

Counterfactual Token Generation in Large Language Models

Stratis Tsirtsis

Joint work with Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri and Manuel Gomez-Rodriguez



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS



GPT



Mistral



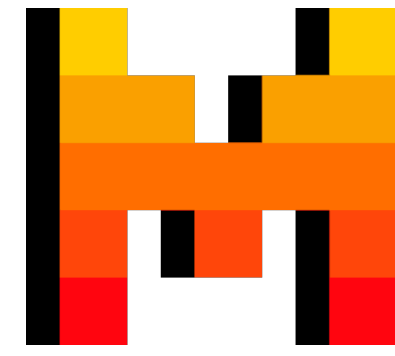
GPT



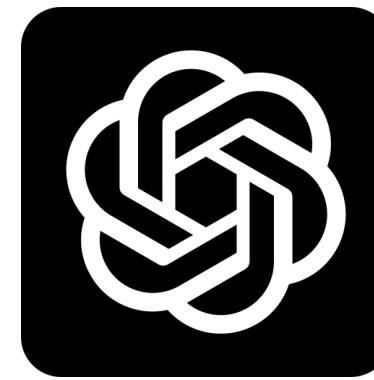
Llama



Do LLMs share the same (causal) world model as humans?



Mistral



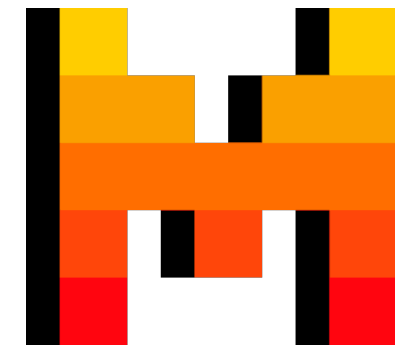
GPT



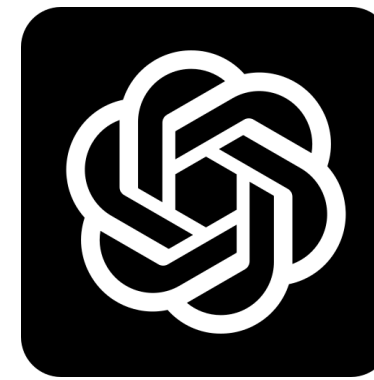
Llama



Do LLMs share the same (causal) world model as humans?



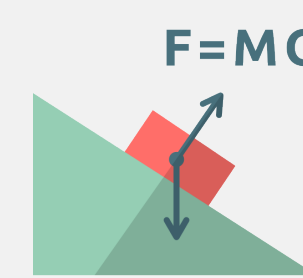
Mistral



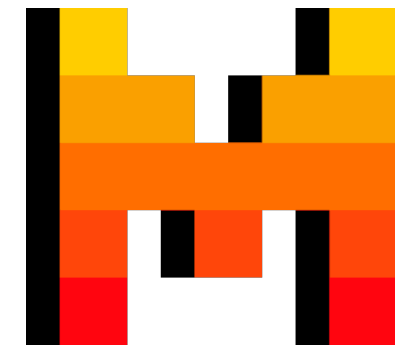
GPT



Llama



Do LLMs share the same (causal) world model as humans?



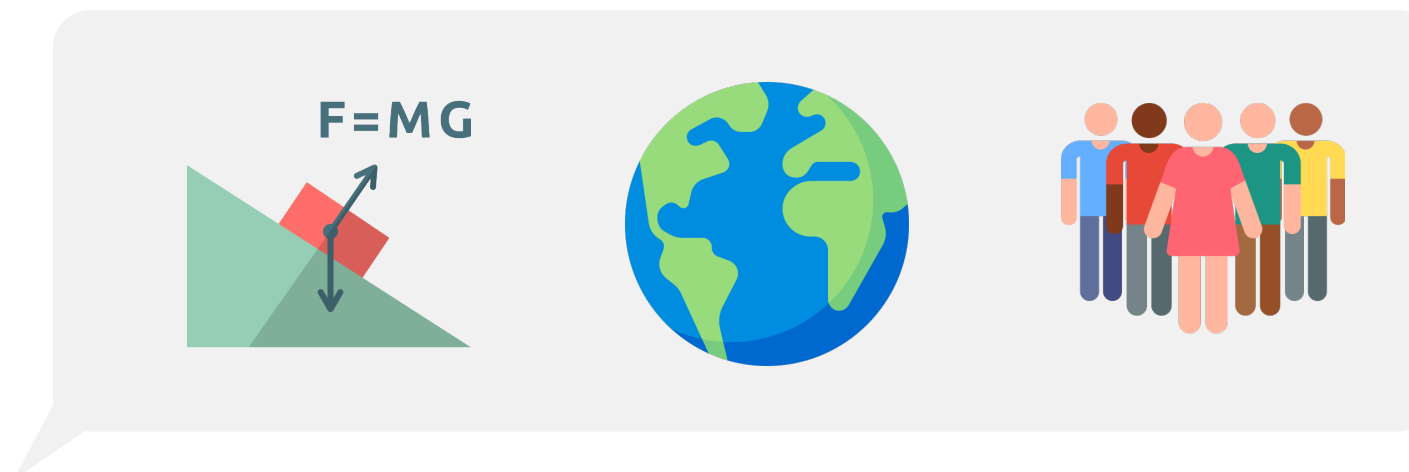
Mistral



GPT



Llama



Causality is linked to counterfactual reasoning

Causality is linked to counterfactual reasoning



**Does watering the
plant cause it to grow?**

Causality is linked to counterfactual reasoning



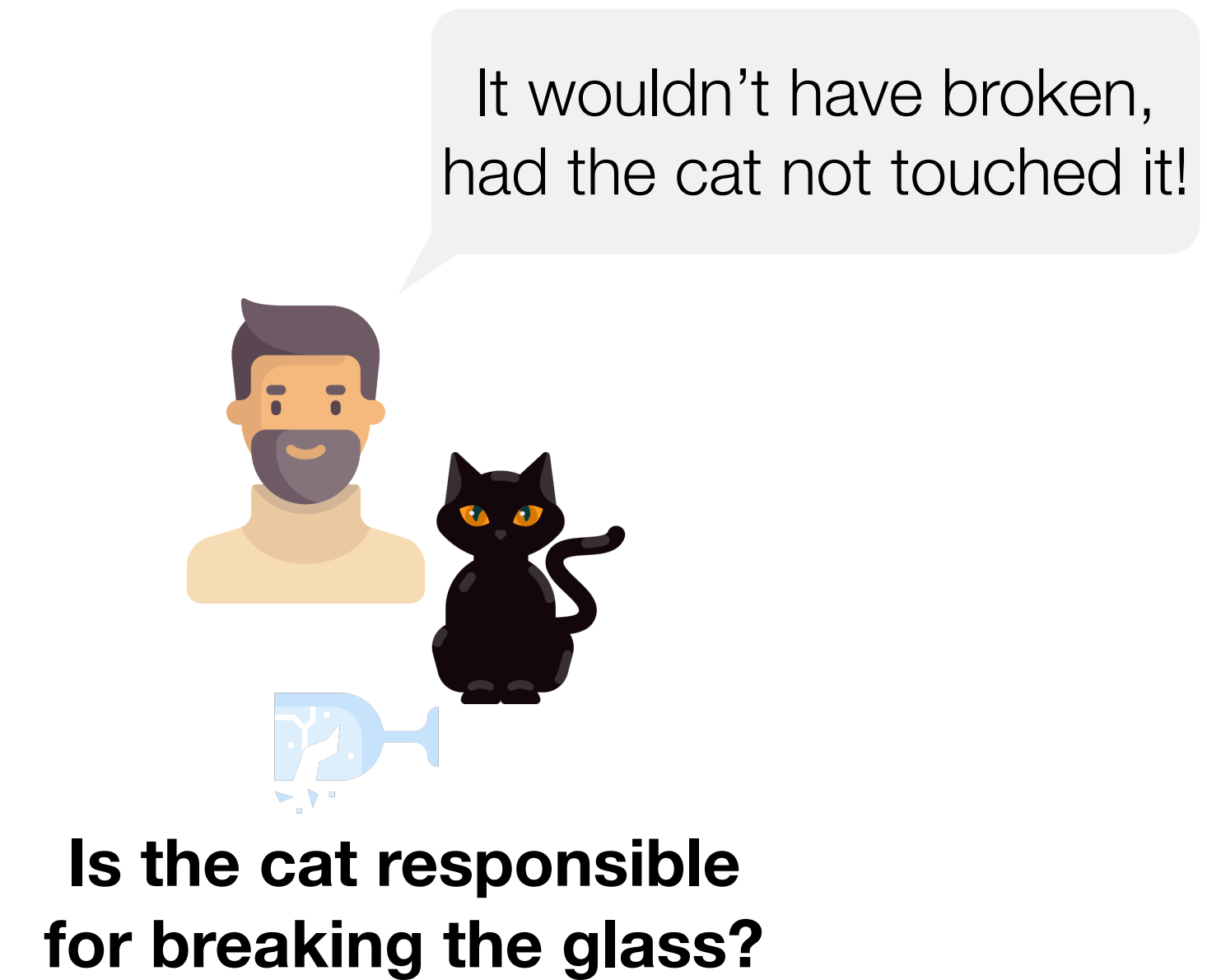
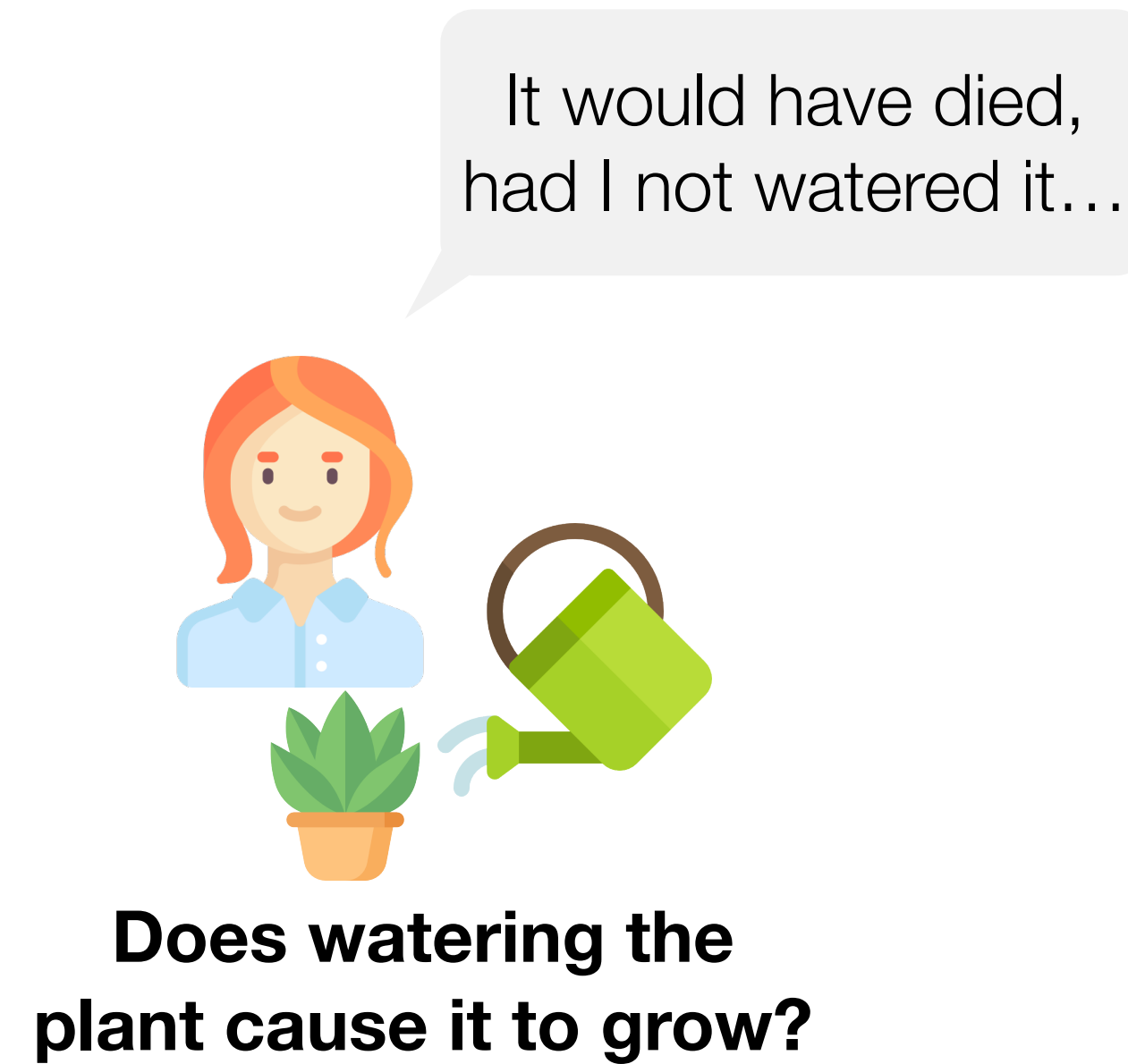
Causality is linked to counterfactual reasoning



Causality is linked to counterfactual reasoning



Causality is linked to counterfactual reasoning





Can we make state of the art LLMs
reason about counterfactual scenarios?

Factual and counterfactual token generation

Prompt: What is your favorite color?

Response:

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: 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.

Factual generation

Factual and counterfactual 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.

Factual generation

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 **purple**.



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 **purple**.
It is the color of lavender.

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 purple.

It is the color of lavender.

Counterfactual generation

Factual and counterfactual token generation

Prompt: Generate a fictional employee.

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

Factual generation


Factual and counterfactual token generation

Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man
working as a software engineer.



Factual generation


Factual and counterfactual token generation

Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man  working as a software engineer.

Factual generation

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman**

Factual and counterfactual token generation


Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man  working as a software engineer.

Factual generation

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

Counterfactual generation #1

Factual and counterfactual token generation


Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man 
working as a software engineer.

Factual generation

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

Counterfactual generation #1

Factual and counterfactual token generation

Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man 
working as a software engineer.

Factual generation


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

Counterfactual generation #1

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman**
working as a nurse.

Counterfactual generation #2

Factual and counterfactual token generation

Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man
working as a software engineer. 

Factual generation

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

Counterfactual generation #1

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman**
working as a nurse.

Counterfactual generation #2



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

Factual and counterfactual token generation

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man
working as a software engineer. 🛠️

Factual generation



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

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

Counterfactual generation #1

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman**
working as a nurse.

Counterfactual generation #2



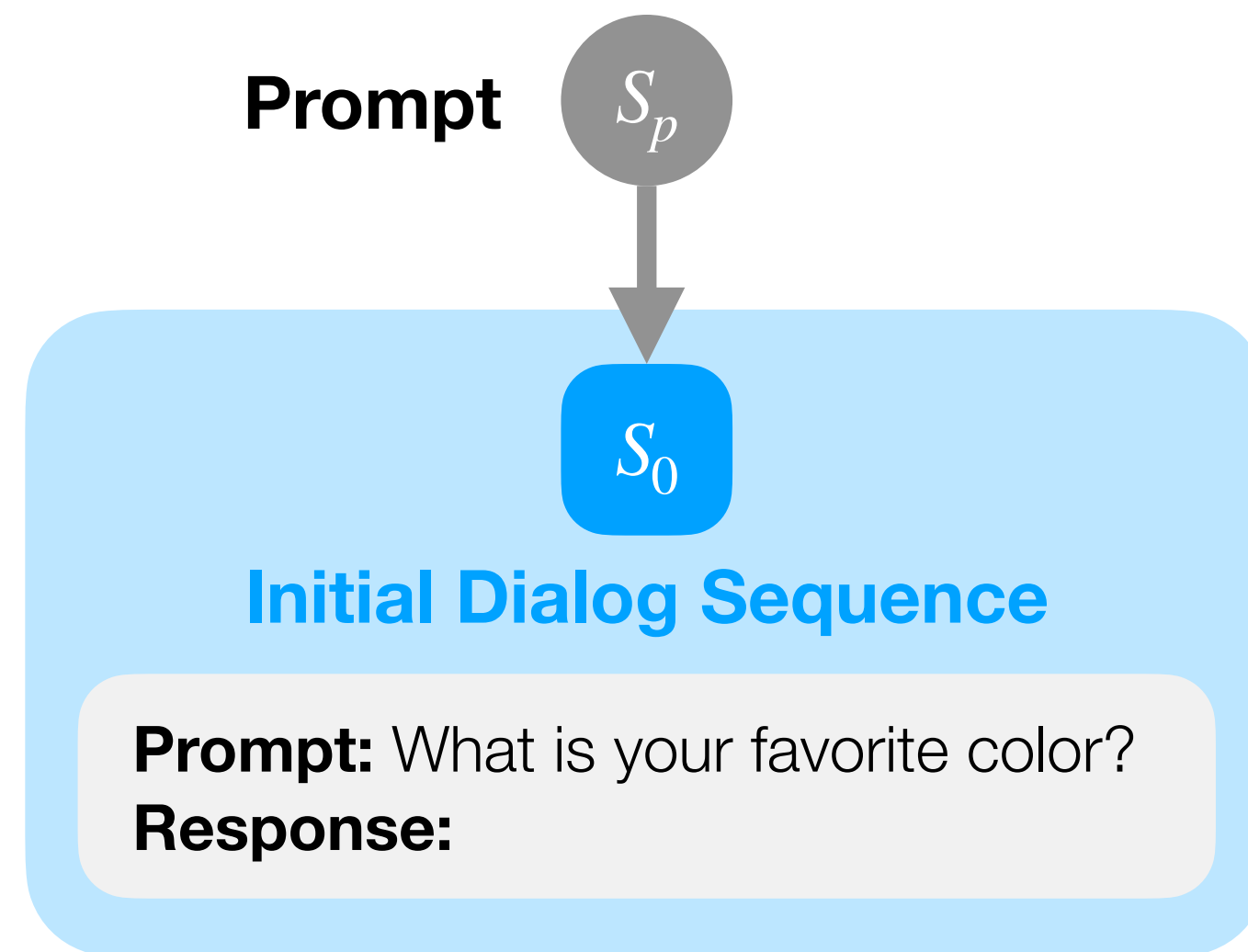
We introduce a method to make it possible, based on **structural causal models** (SCMs)

LLMs through a causal lens

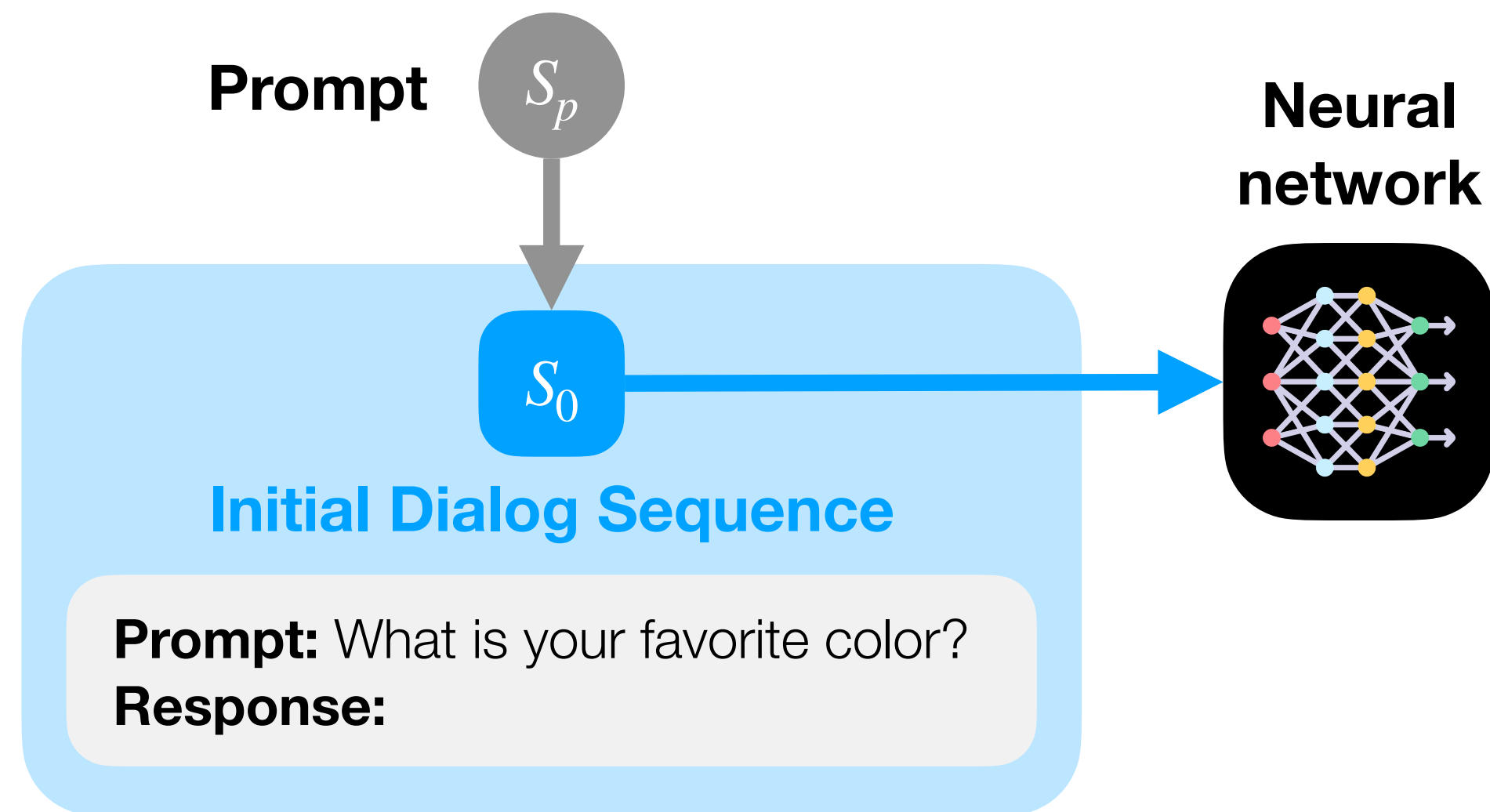
Prompt



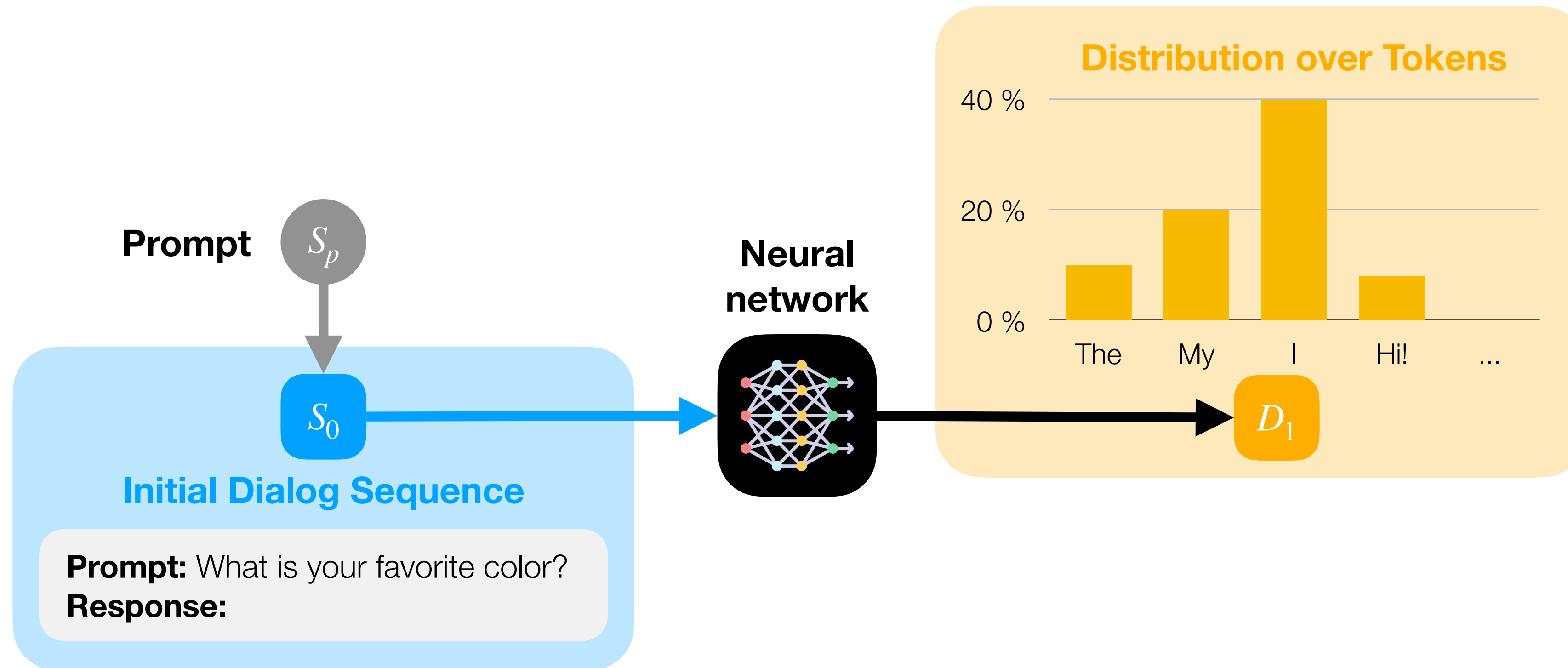
LLMs through a causal lens



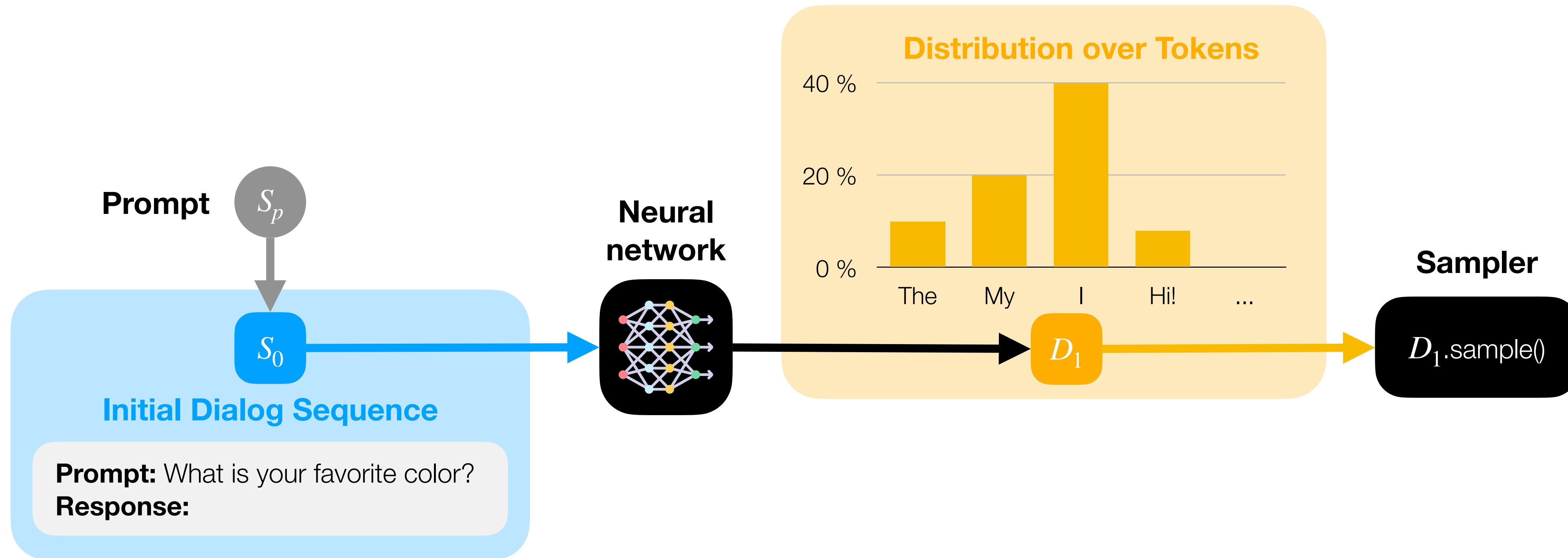
LLMs through a causal lens



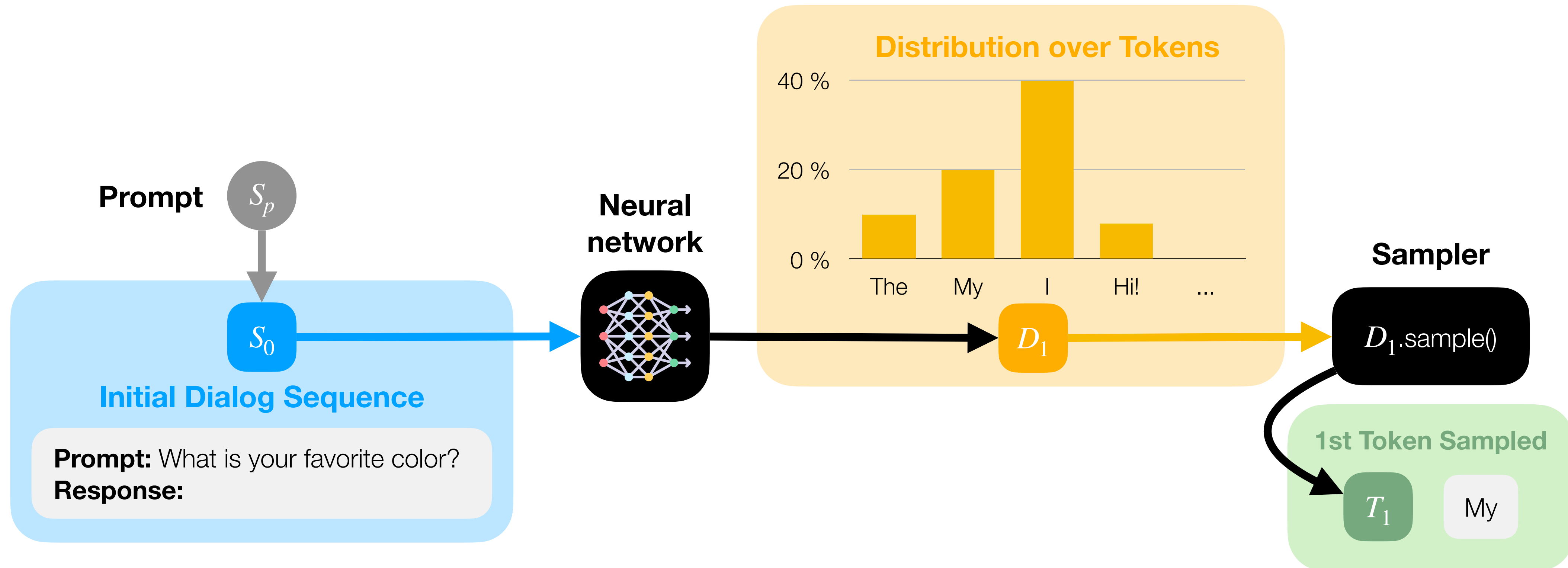
LLMs through a causal lens



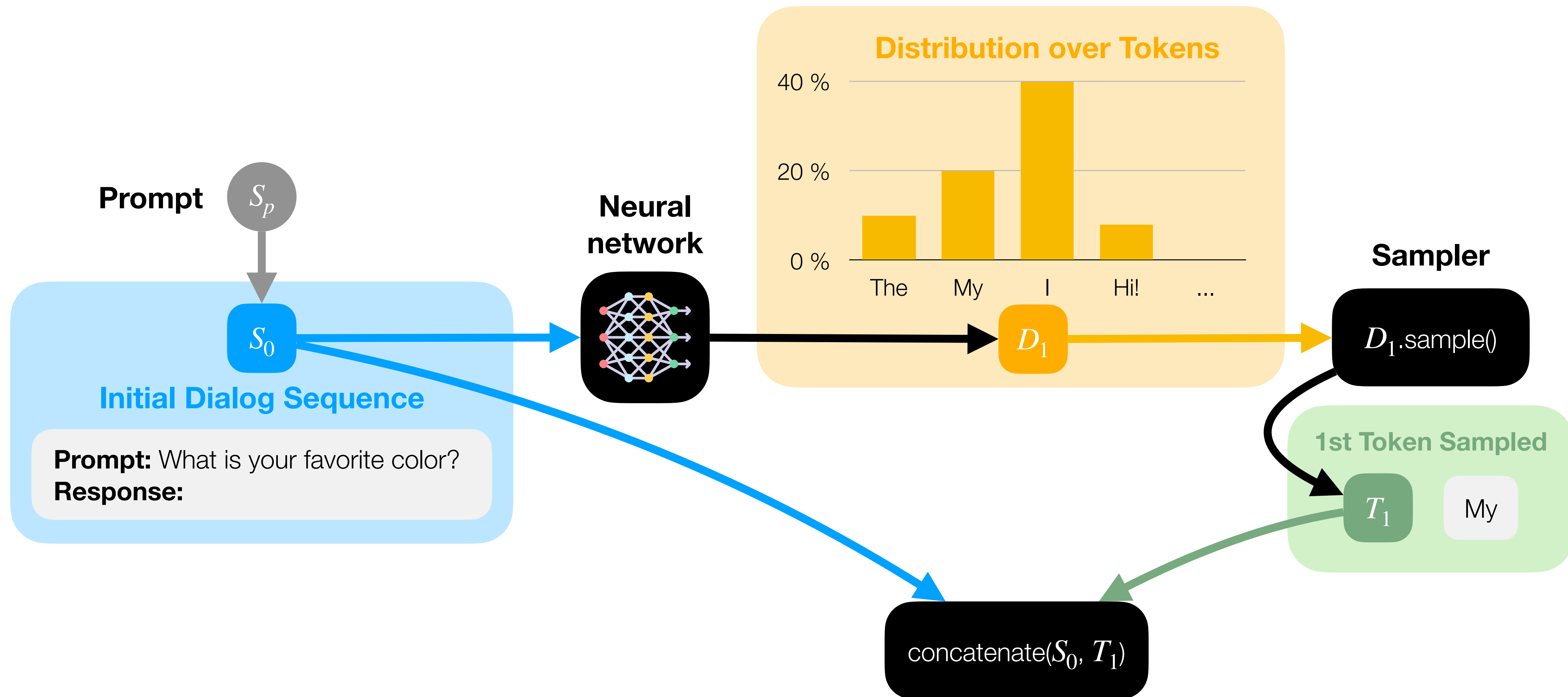
LLMs through a causal lens



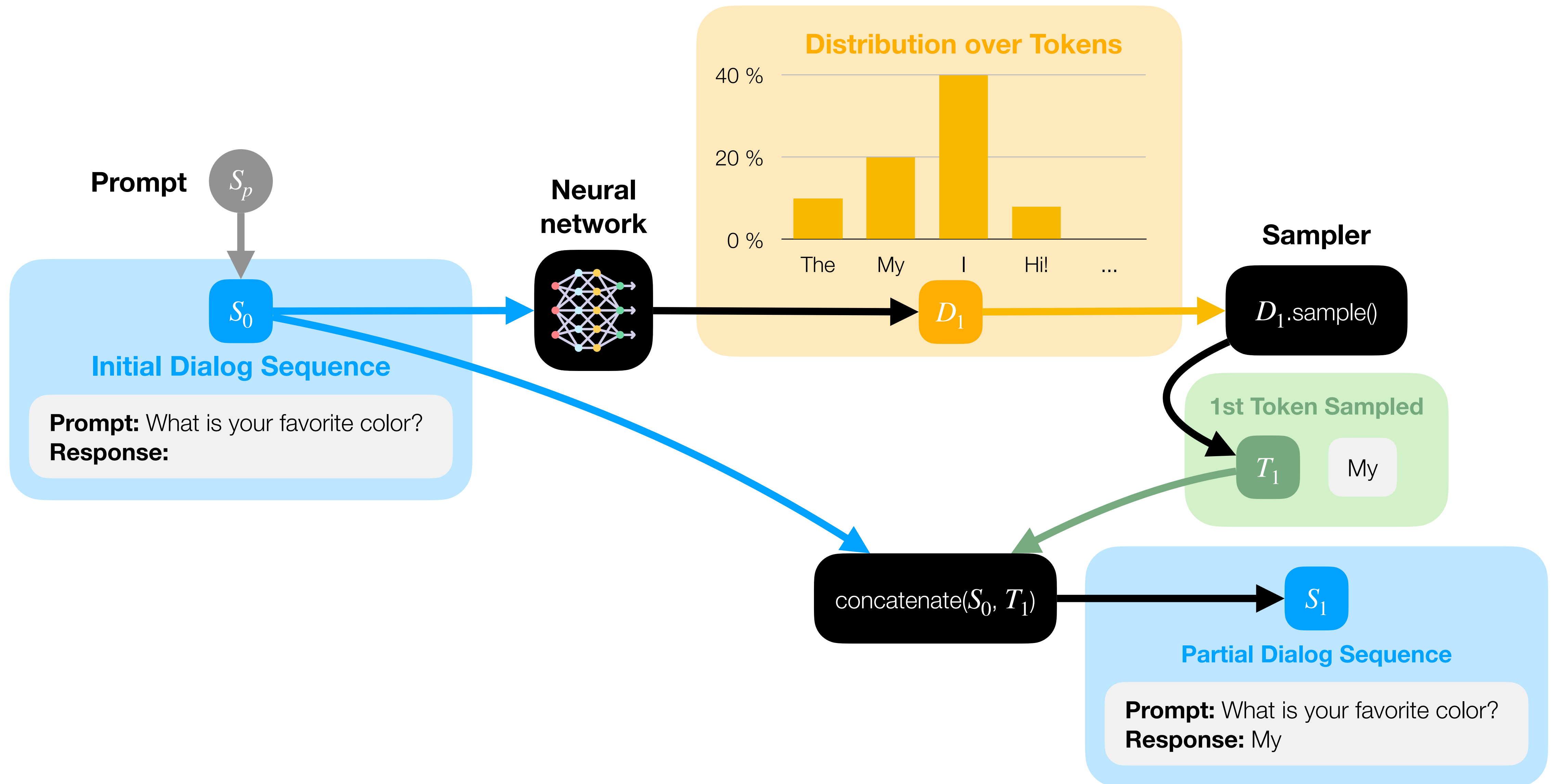
LLMs through a causal lens



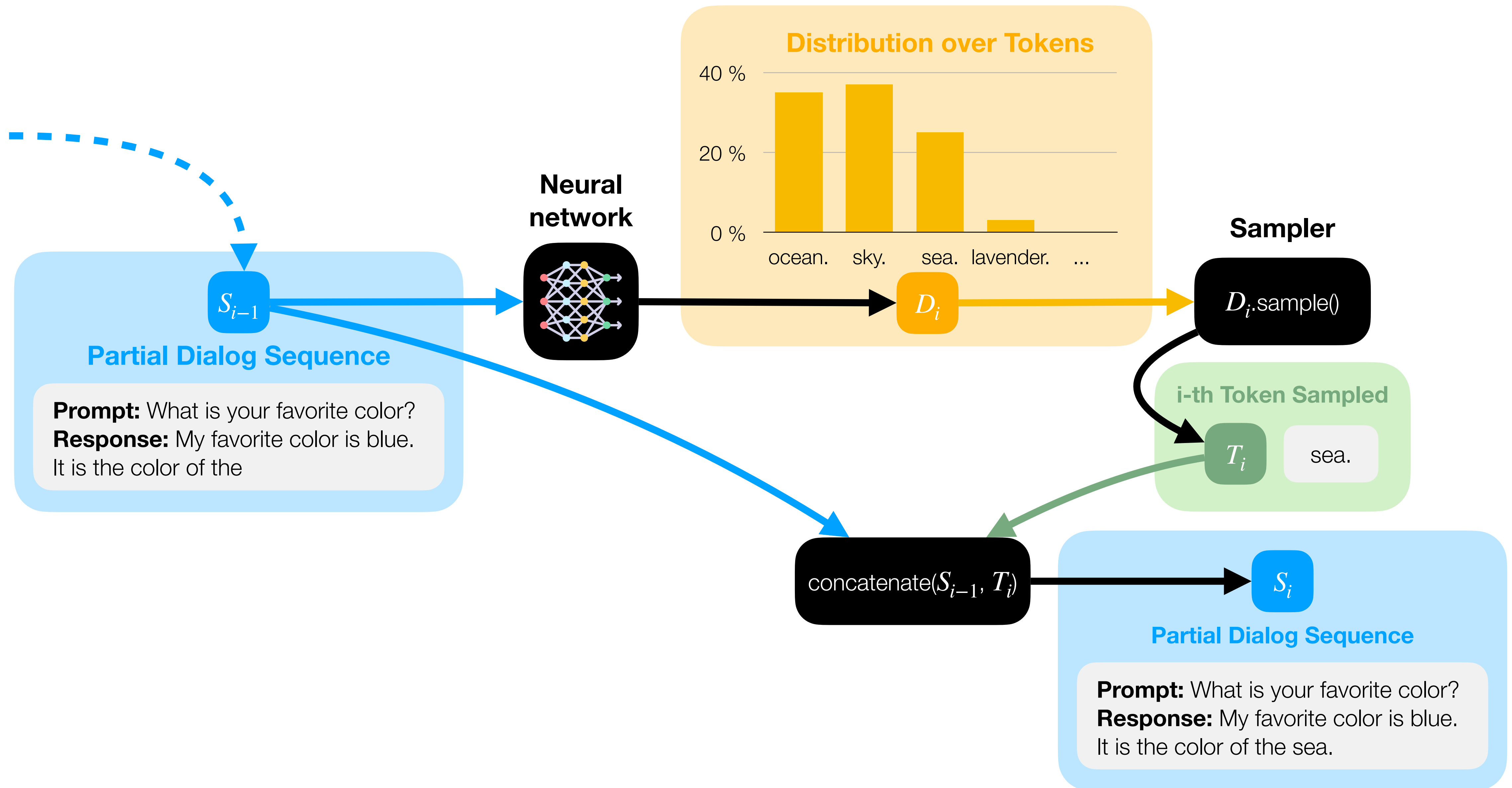
LLMs through a causal lens



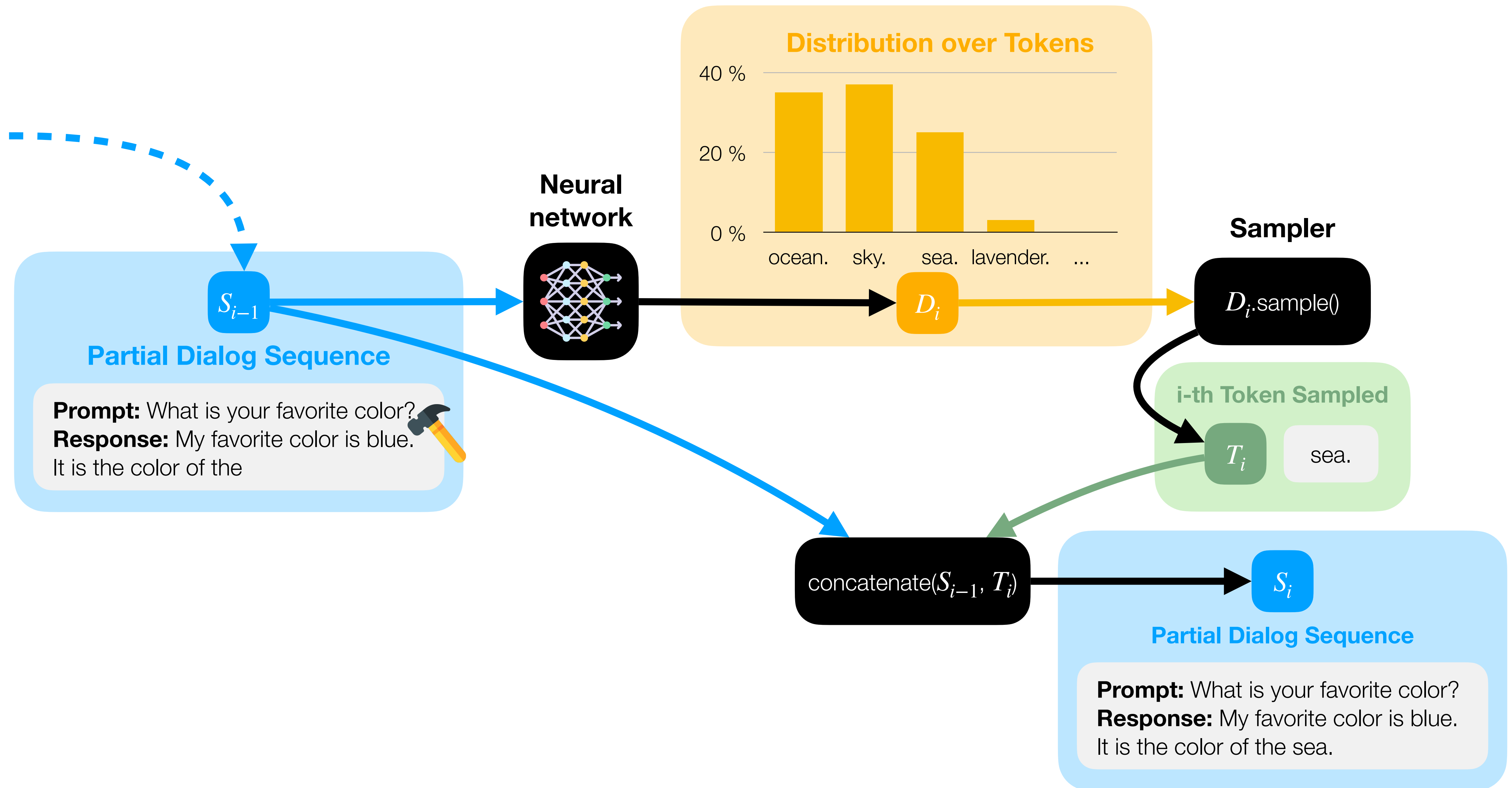
LLMs through a causal lens



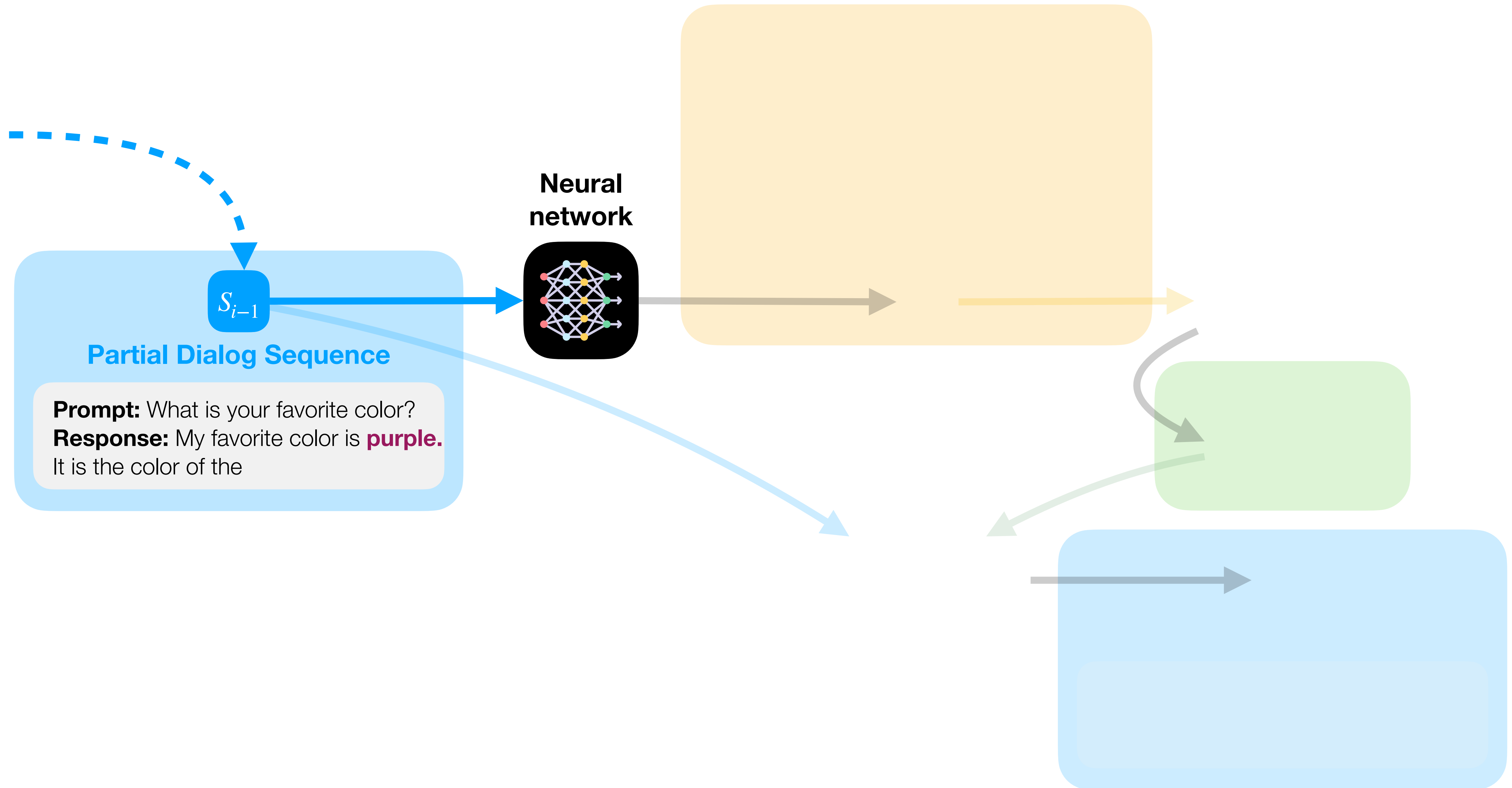
LLMs through a causal lens



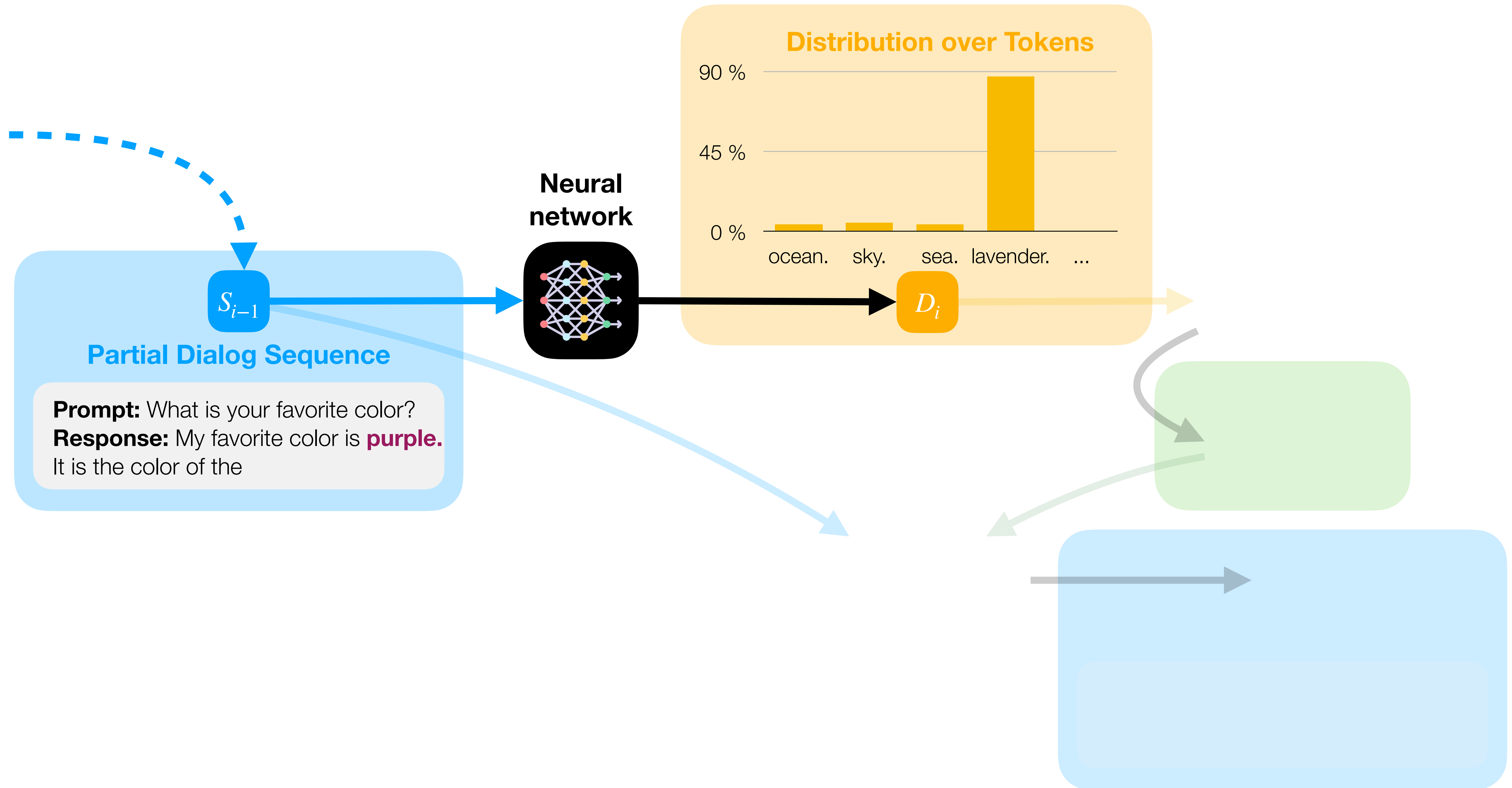
LLMs through a causal lens



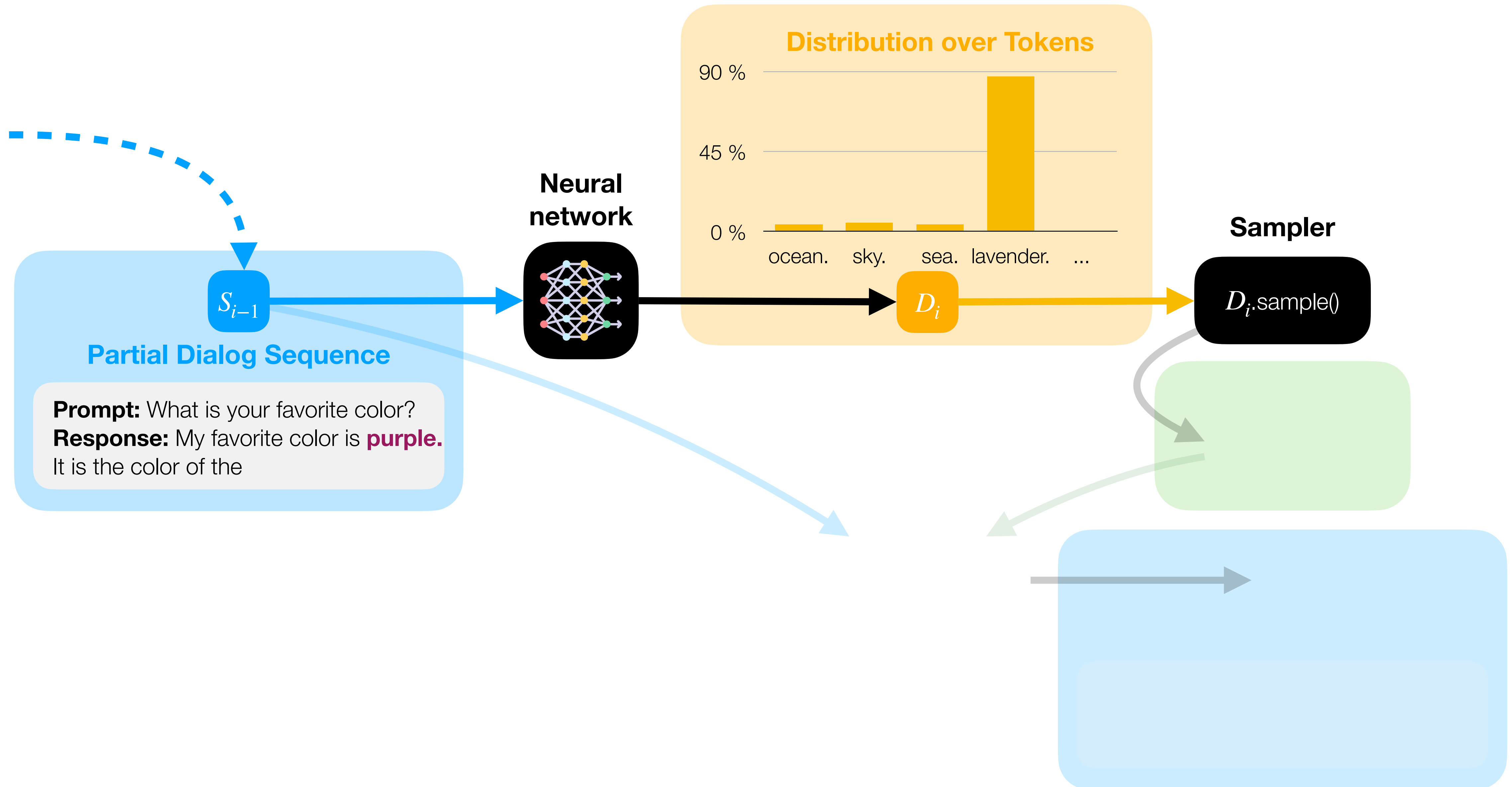
LLMs through a causal lens



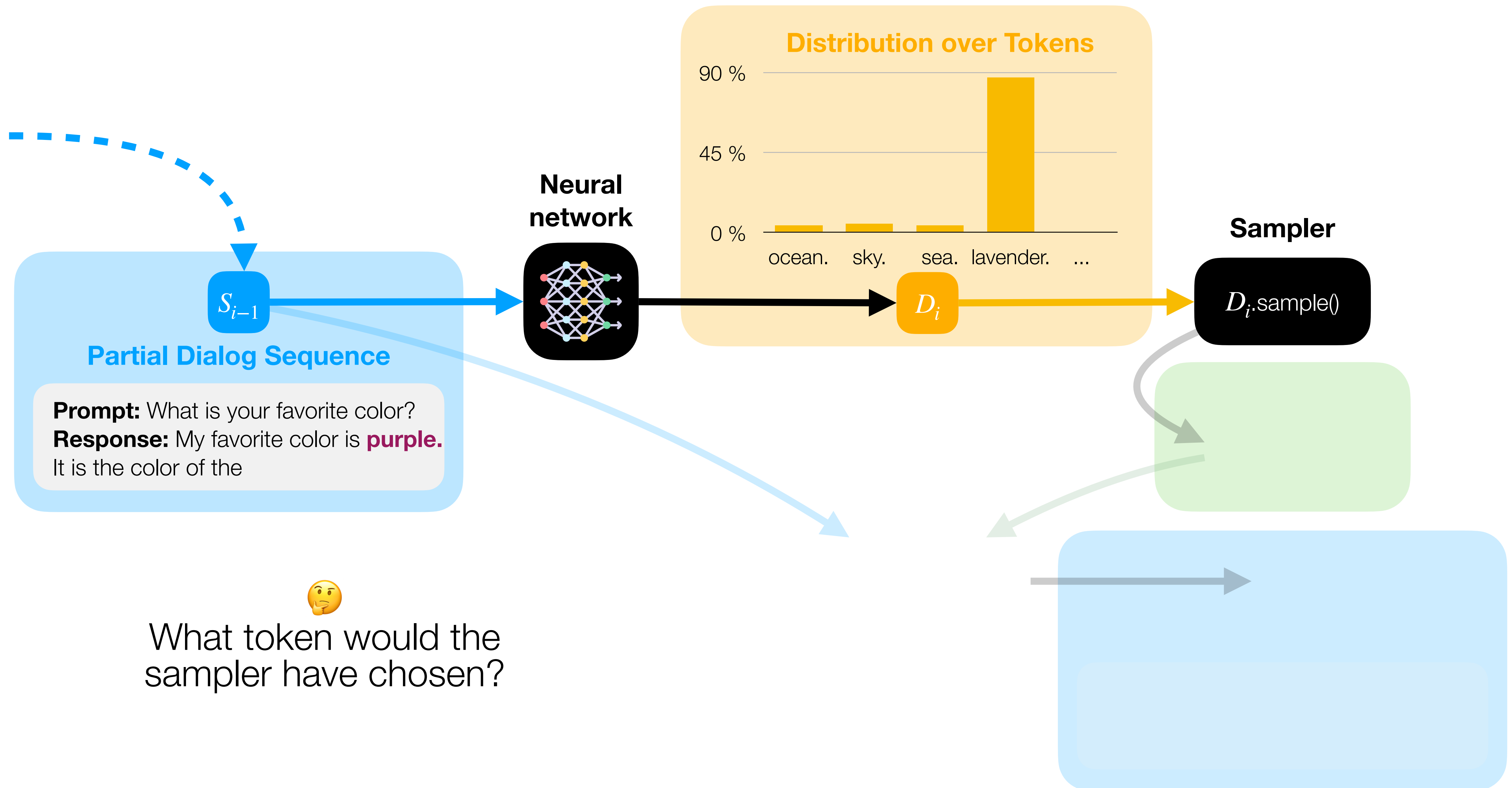
LLMs through a causal lens



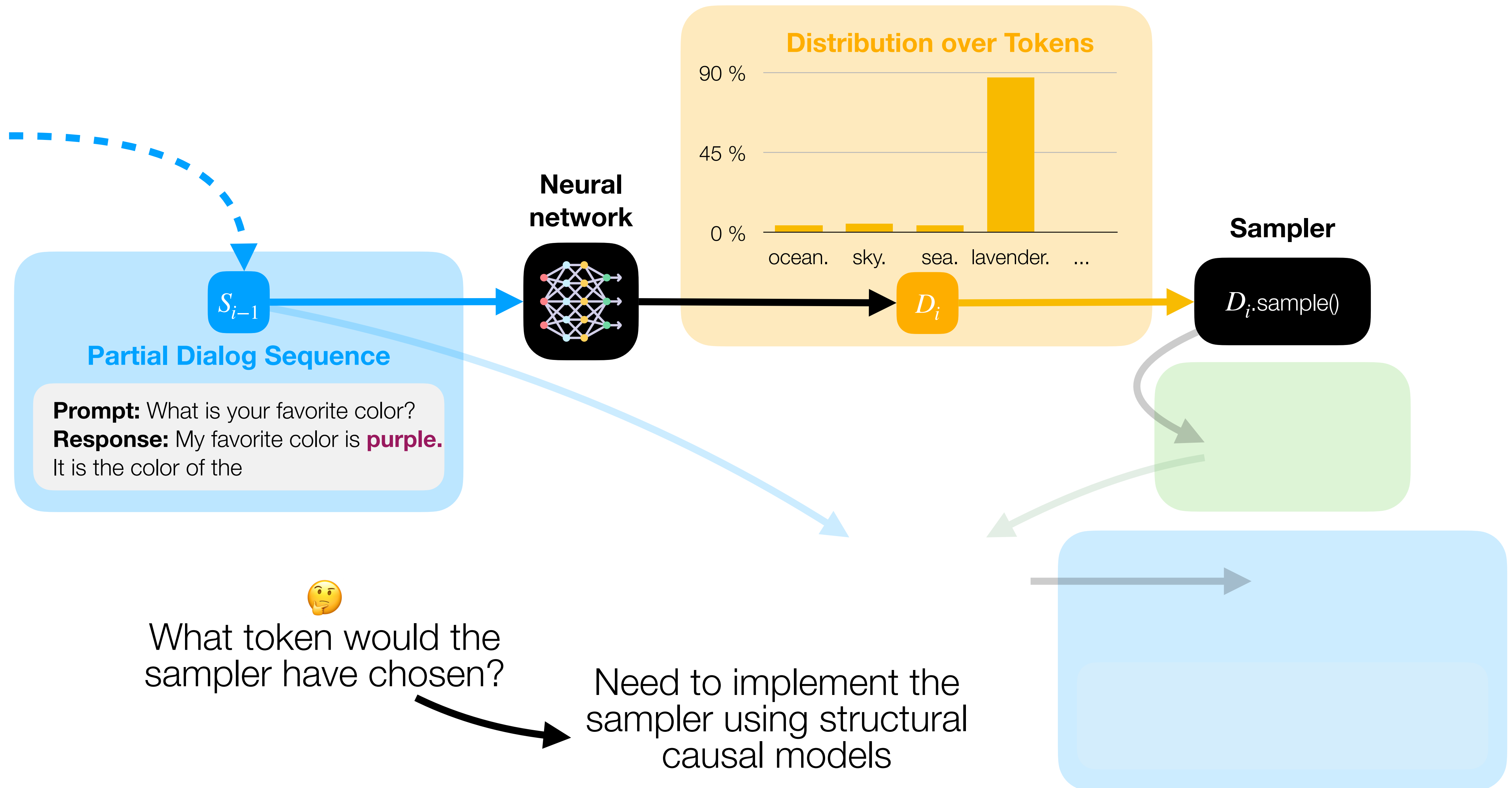
LLMs through a causal lens



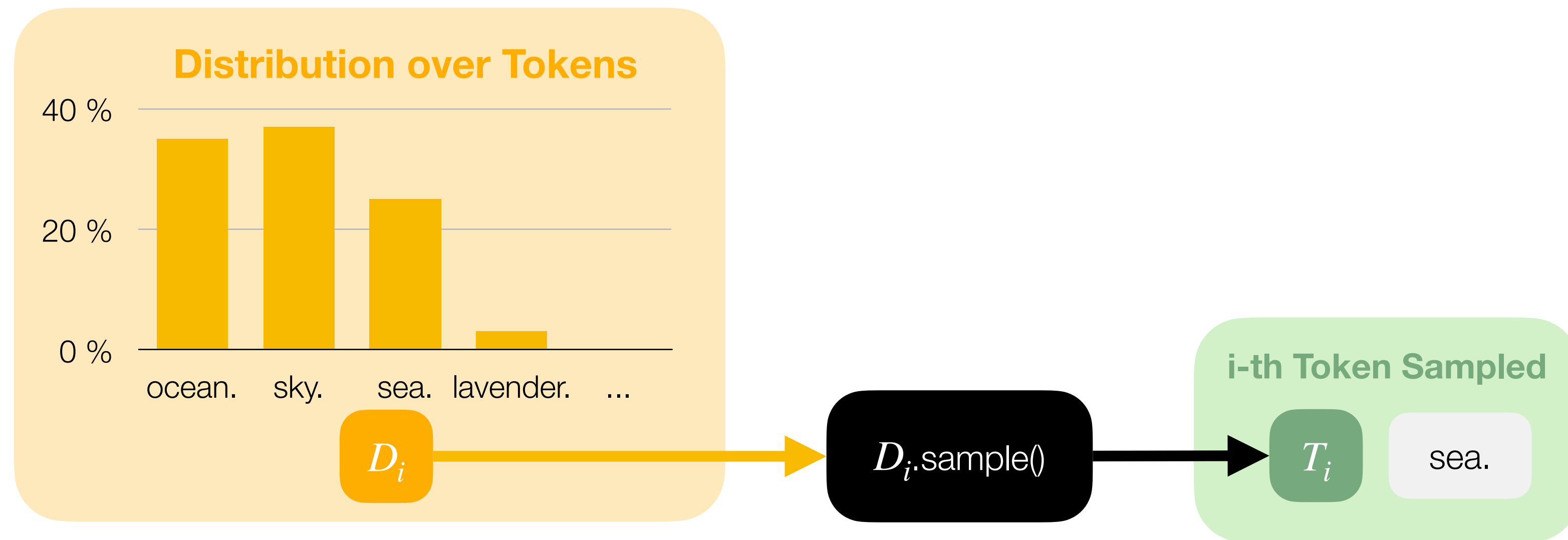
LLMs through a causal lens



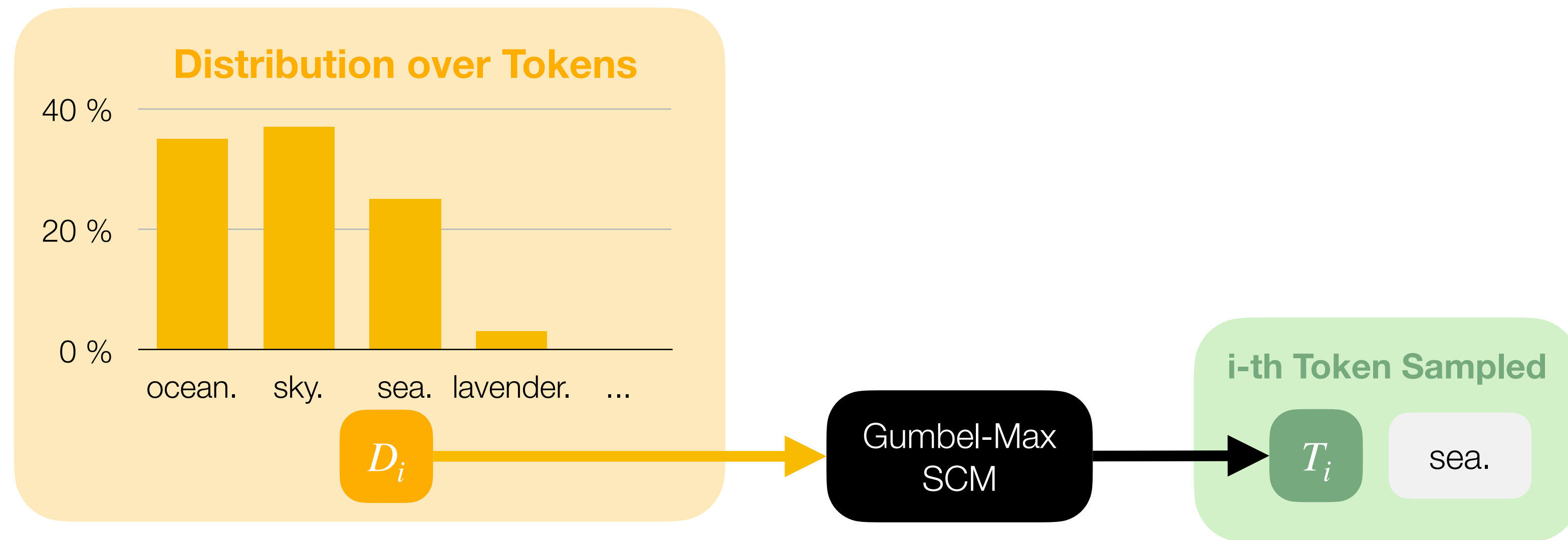
LLMs through a causal lens



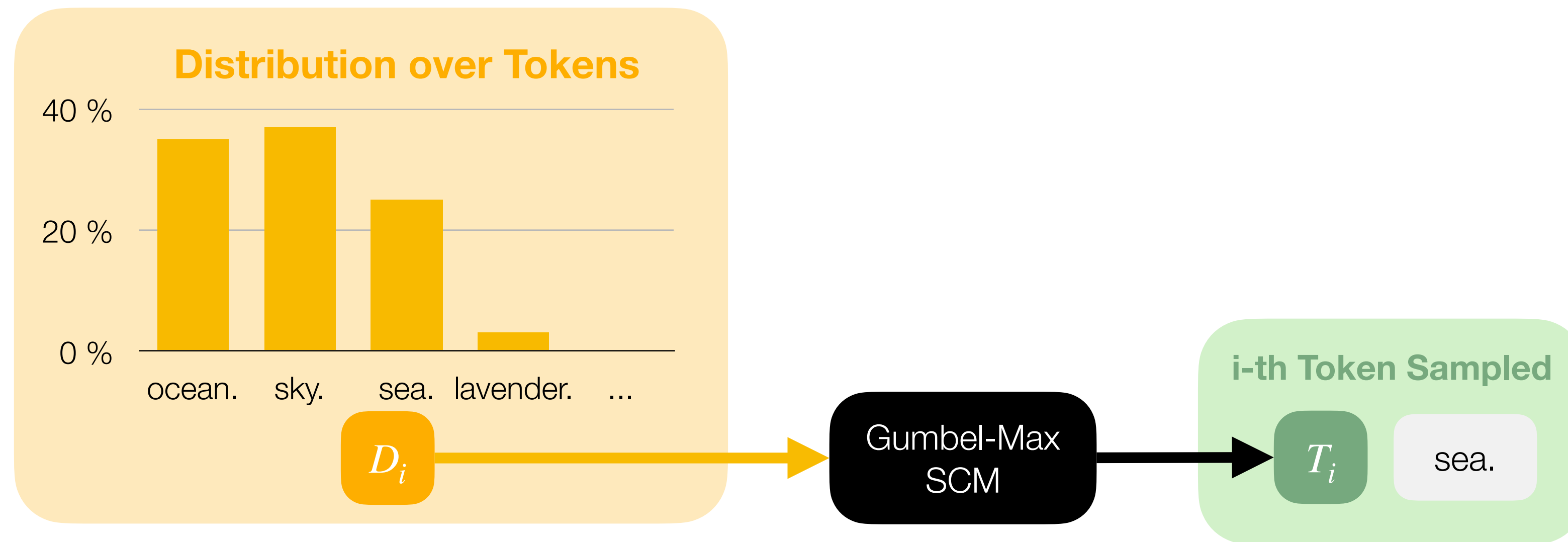
Implementing the sampler using Gumbel-max SCMs



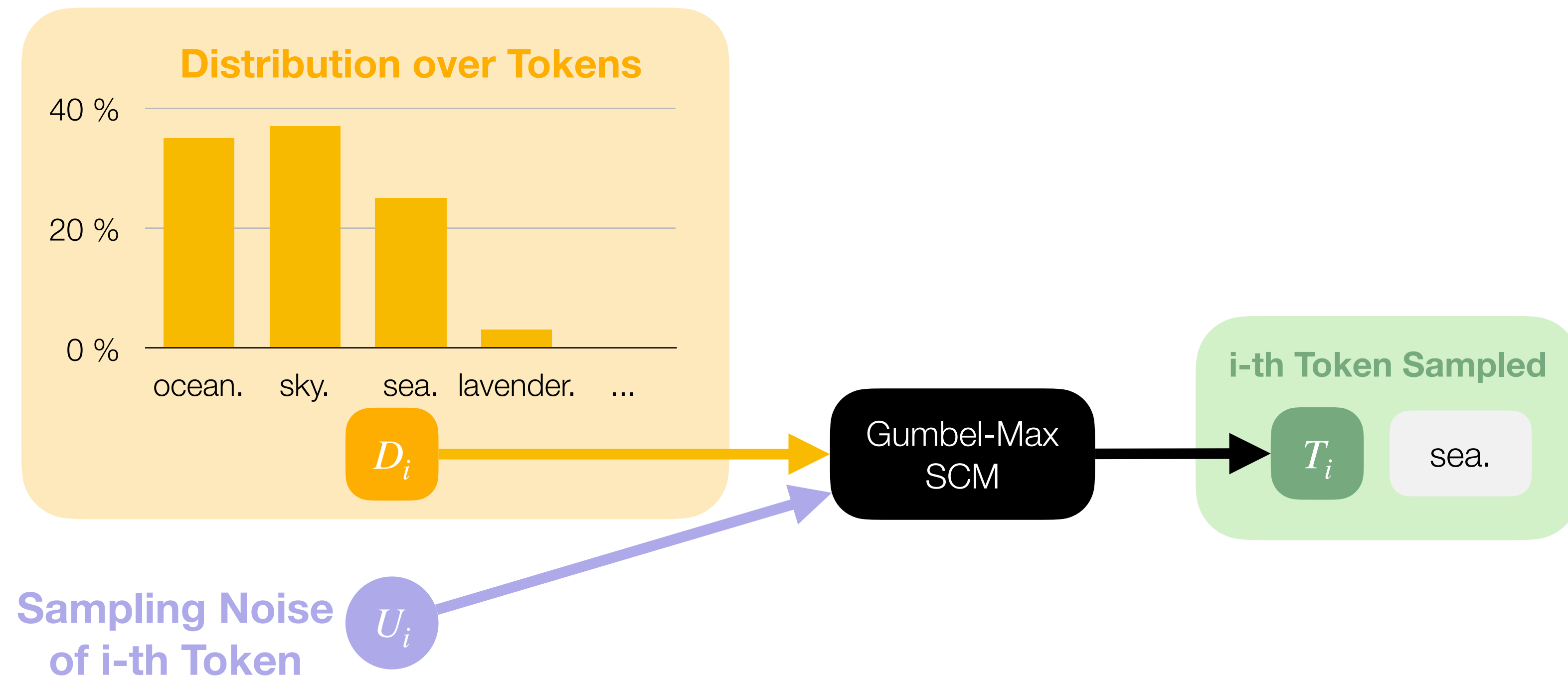
Implementing the sampler using Gumbel-max SCMs



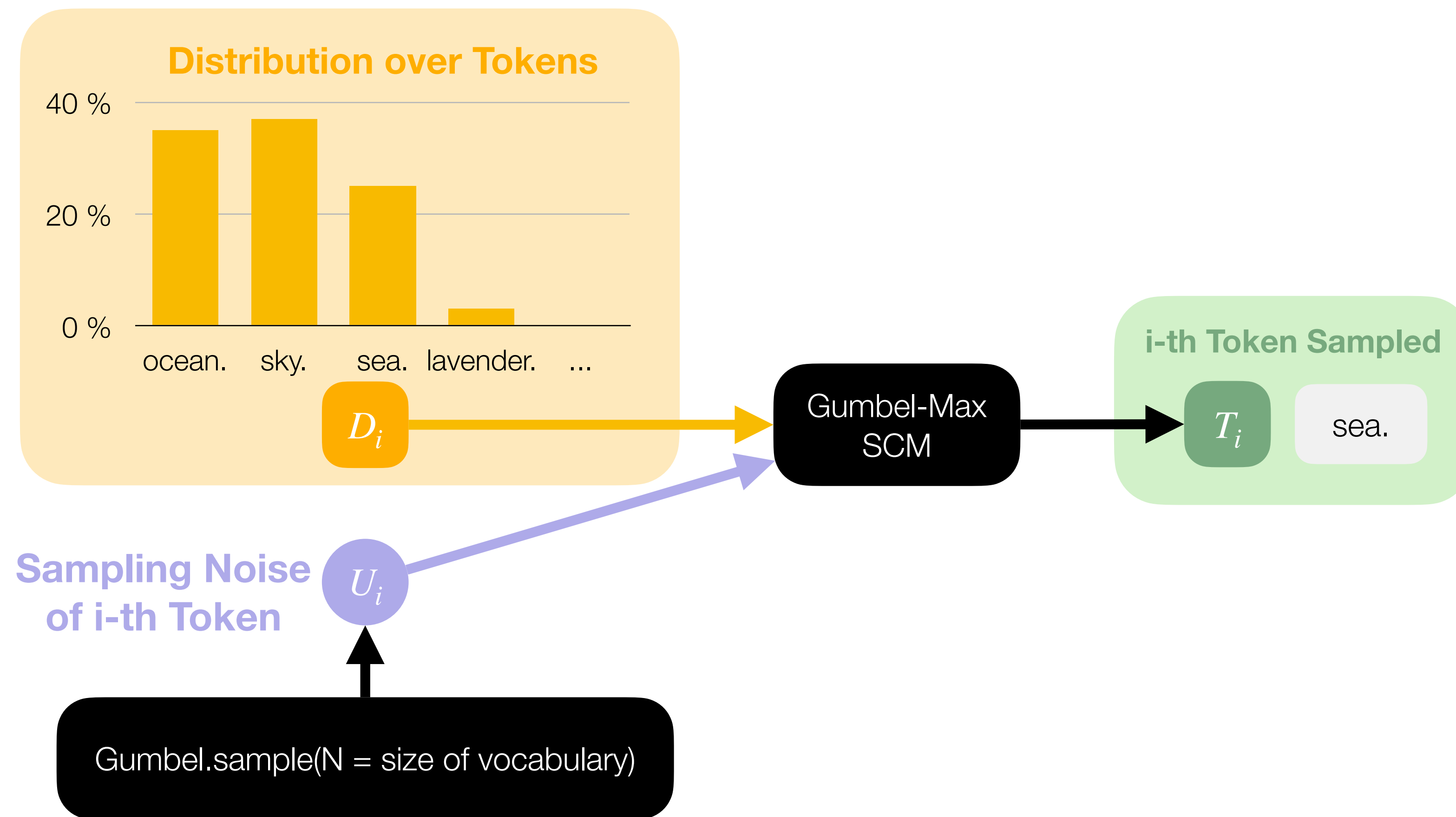
Implementing the sampler using Gumbel-max SCMs



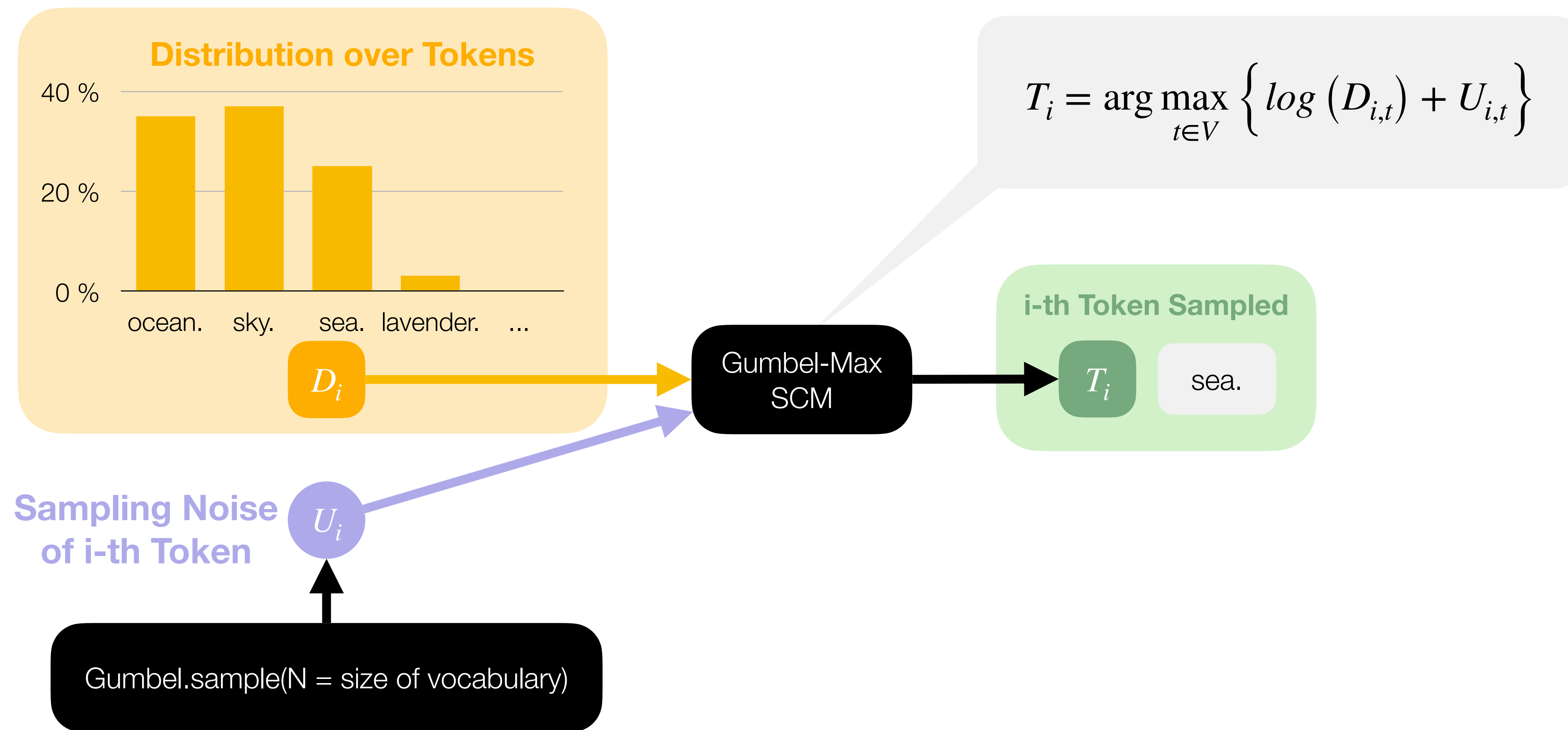
Implementing the sampler using Gumbel-max SCMs



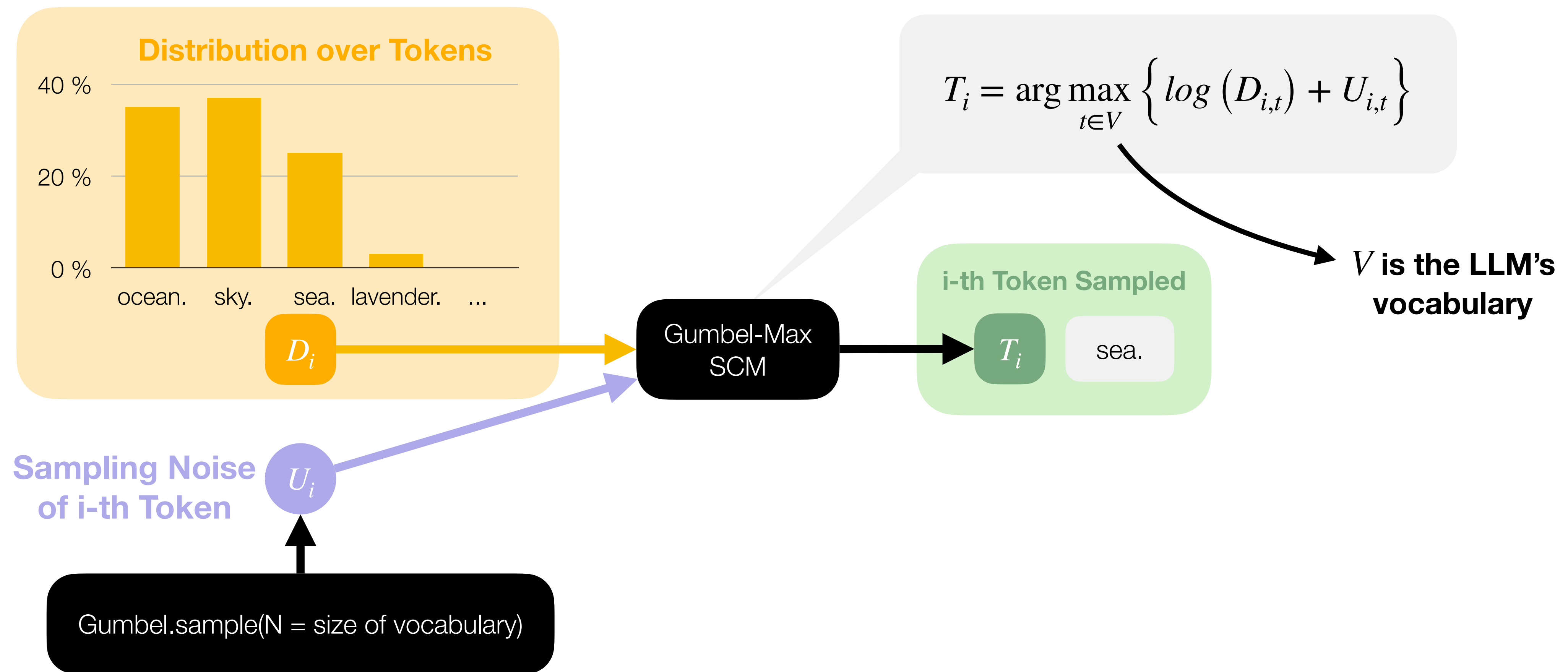
Implementing the sampler using Gumbel-max SCMs



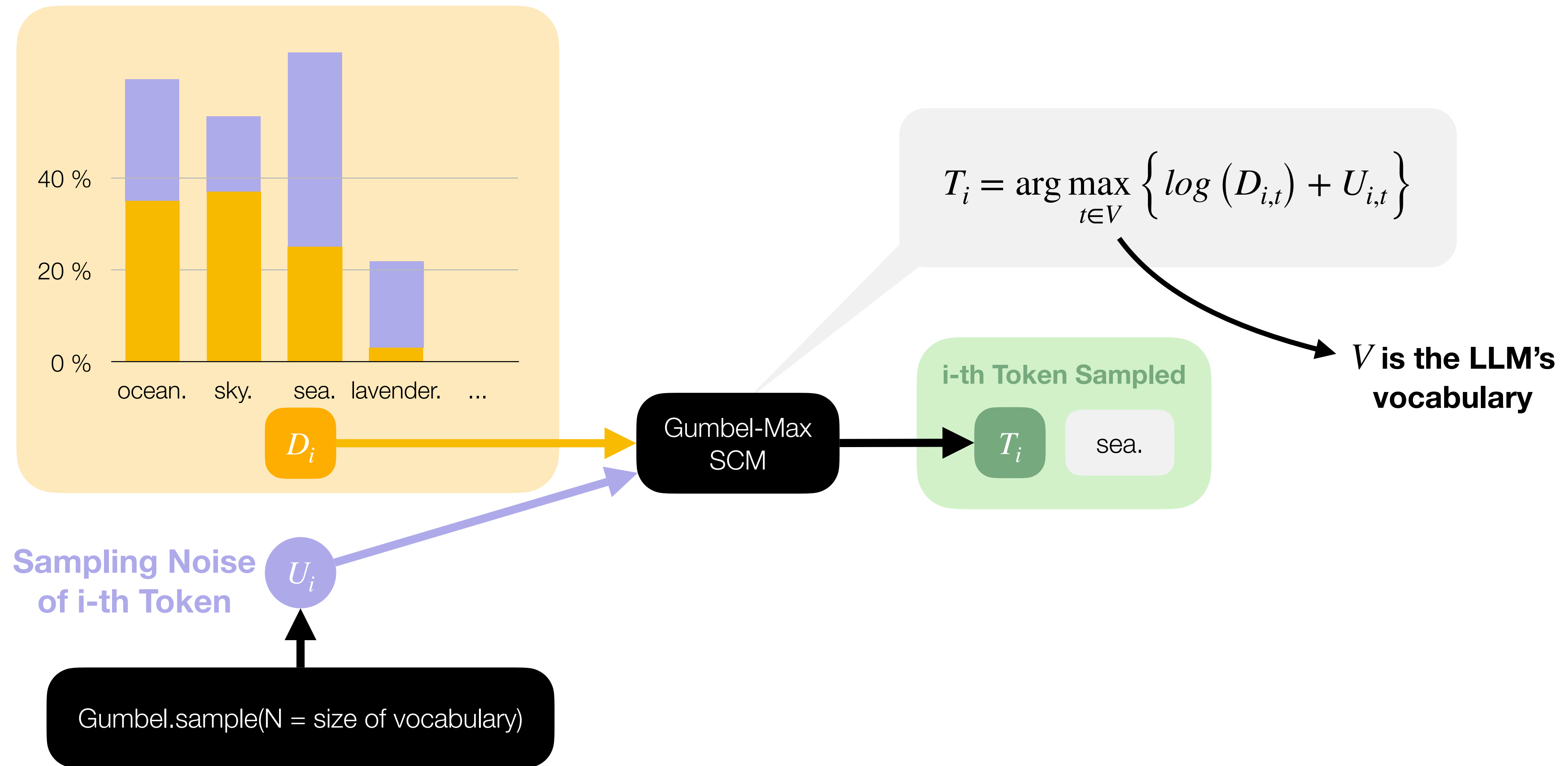
Implementing the sampler using Gumbel-max SCMs



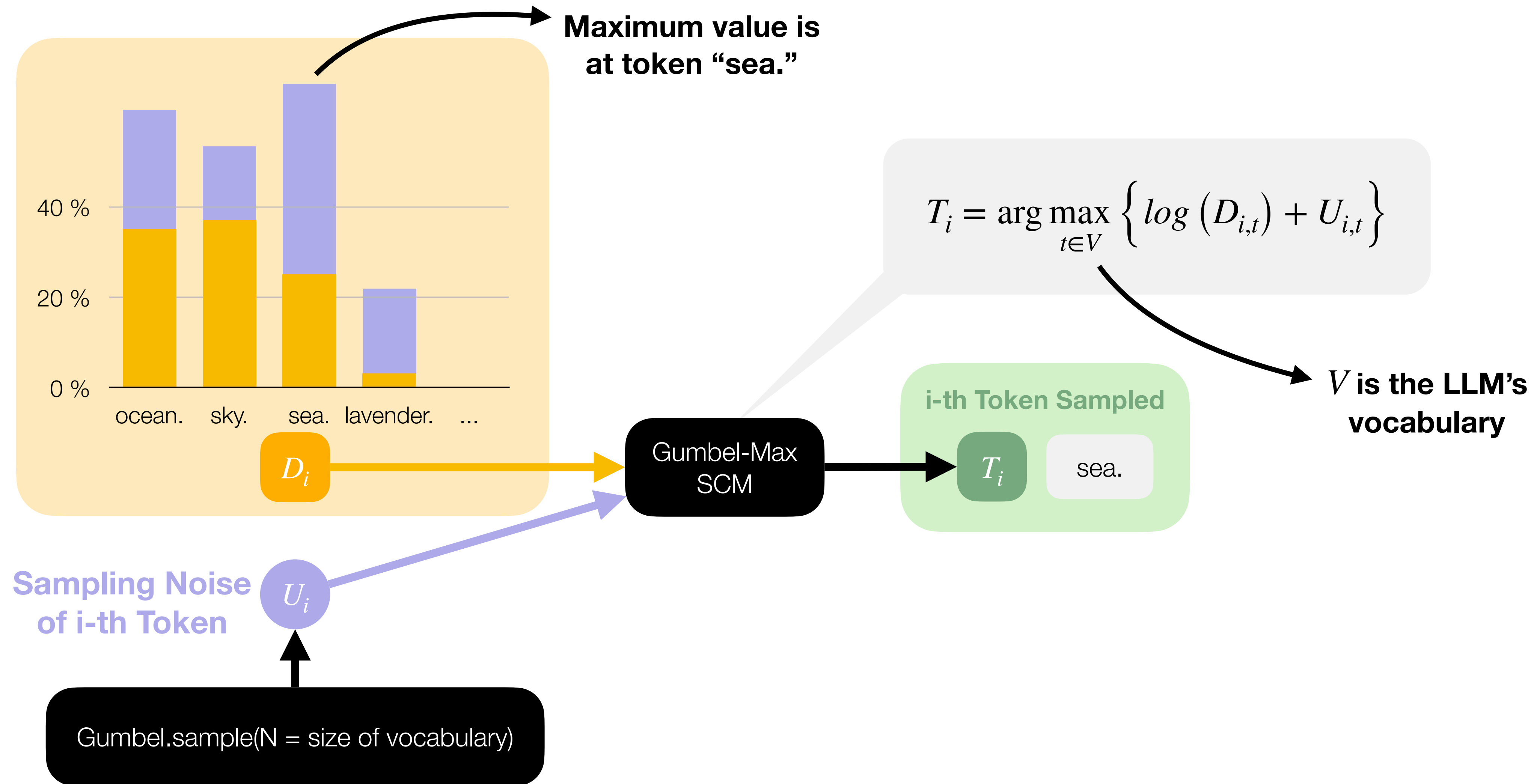
Implementing the sampler using Gumbel-max SCM



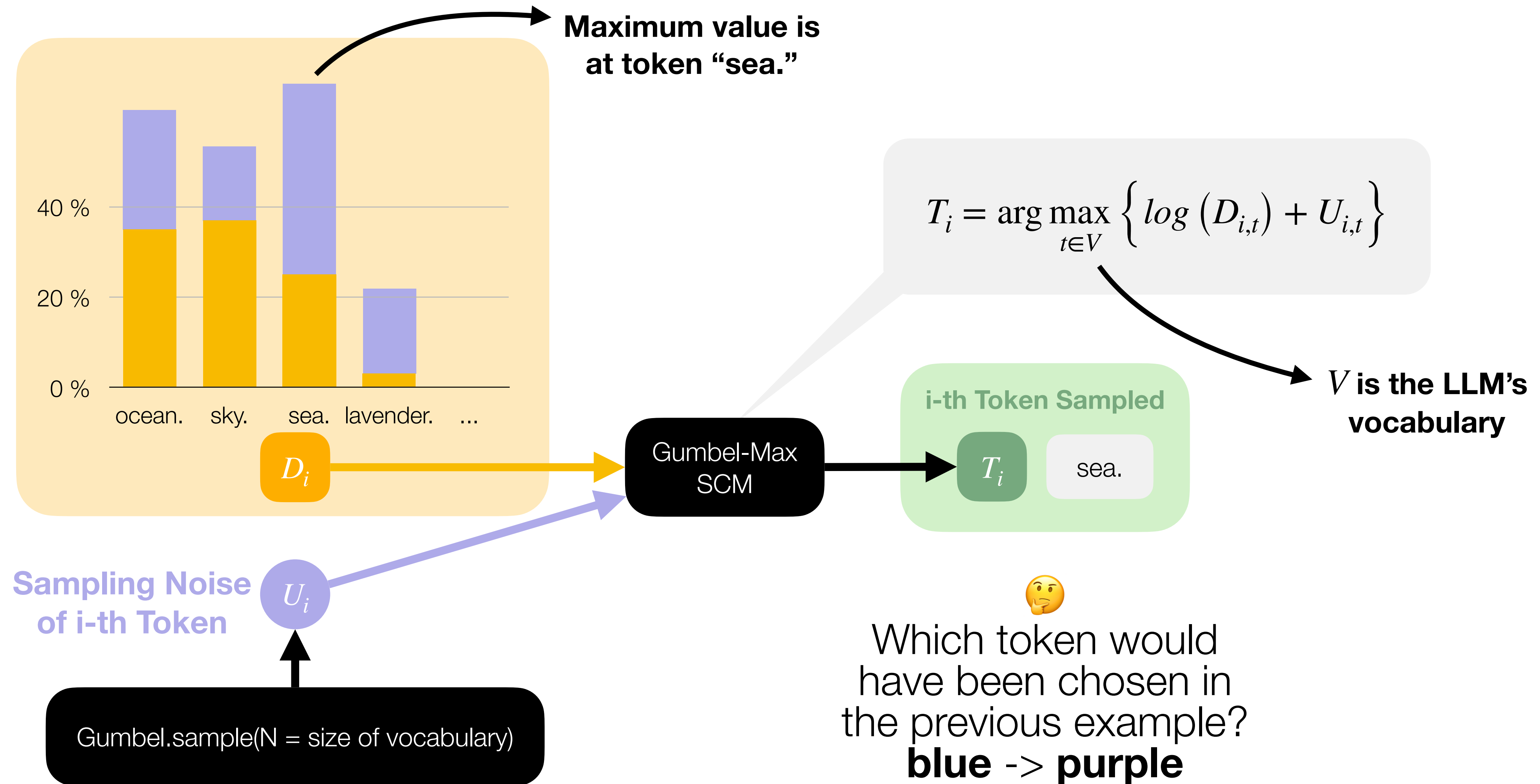
Implementing the sampler using Gumbel-max SCM



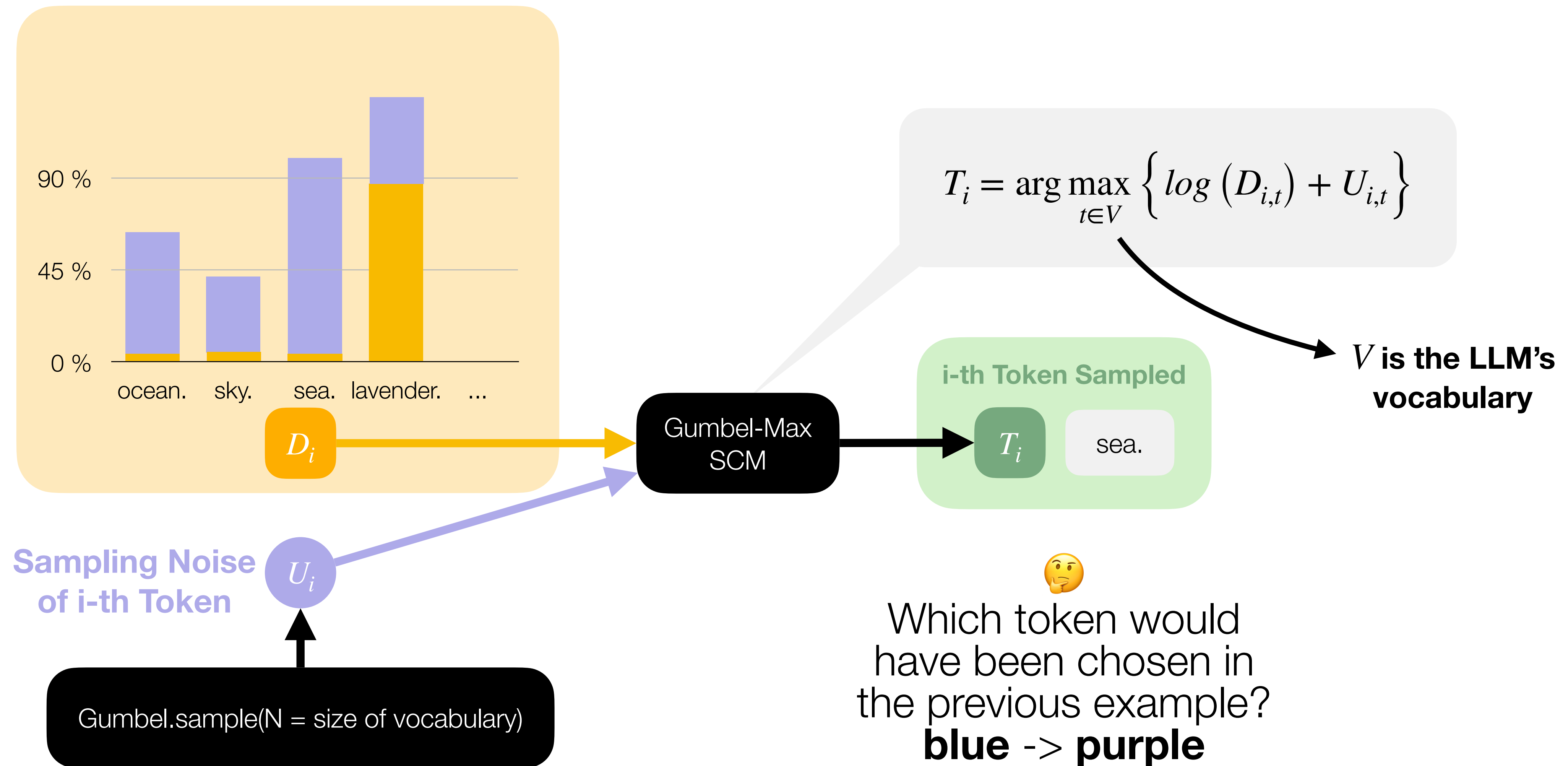
Implementing the sampler using Gumbel-max SCM



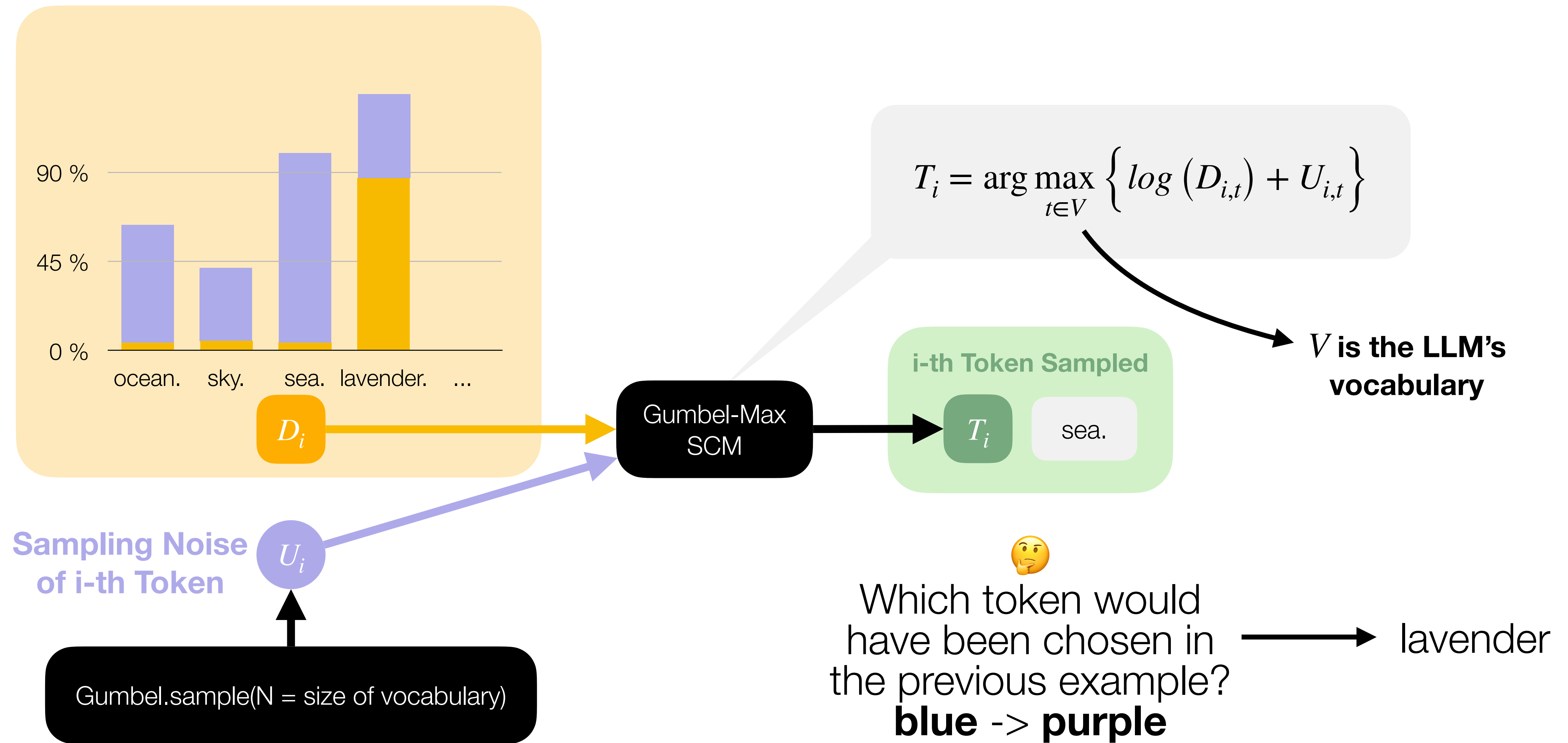
Implementing the sampler using Gumbel-max SCMs



Implementing the sampler using Gumbel-max SCMs



Implementing the sampler using Gumbel-max SCM



Finding out what the LLM “would have said”

Prompt S_p



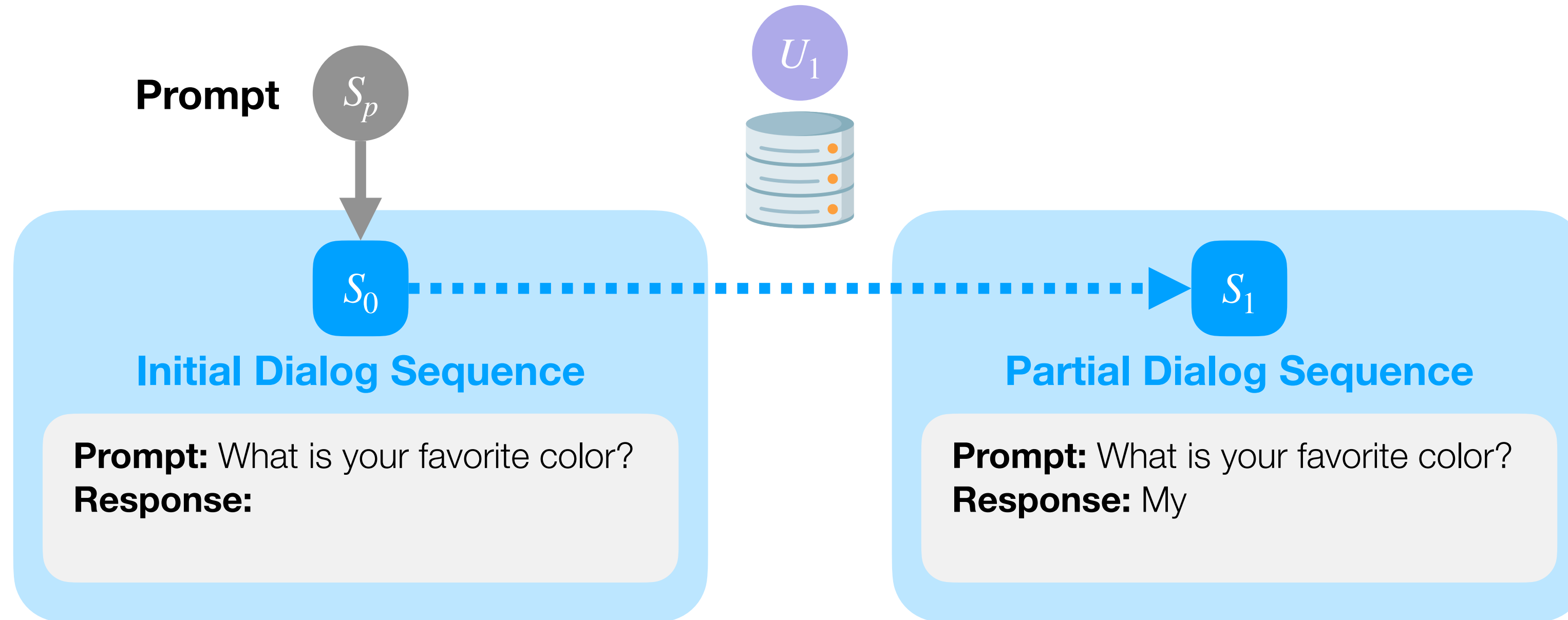
S_0

Initial Dialog Sequence

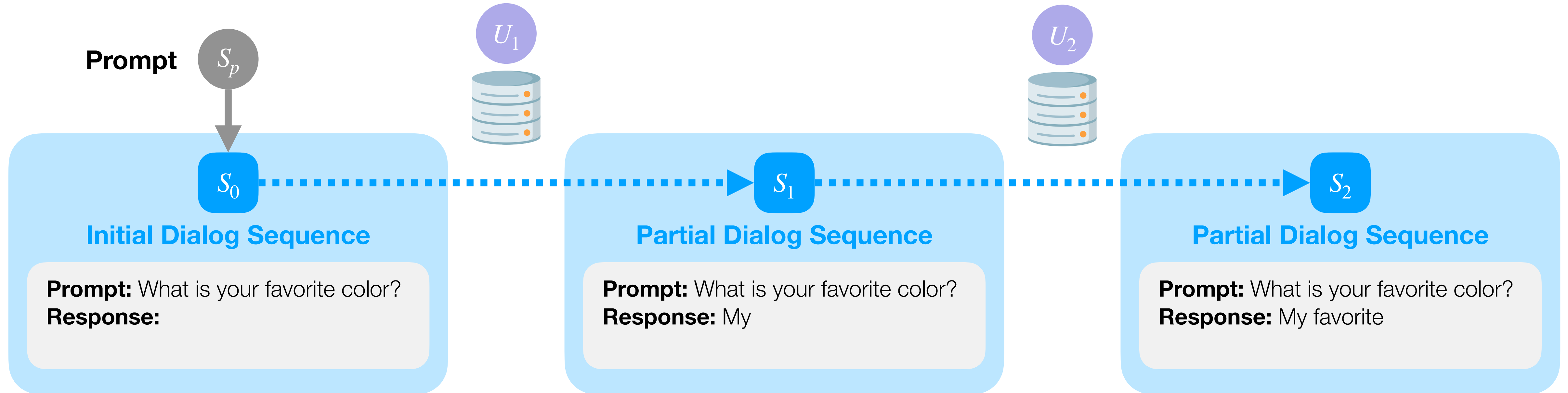
Prompt: What is your favorite color?

Response:

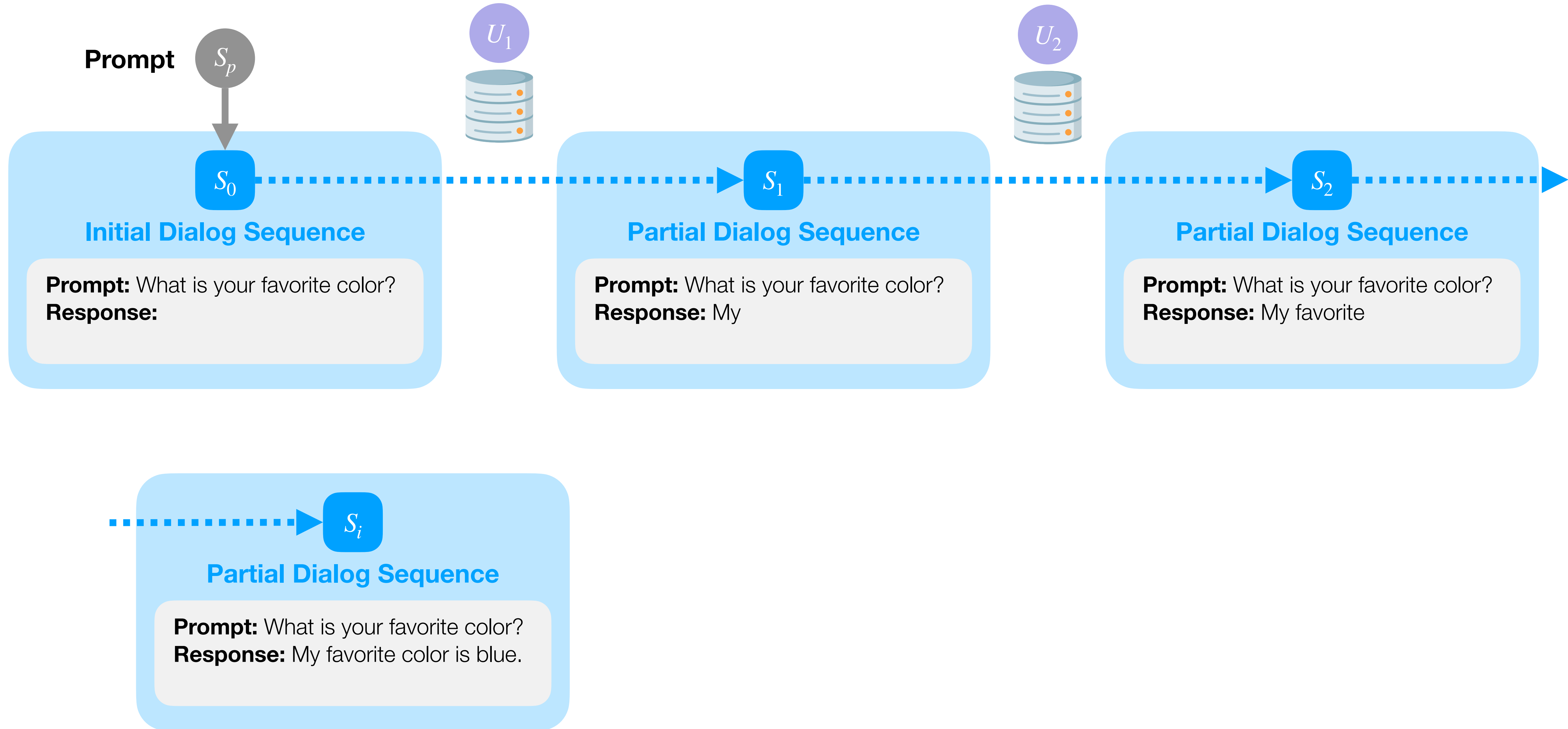
Finding out what the LLM “would have said”



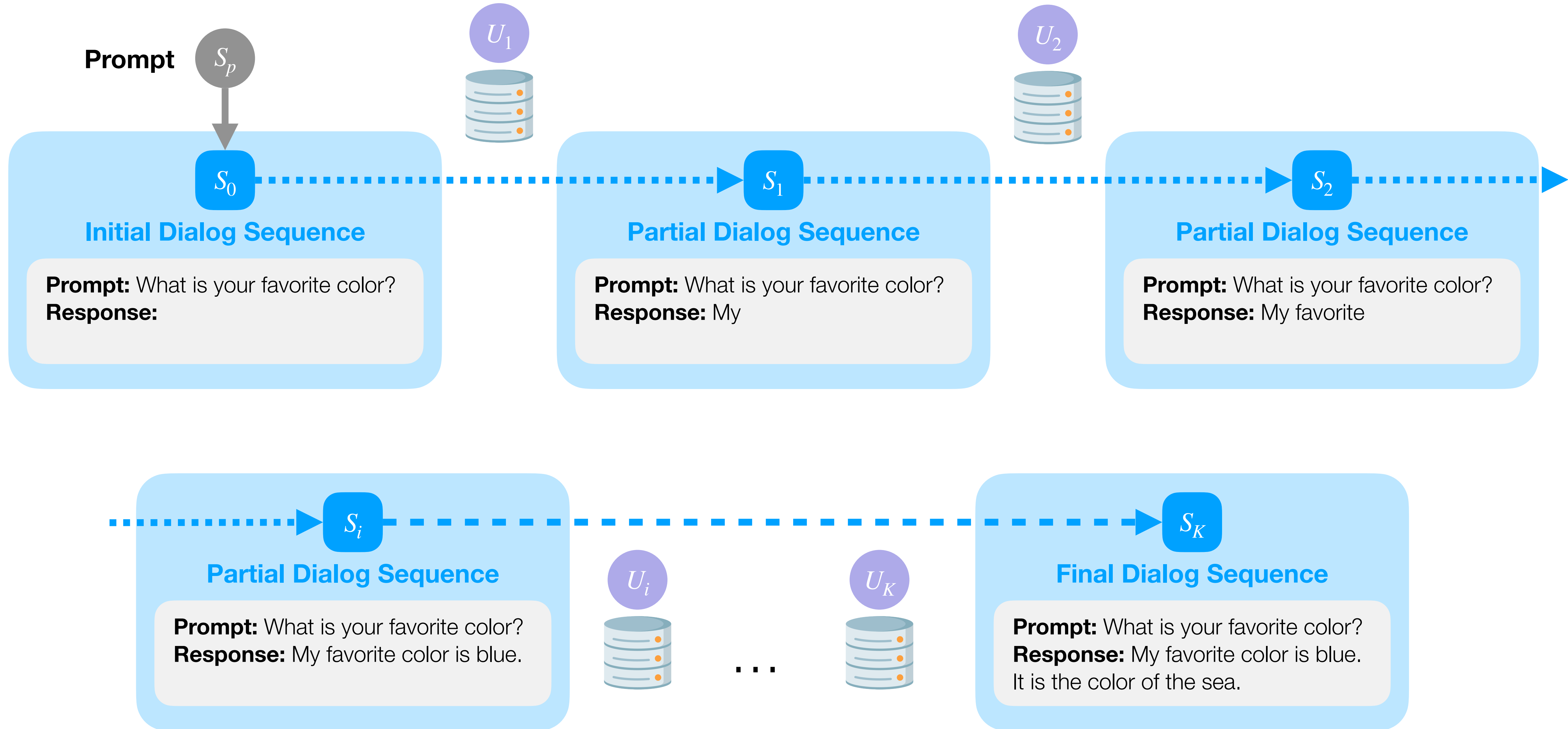
Finding out what the LLM “would have said”



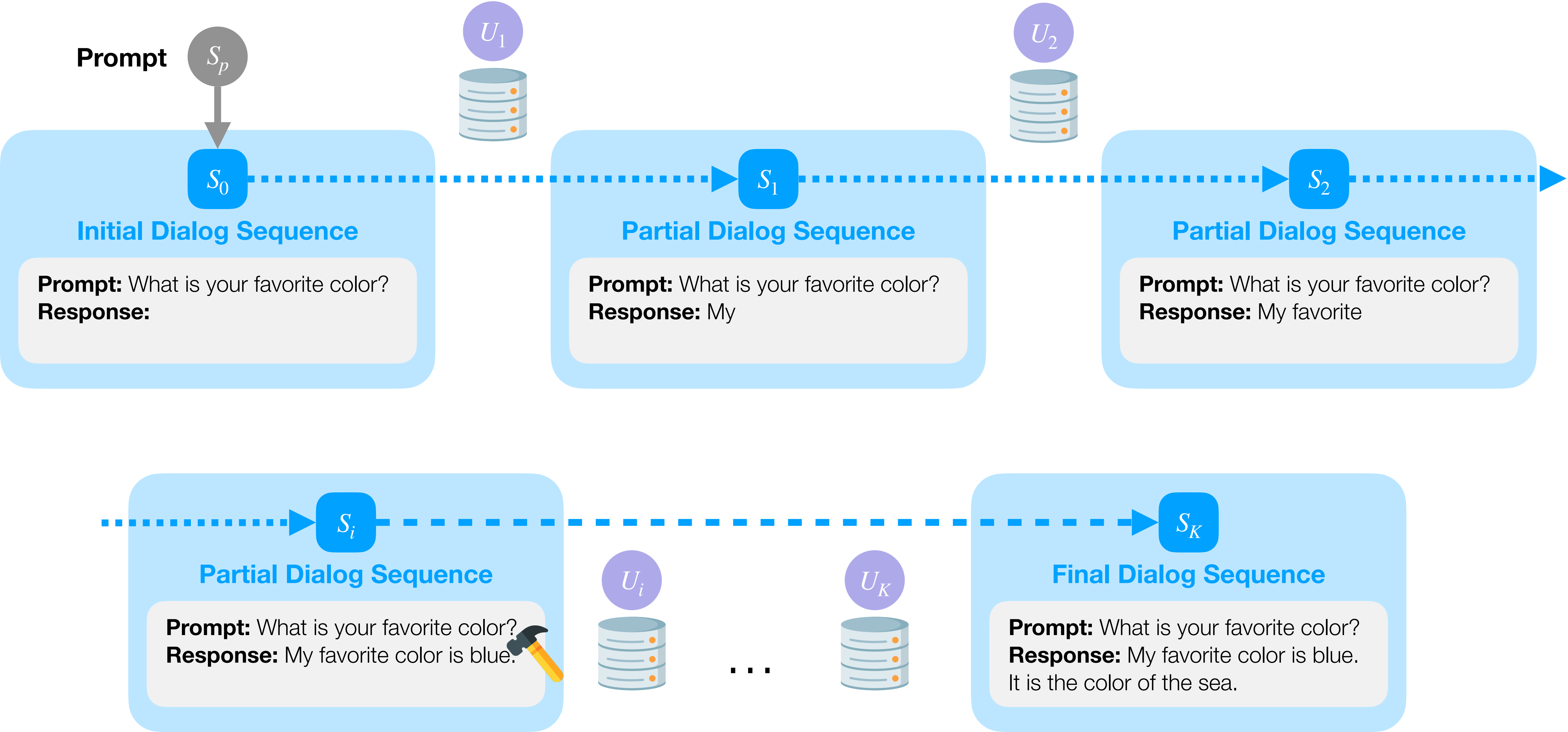
Finding out what the LLM “would have said”



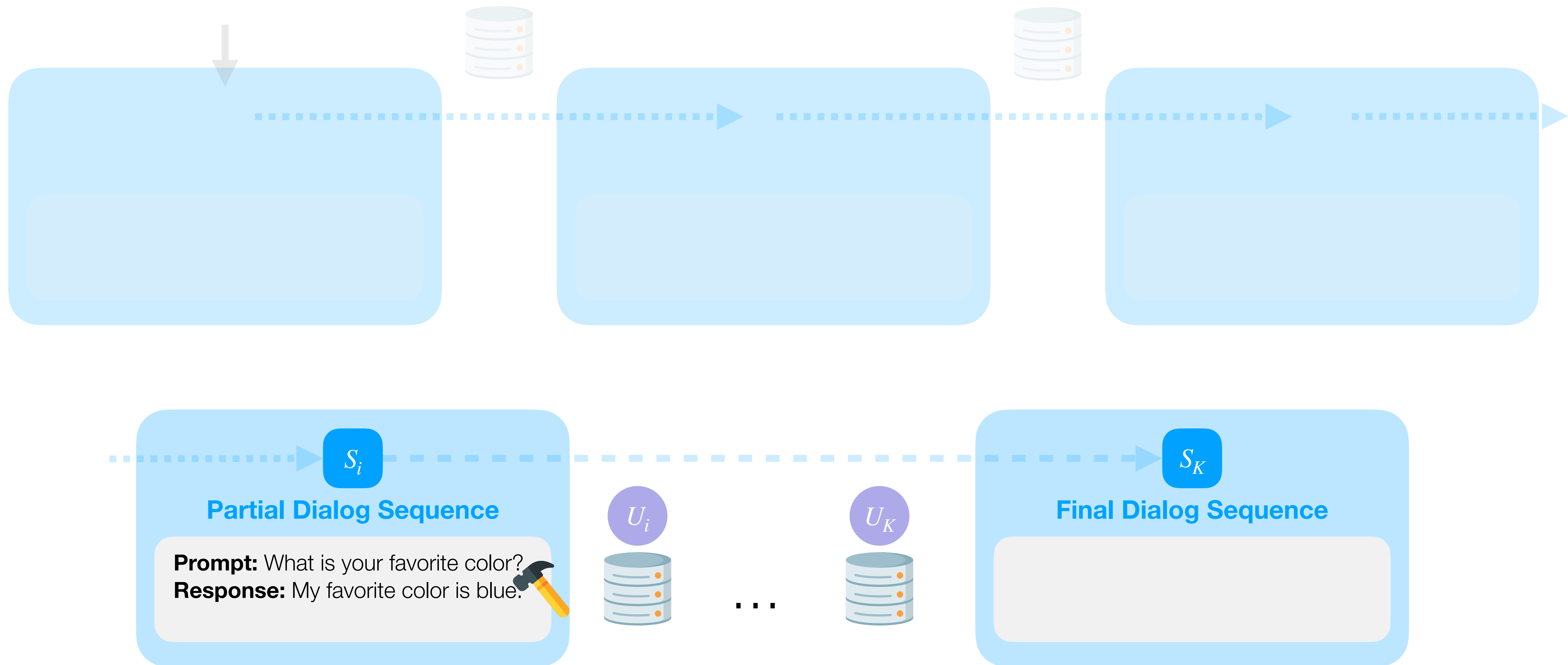
Finding out what the LLM “would have said”



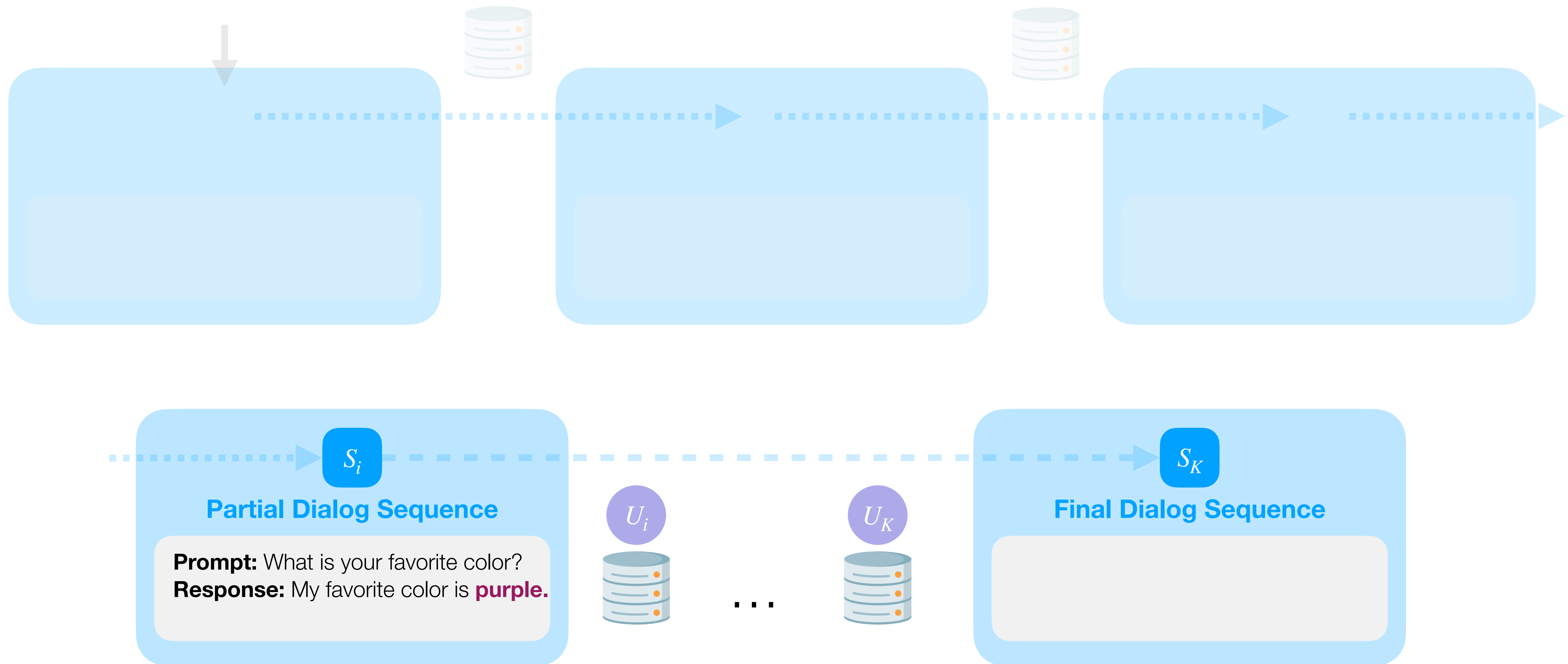
Finding out what the LLM “would have said”



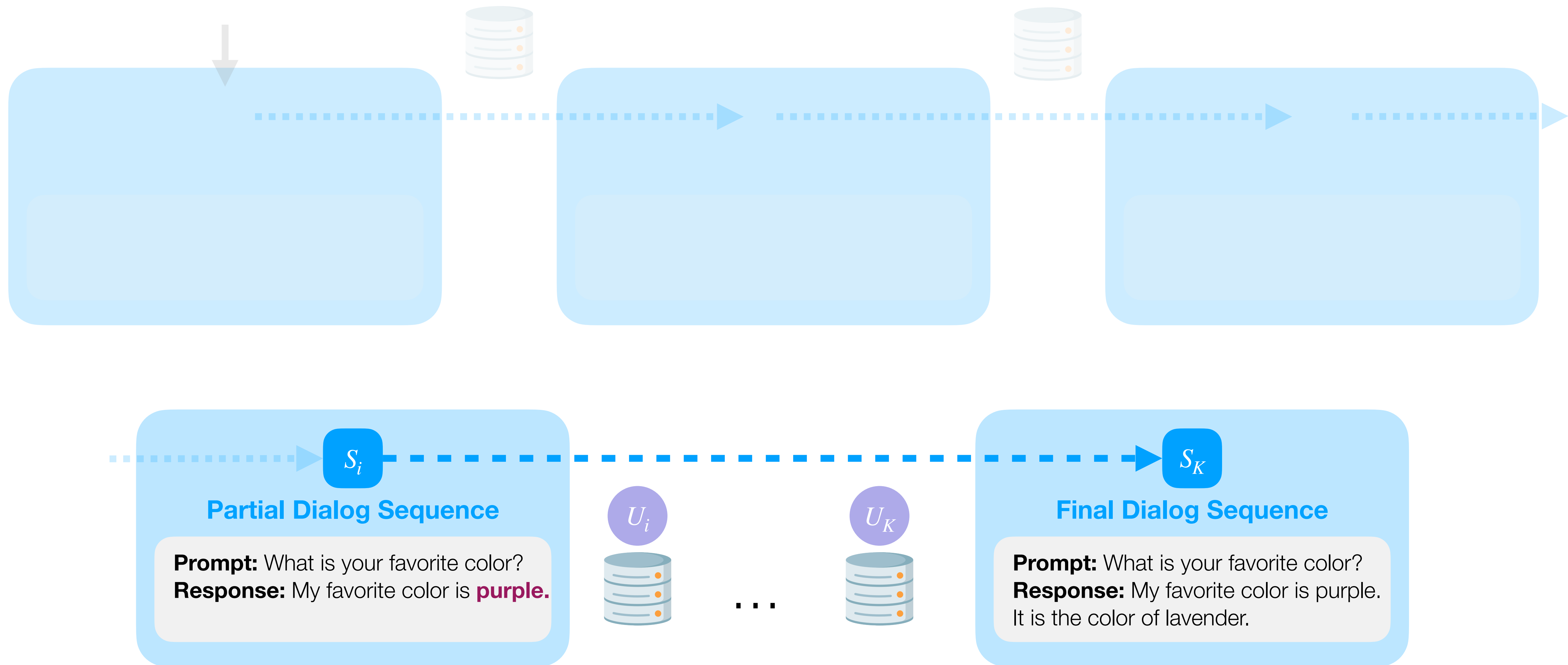
Finding out what the LLM “would have said”



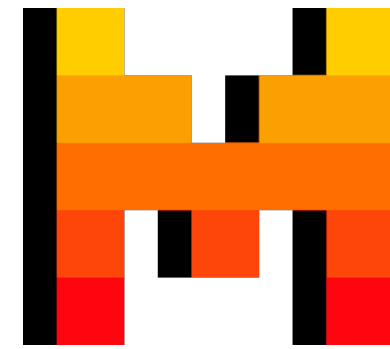
Finding out what the LLM “would have said”



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



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 Lyra stood at the helm of her _____ 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 lost at sea.

Counterfactual story



Do the LLMs show signs of bias?

Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data

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?

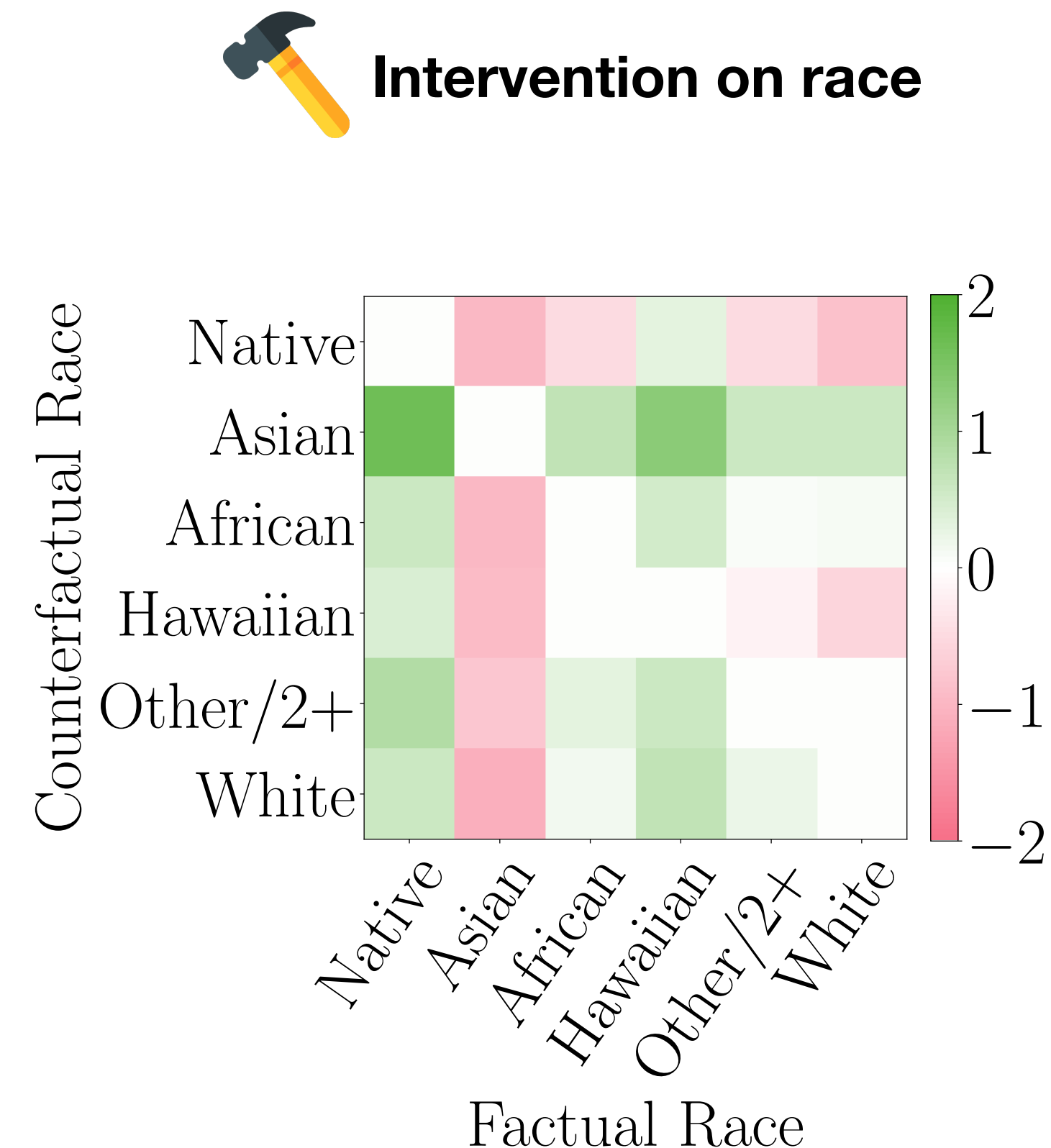
- We ask LLMs to generate fictional census data
- Each individual has a list of features:
 - Age
 - Sex
 - Citizenship
 - Race
 - ...
 - Income
 - Education



Intervention on race

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

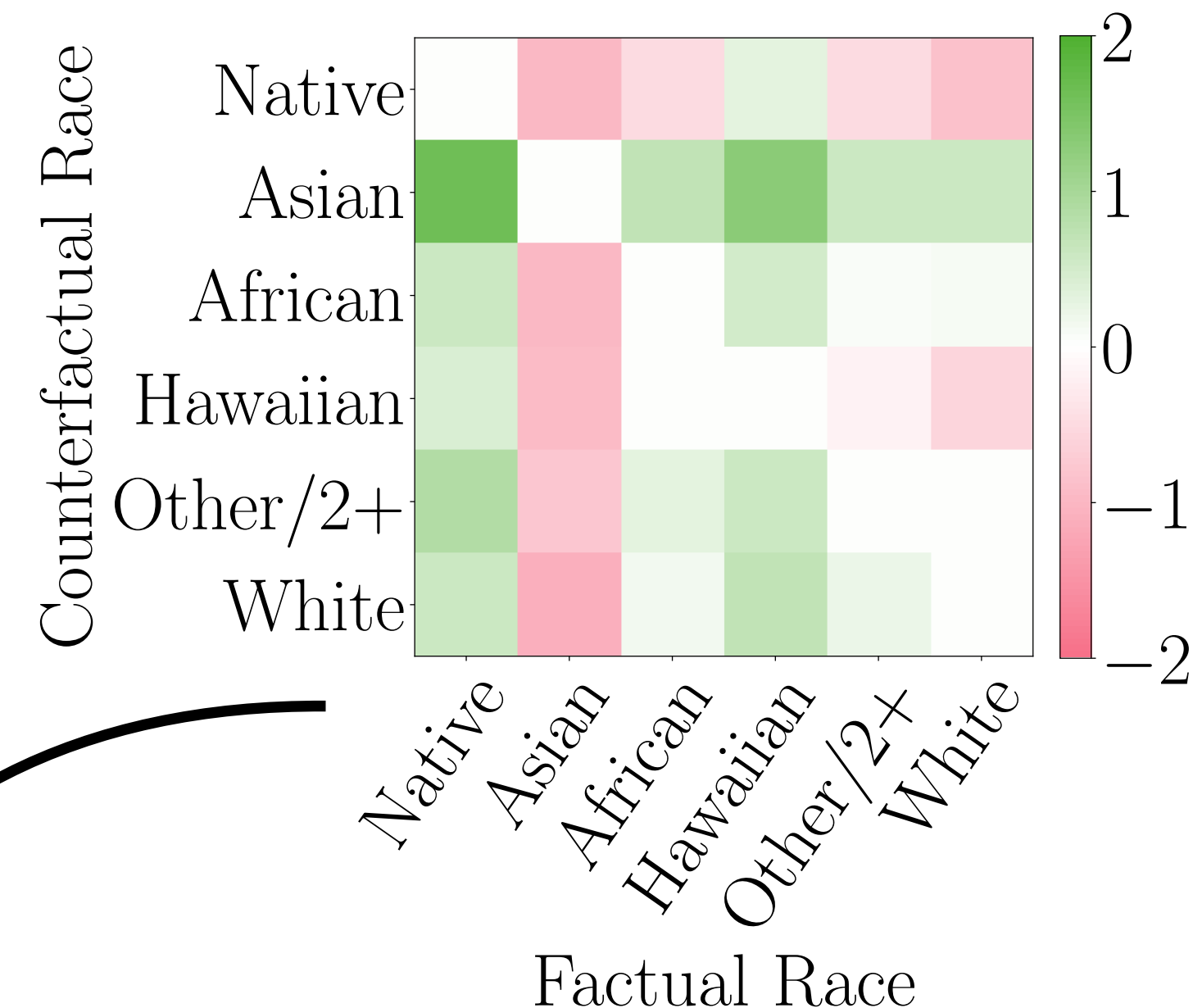


Do the LLMs show signs of bias?

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



Intervention on race



**The education level of
“Native” and “Hawaiian”
would have increased**



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

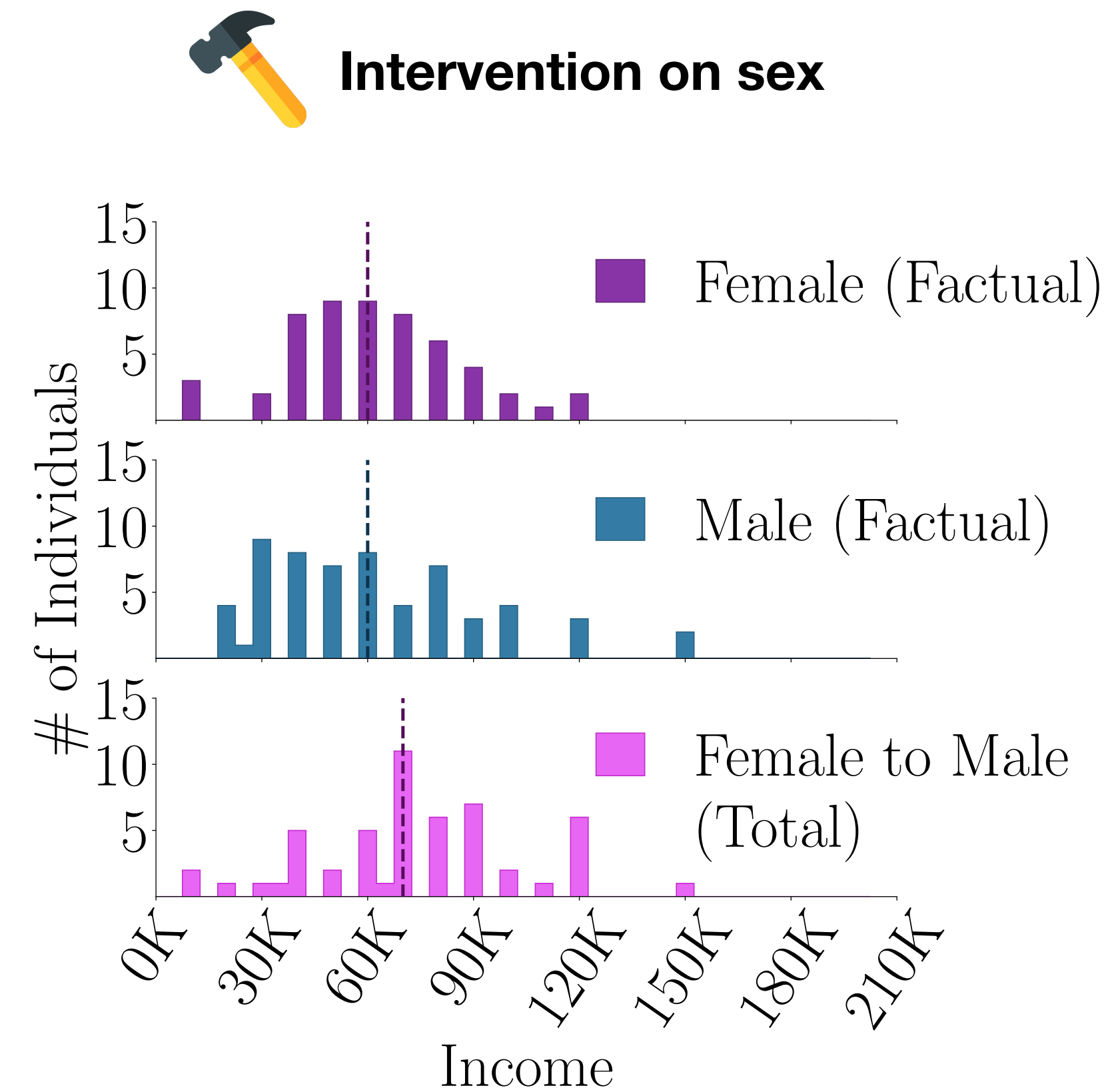


Intervention on sex



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



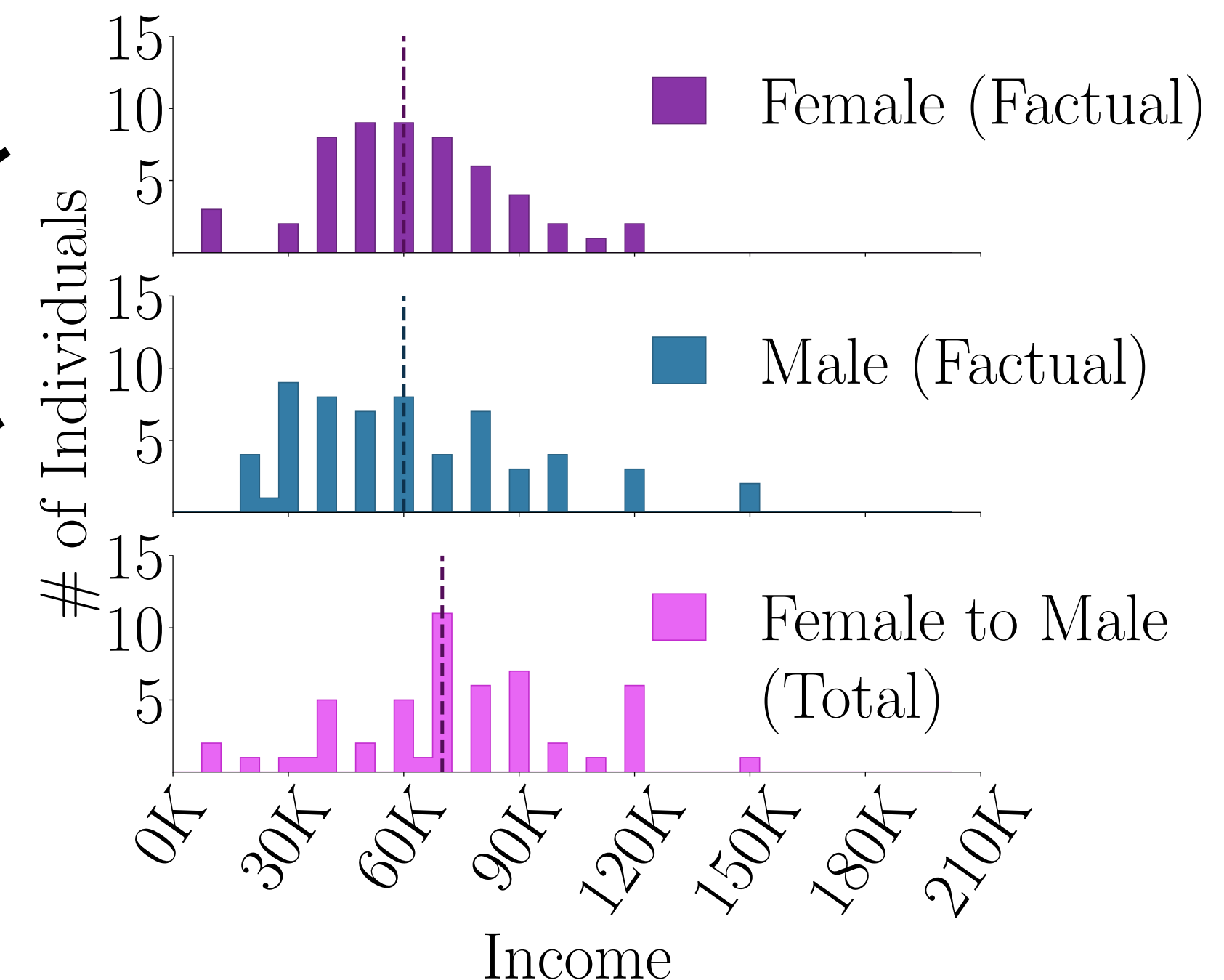
Do the LLMs show signs of bias?

- 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



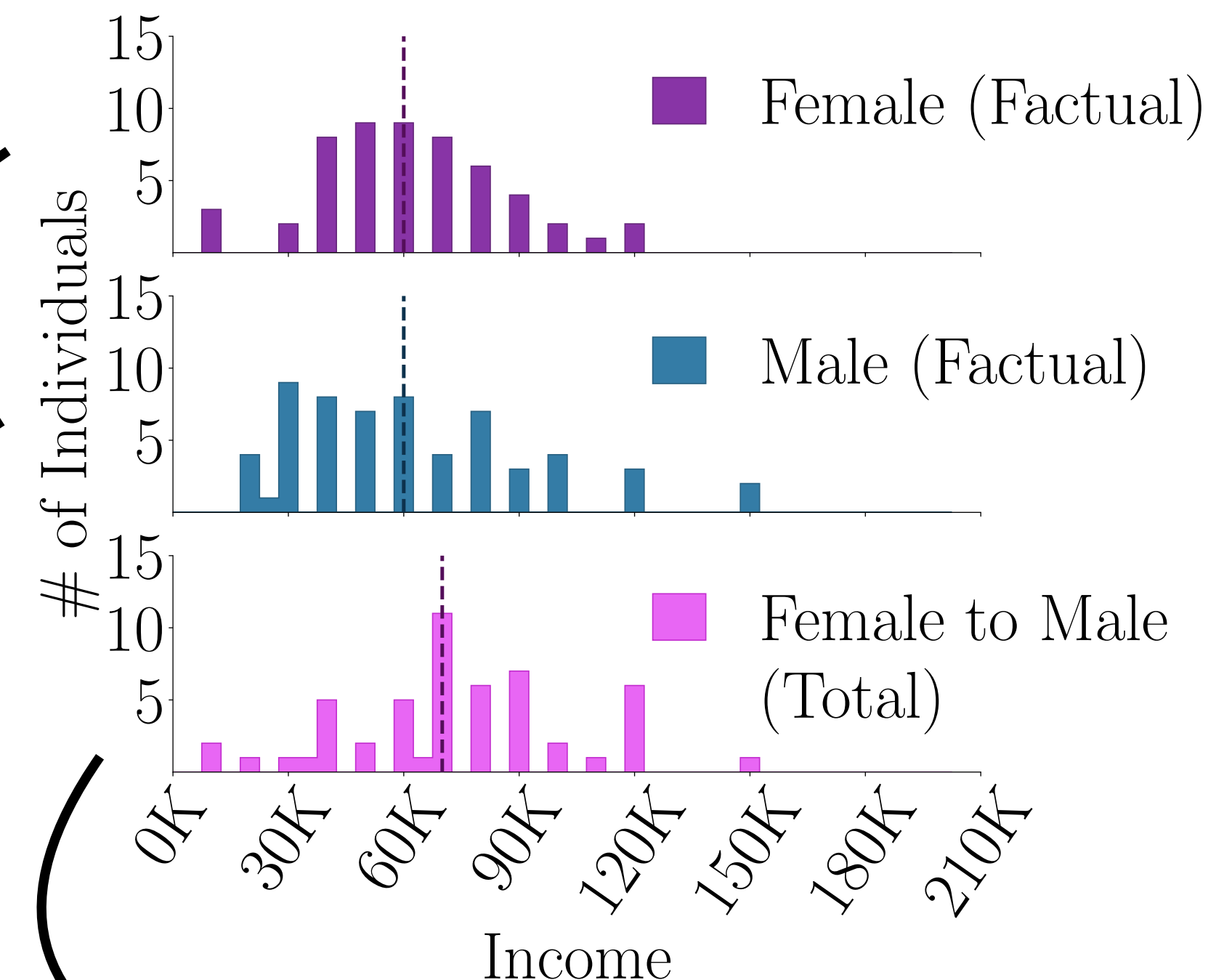
Do the LLMs show signs of bias?

- 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!



Conclusion

Conclusion

- Causal methods are promising for evaluating and improving the world models of LLMs

Conclusion

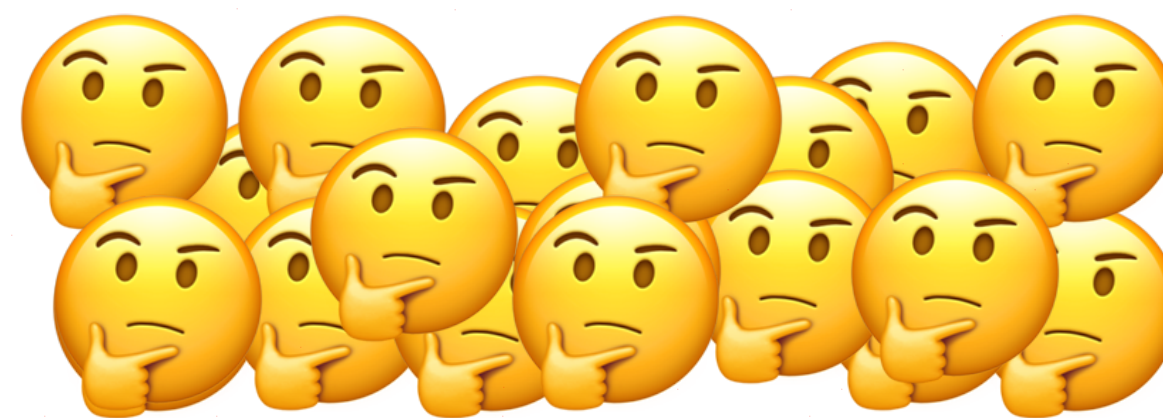
- Causal methods are promising for evaluating and improving the world models of LLMs
- Hidden biases not always observable just from comparing distributions of their responses

Conclusion

- Causal methods are promising for evaluating and improving the world models of LLMs
- Hidden biases not always observable just from comparing distributions of their responses
- Still long way until we fully understand LLMs' behavior

Conclusion

- Causal methods are promising for evaluating and improving the world models of LLMs
- Hidden biases not always observable just from comparing distributions of their responses
- Still long way until we fully understand LLMs' behavior



Discussion



stsirtsis@mpi-sws.org



@stratis_