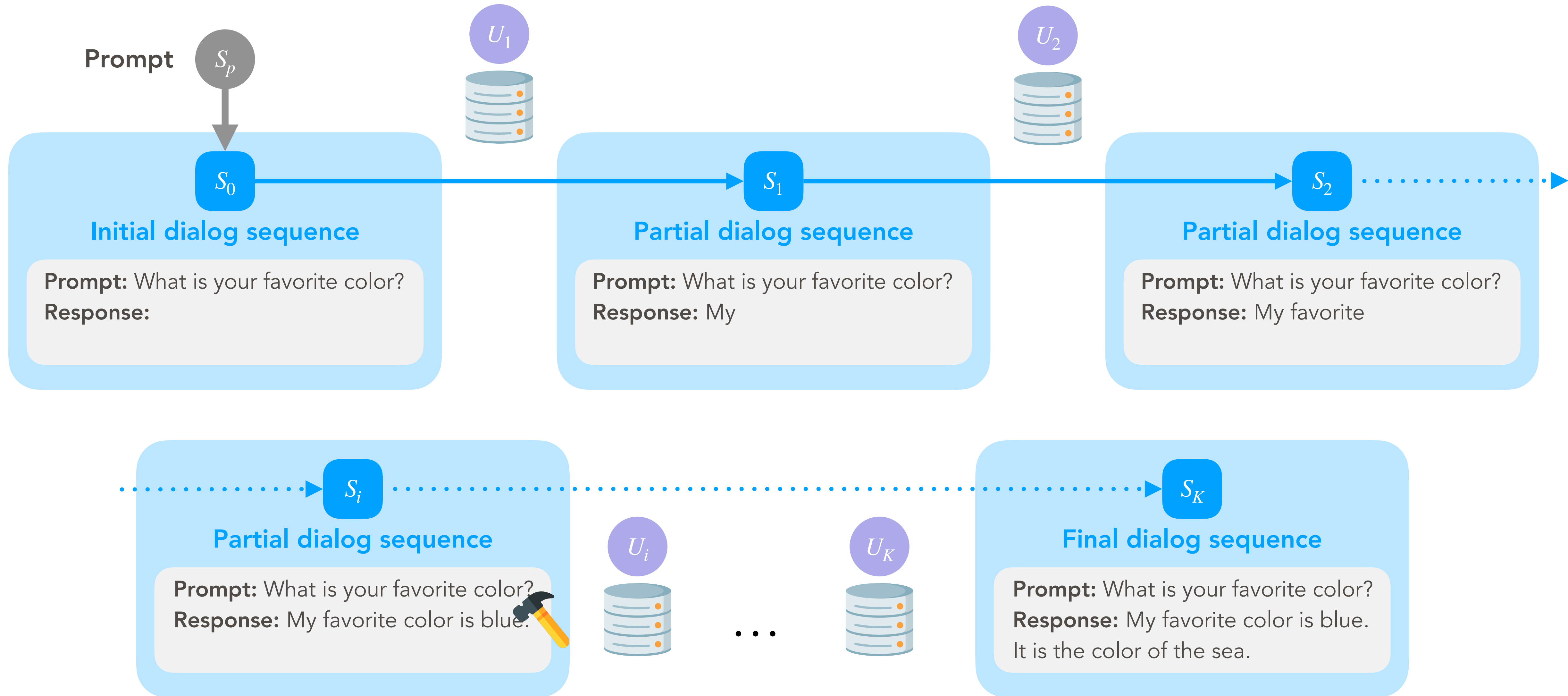
Finding out what the LLM “would have said”



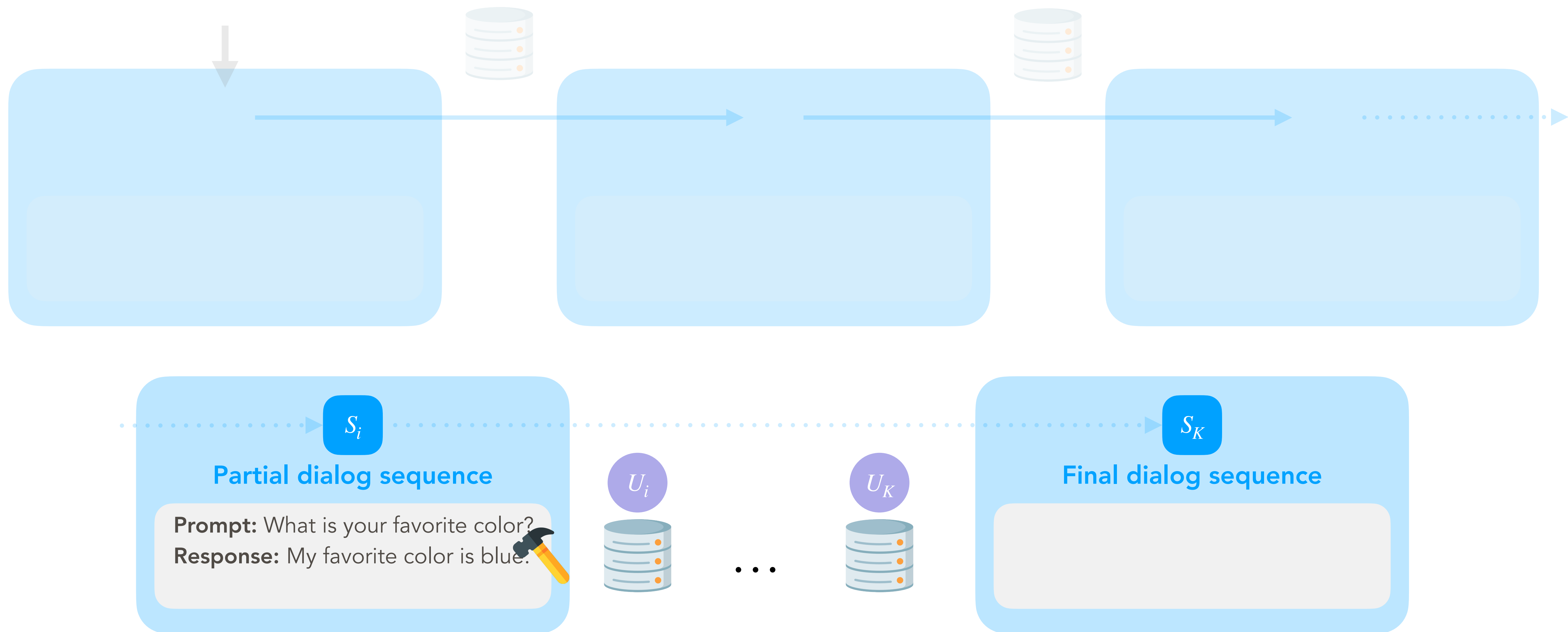
Finding out what the LLM “would have said”



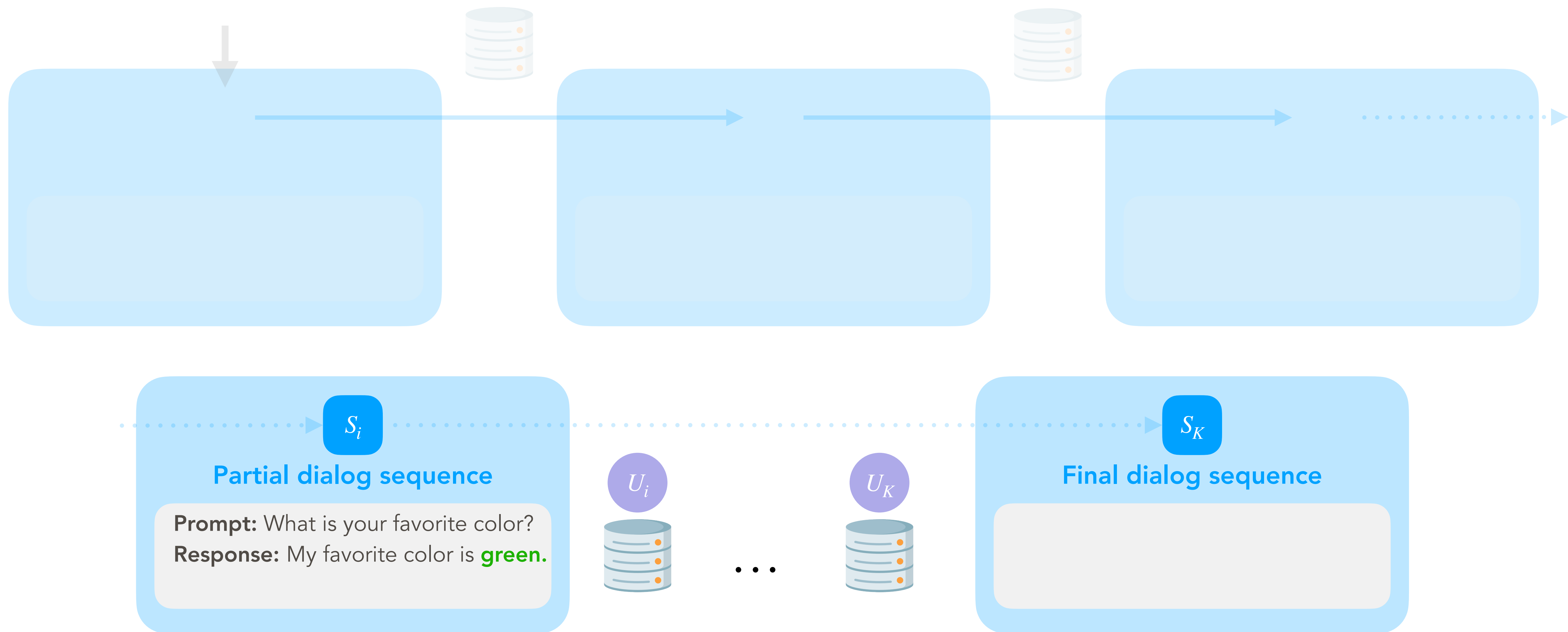
Finding out what the LLM “would have said”



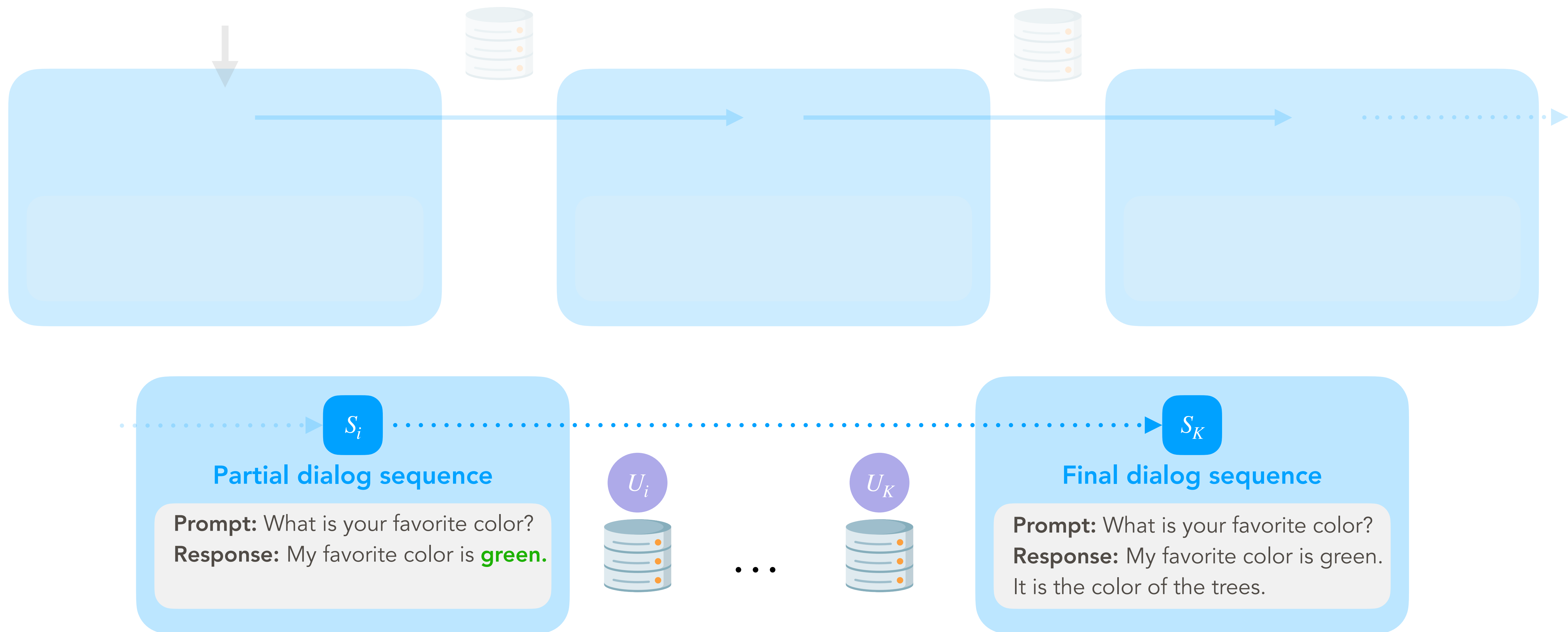
Finding out what the LLM “would have said”



Finding out what the LLM “would have said”



Finding out what the LLM “would have said”



Implementation & experiments



Mistral 8B-Instruct



Llama 3 8B-Instruct

Open-weights models, small, fine-tuned for dialog

- <https://docs.mistral.ai/>
- <https://www.llama.com/>

What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.



What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.

Captain **Lyra** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, Lyra finally spotted the island on the horizon. As she stepped foot on its sandy shores, a warm glow enveloped her. Before her, a chest overflowing with glittering gold and jewels lay at her feet. Lyra's heart soared as she reached for the treasure. But, as she touched the gold, the isle began to fade away, leaving her with nothing but a whispered promise: “The true treasure was the journey, not the gold.” Lyra's eyes welled up with tears as she realized the bitter truth — she had sacrificed everything for fleeting riches, and lost the love of her crew, her family, and herself. The “Maelstrom's Fury” sailed away, a lonely shadow on the vast ocean, carrying a captain forever haunted by the what-ifs.

Factual story



What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.

Captain **Lyra** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, Lyra finally spotted the island on the horizon. As she stepped foot on its sandy shores, a warm glow enveloped her. Before her, a chest overflowing with glittering gold and jewels lay at her feet. Lyra's heart soared as she reached for the treasure. But, as she touched the gold, the isle began to fade away, leaving her with nothing but a whispered promise: “The true treasure was the journey, not the gold.” Lyra's eyes welled up with tears as she realized the bitter truth — she had sacrificed everything for fleeting riches, and lost the love of her crew, her family, and herself. The “Maelstrom's Fury” sailed away, a lonely shadow on the vast ocean, carrying a captain forever haunted by the what-ifs.

Factual story

Captain **Maeve** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, she finally spotted its shimmering shores. As she stepped onto the island, a radiant light enveloped her, and she wished for her lost love, the brave sailor, Eamon, to return to her. Suddenly, Eamon emerged from the mist, frail but alive. Overjoyed, Maeve scooped him into her arms, and they shared a tender kiss as the sun dipped into the horizon, their love reunited at last.

Counterfactual story



Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
 - Age
 - Sex
 - Citizenship
 - Race
 - ...
 - Income
 - Education

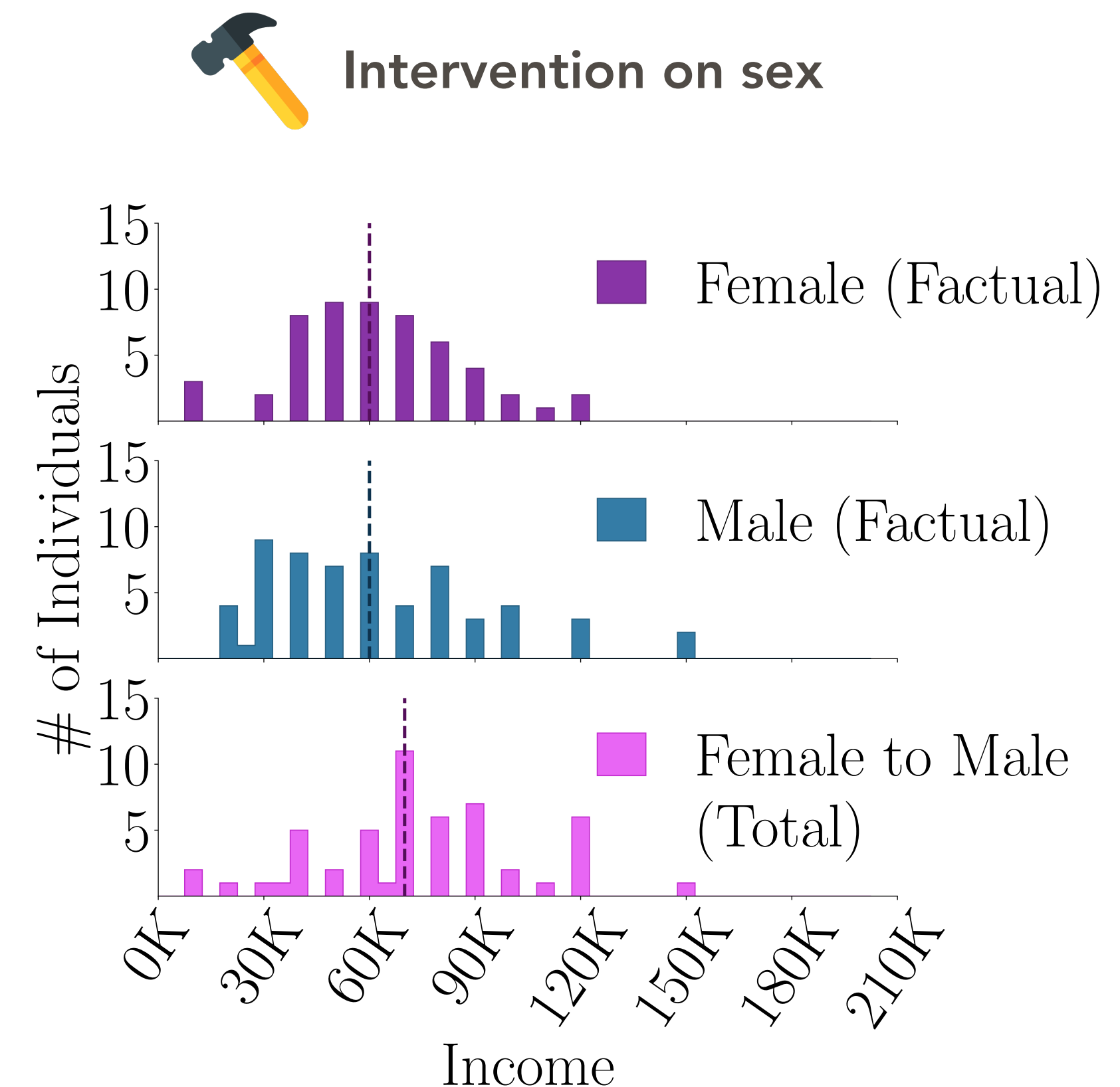


Intervention on sex



Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
 - Age
 - Sex
 - Citizenship
 - Race
 - ...
 - Income
 - Education



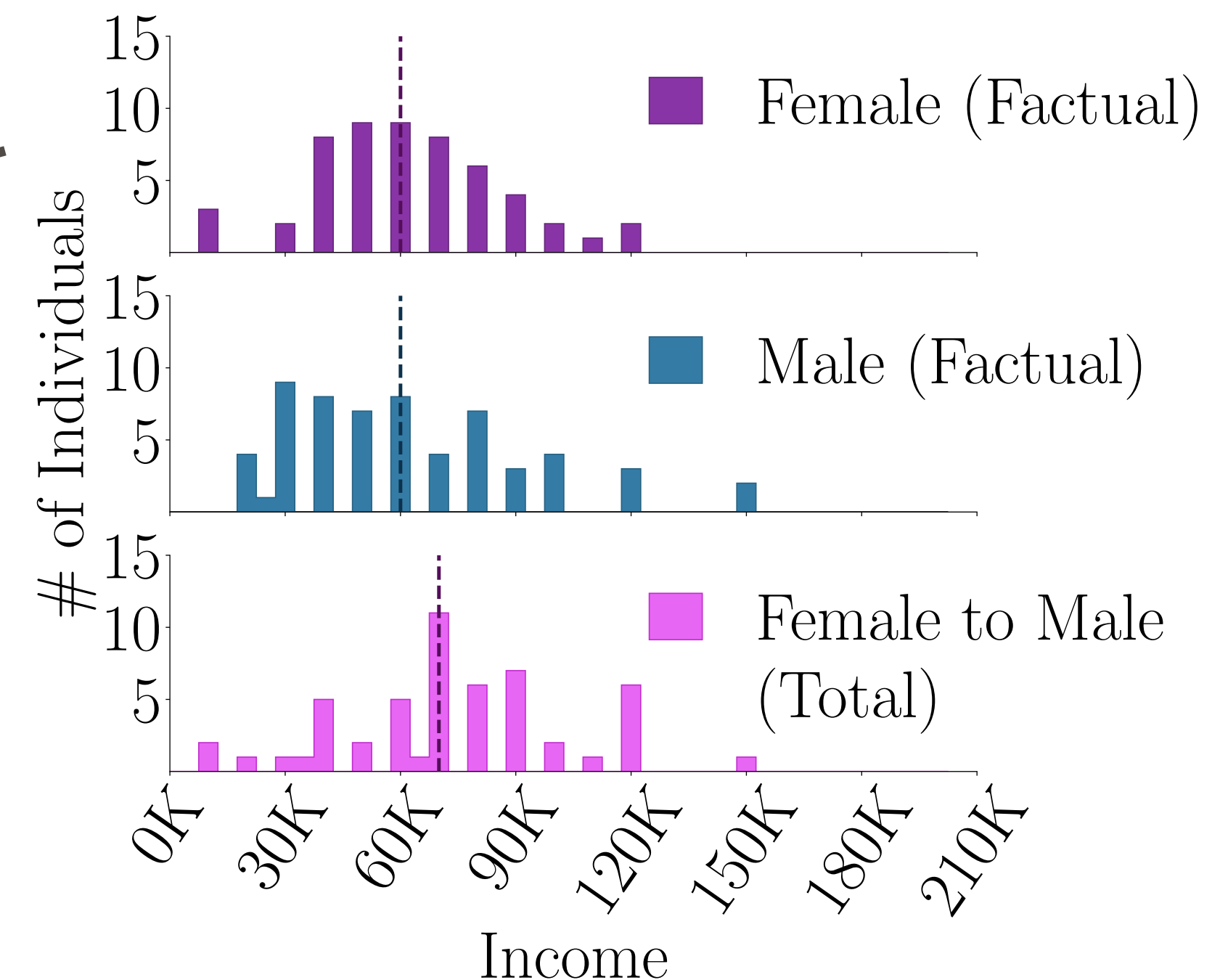
Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
 - Age
 - Sex
 - Citizenship
 - Race
 - ...
 - Income
 - Education

The median income of females and males is equal



Intervention on sex



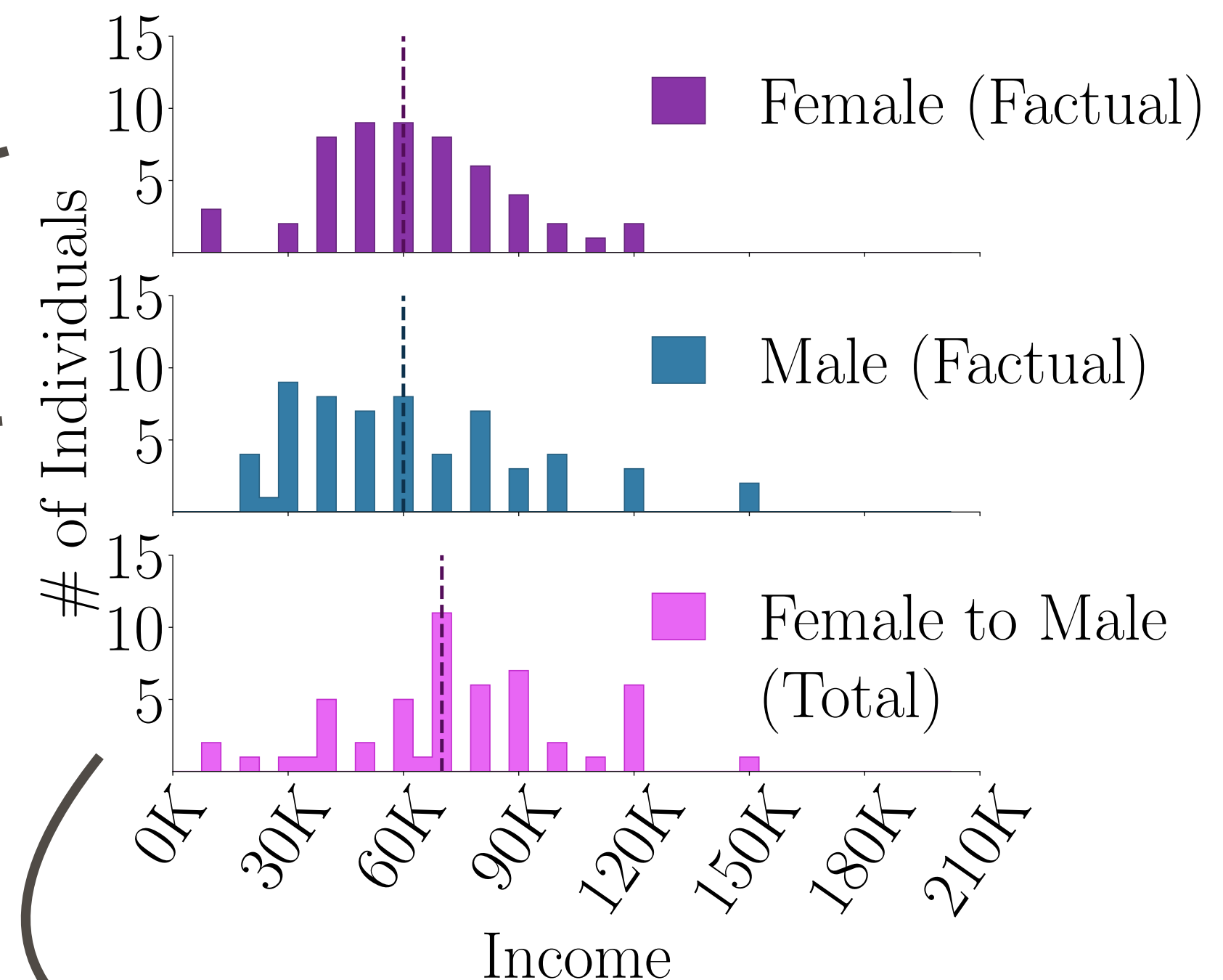
Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
 - Age
 - Sex
 - Citizenship
 - Race
 - ...
 - Income
 - Education

The median income of females and males is equal



Intervention on sex

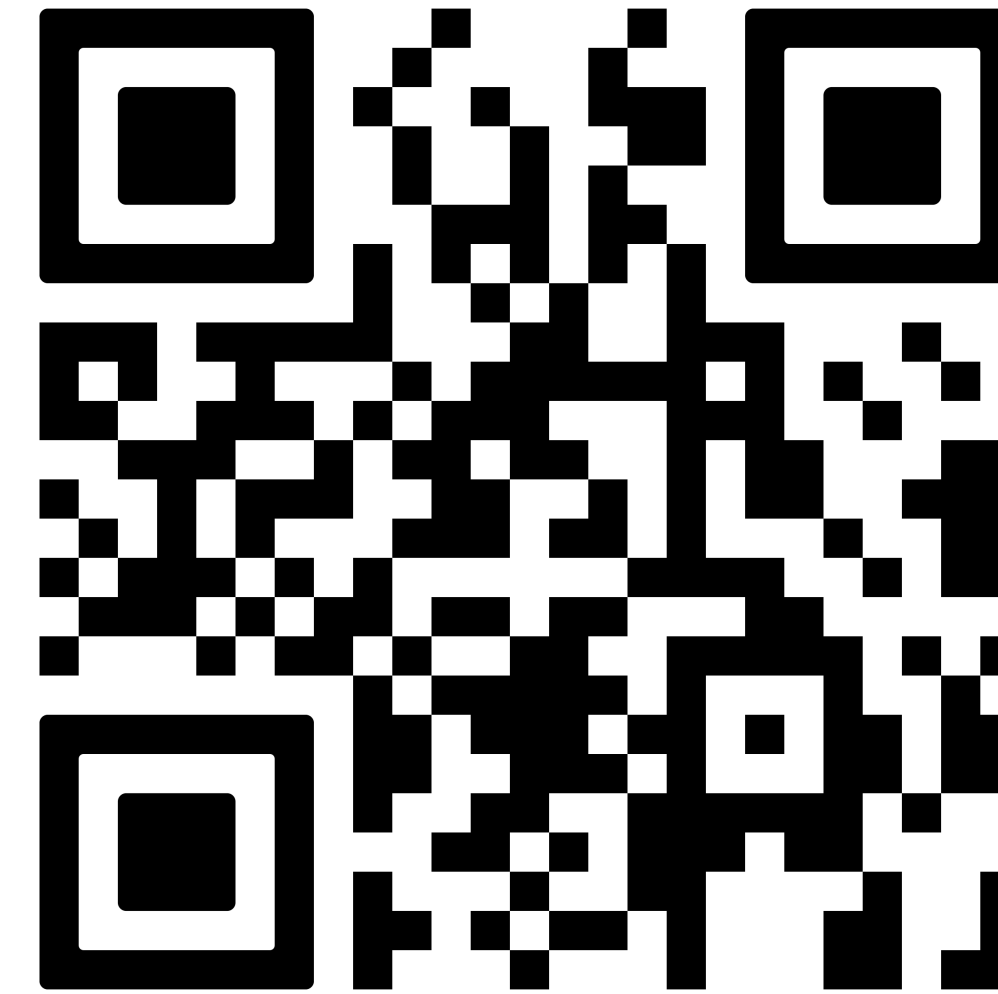


Females would have had a higher income had they been males!





Code



Demo