Counterfactual Token Generation in Large Language Models

Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez

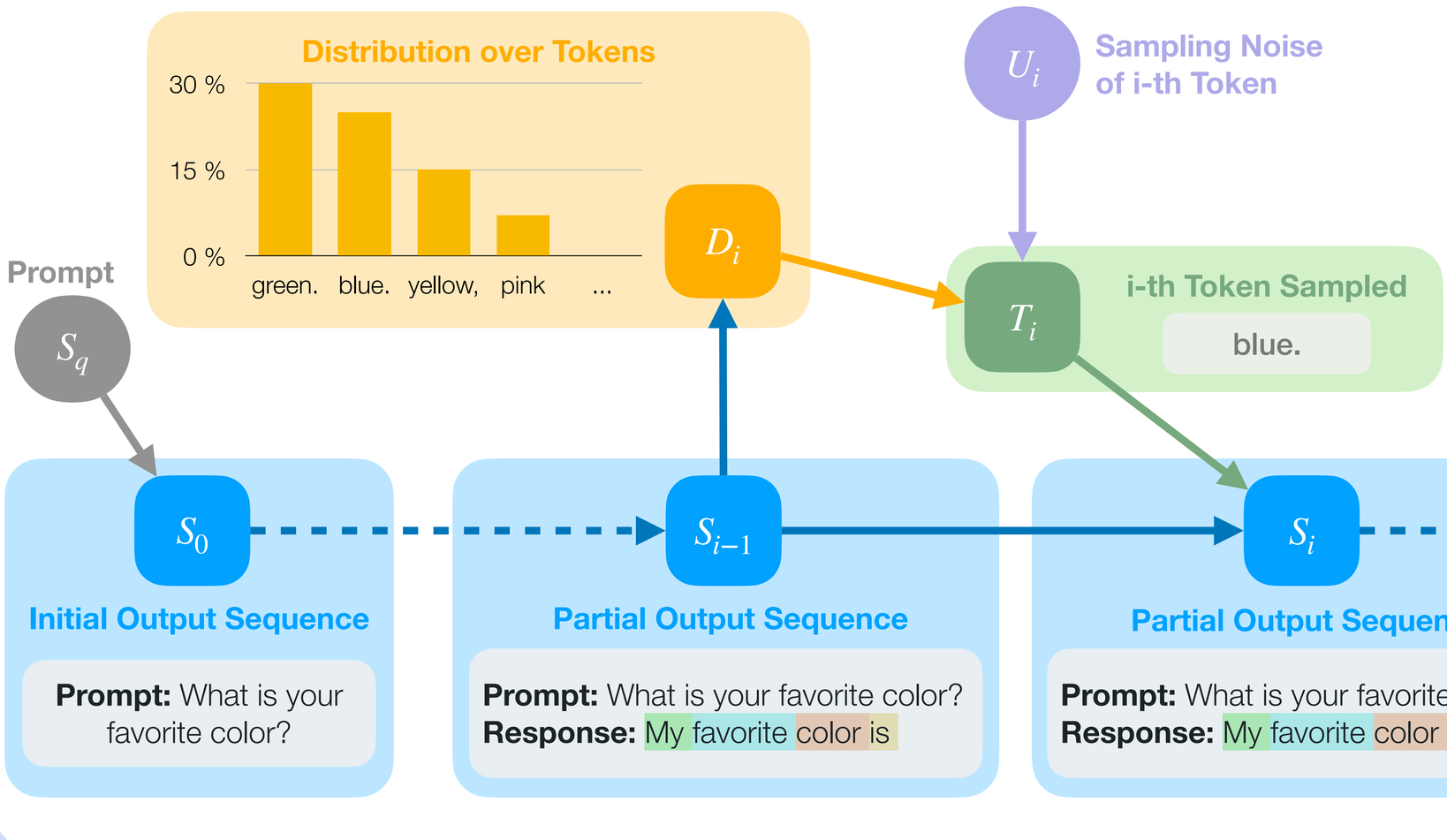
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

ETH zürich

Paper

Code

LLMs as Structural Causal Models



Interventional generation with unmodified input

Prompt: What is your favorite color?
Response: My favorite color is **blue**. Thanks for asking. Do you also like it?

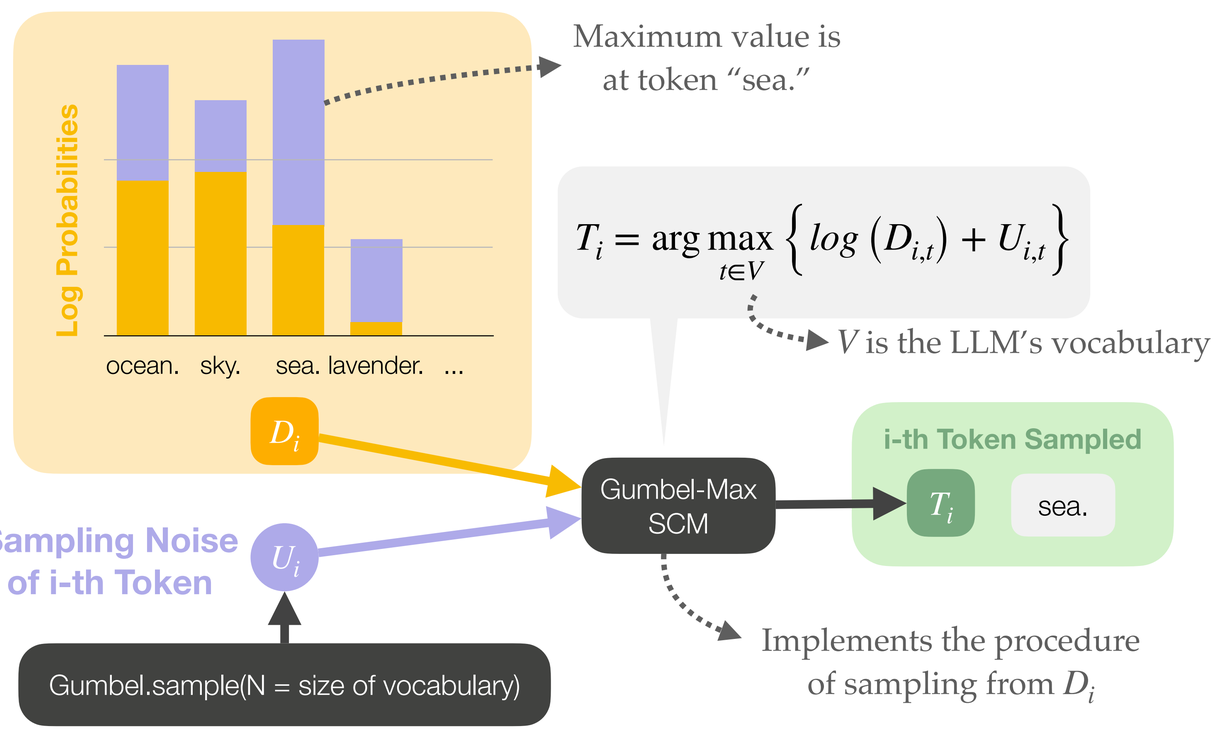
Interventional generation with modified input

Prompt: What is your favorite color?
Response: My favorite color is **purple**. Thanks for asking. What is your favorite?

Counterfactual generation with modified input

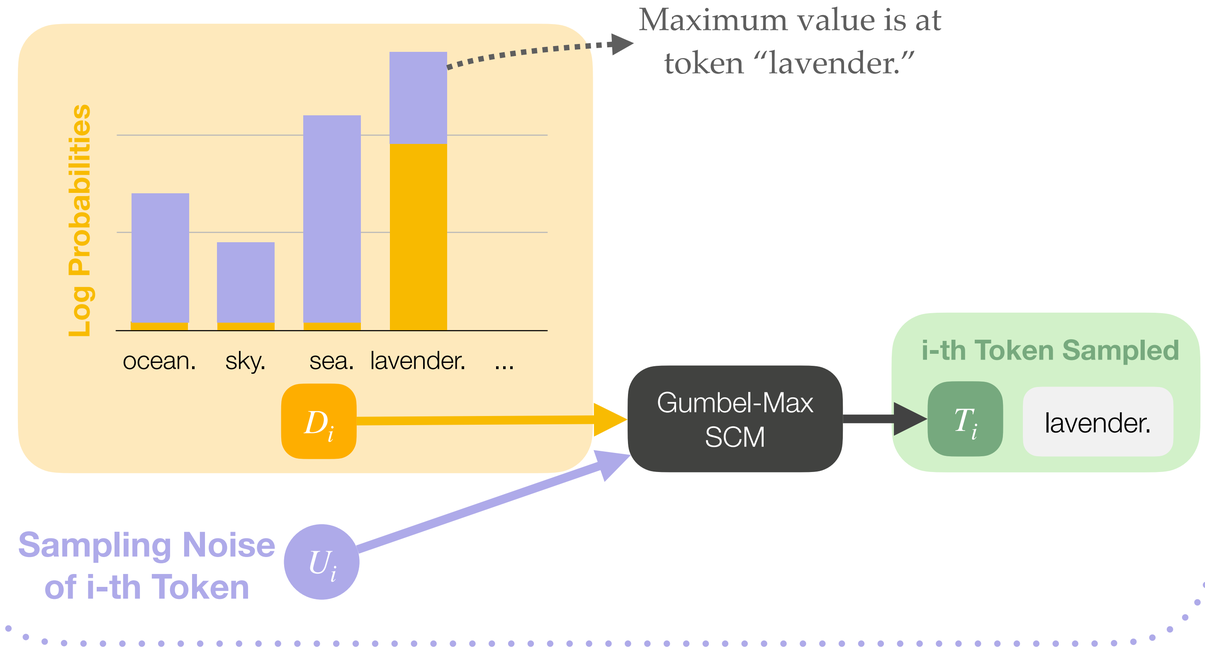
Prompt: What is your favorite color?
Response: My favorite color is **purple**. It is the color of fresh lavender.

Counterfactual Token Generation Using Gumbel-Max SCMs



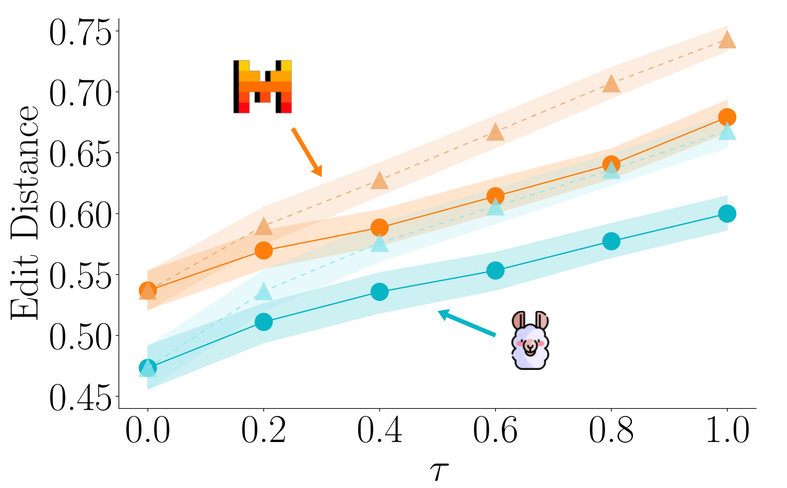
Which token would have been chosen, **counterfactually**?

🤔 **blue → purple**



Experiments `Llama 3 8B-Instruct & Ministral-8B-Instruct`

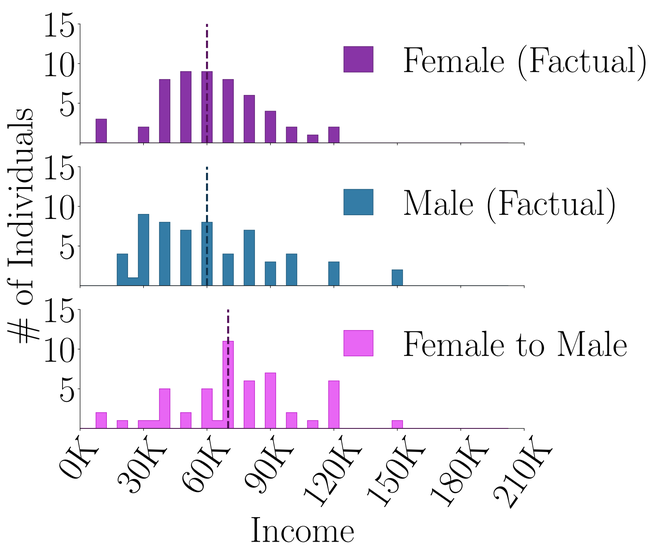
Text Similarity



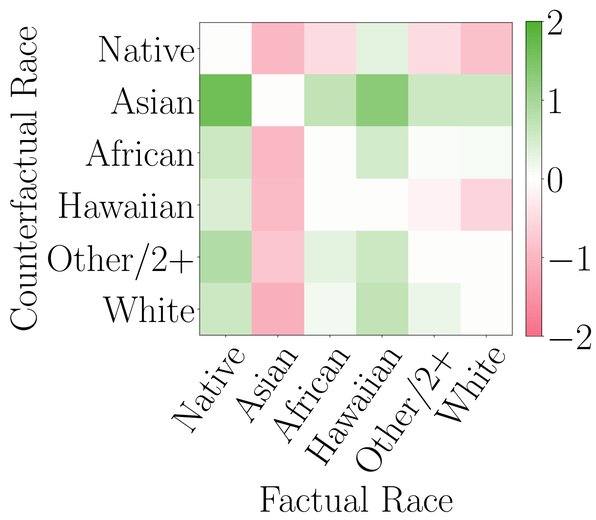
Interventional Token Generation

Counterfactual Token Generation

Discovering Model Biases



`Llama:` Distribution of factual and counterfactual income upon intervention on sex



`Ministral:` Change in education level upon intervention on race