Large language models

simplify building

natural

language

interfaces

Agenda

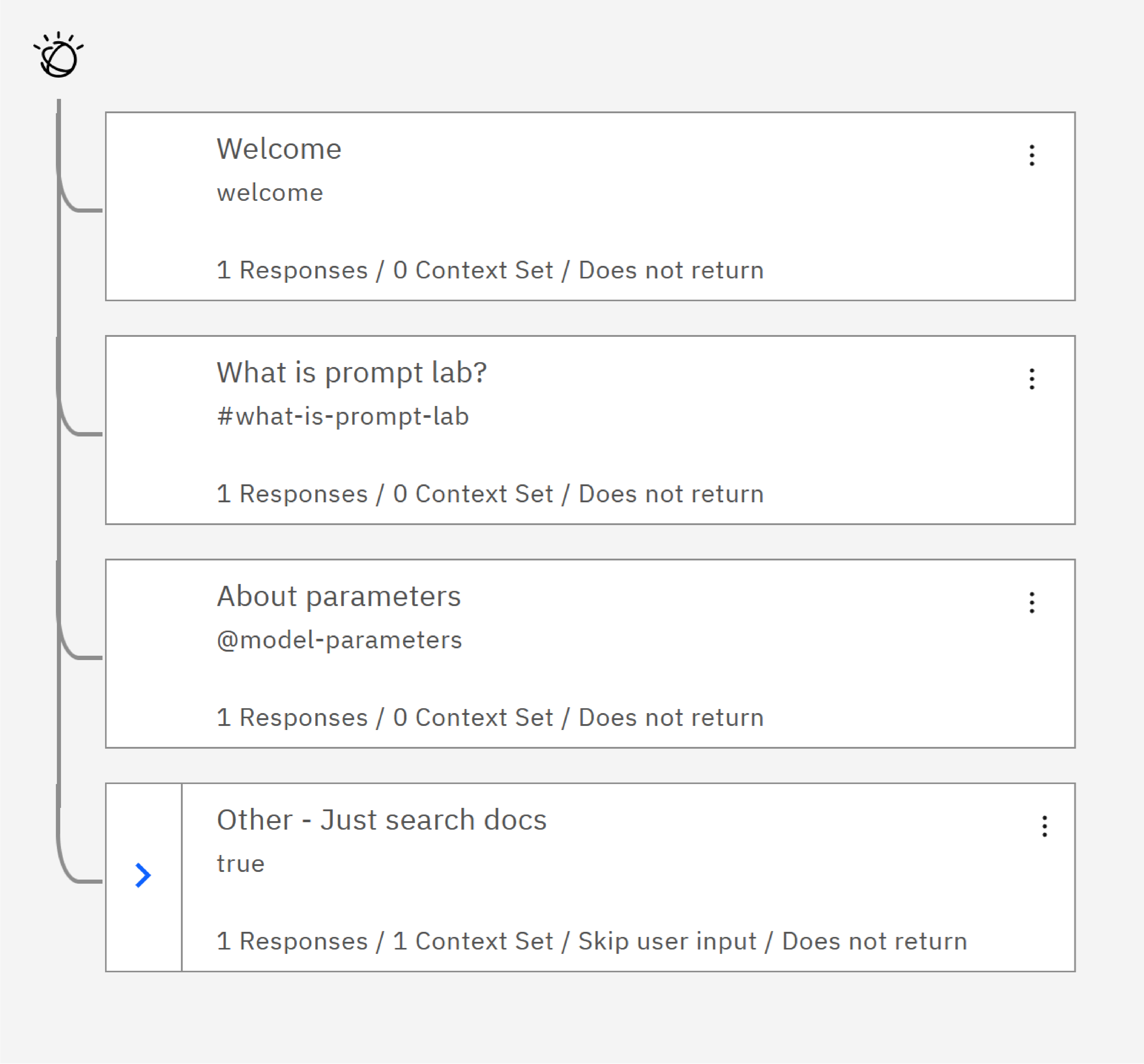
1. Before LLMs
2. After LLMs

Use case

Create a question-answering solution to help users of watsonx.ai be more productive.

Knowledge base: Product documentation

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

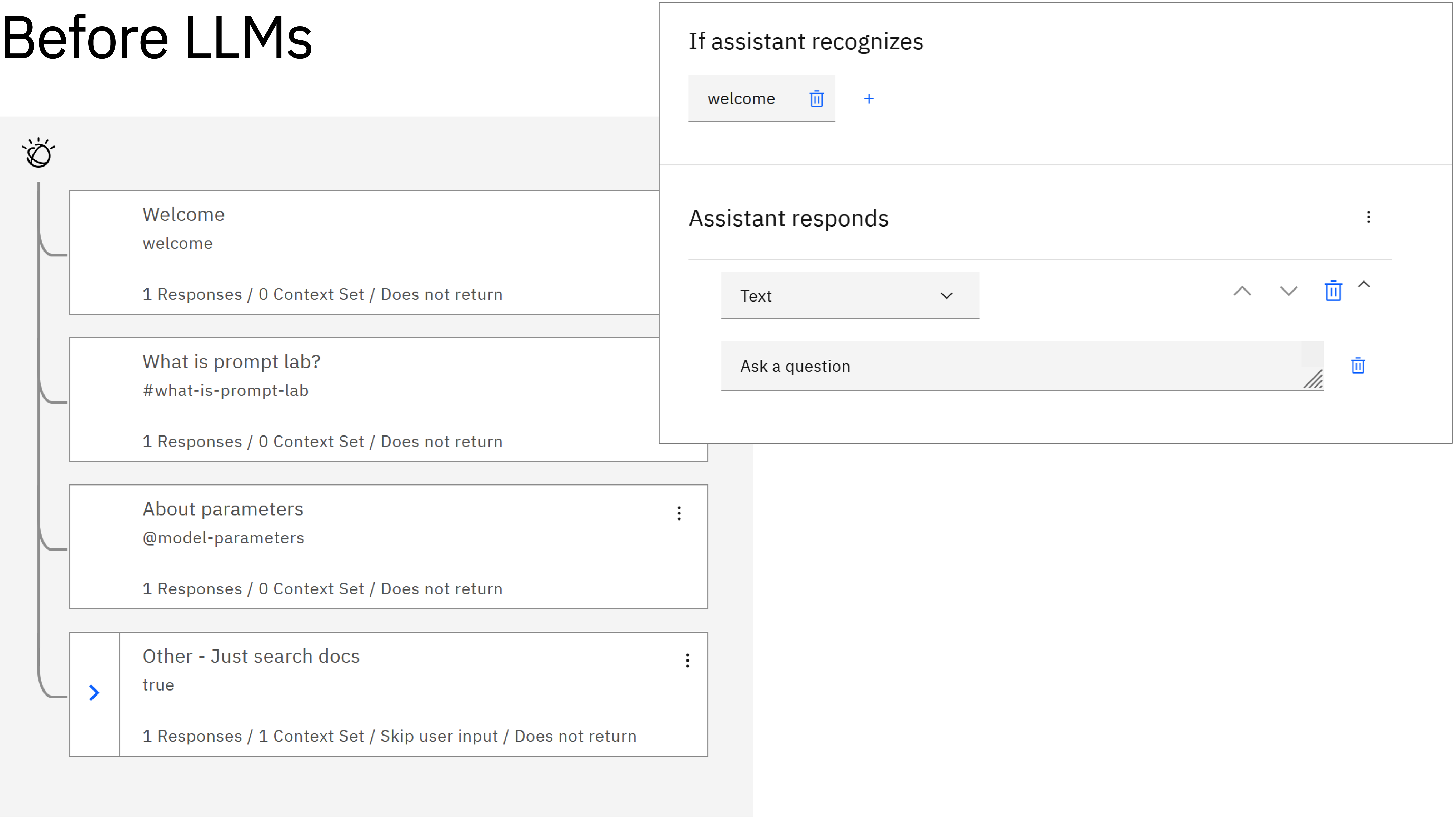
Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

To understand user input required using NLP techniques:

- Stop word removal
- Stemming
- Classification
- Part of speech detection
- Entity extraction
- …

Building a solution to tease apart the meaning of user input this way is time-consuming, and the result can be disappointing.

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

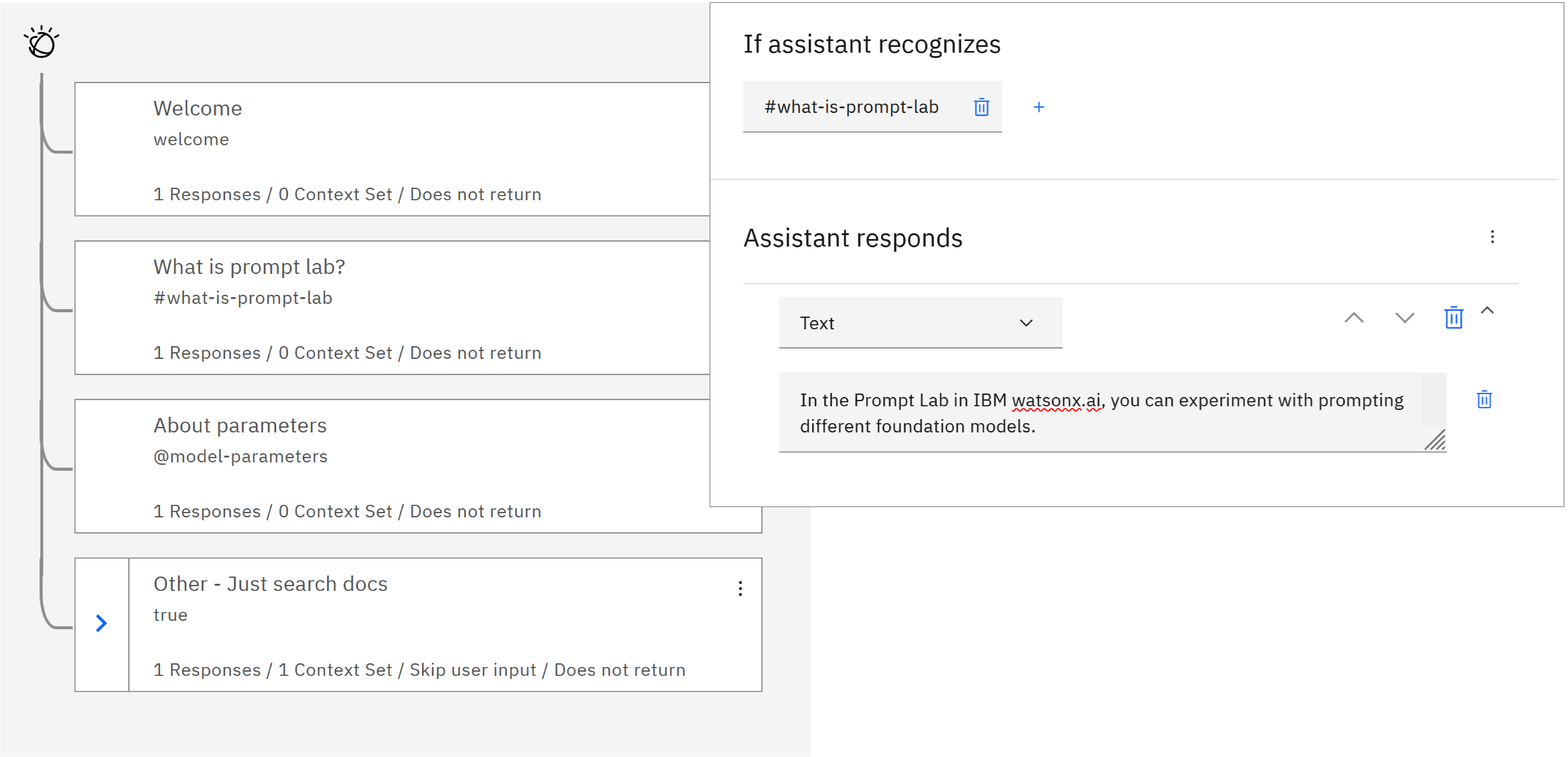
About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

If assistant recognizes
welcome

Assistant responds
Text
Ask a question

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

If assistant recognizes

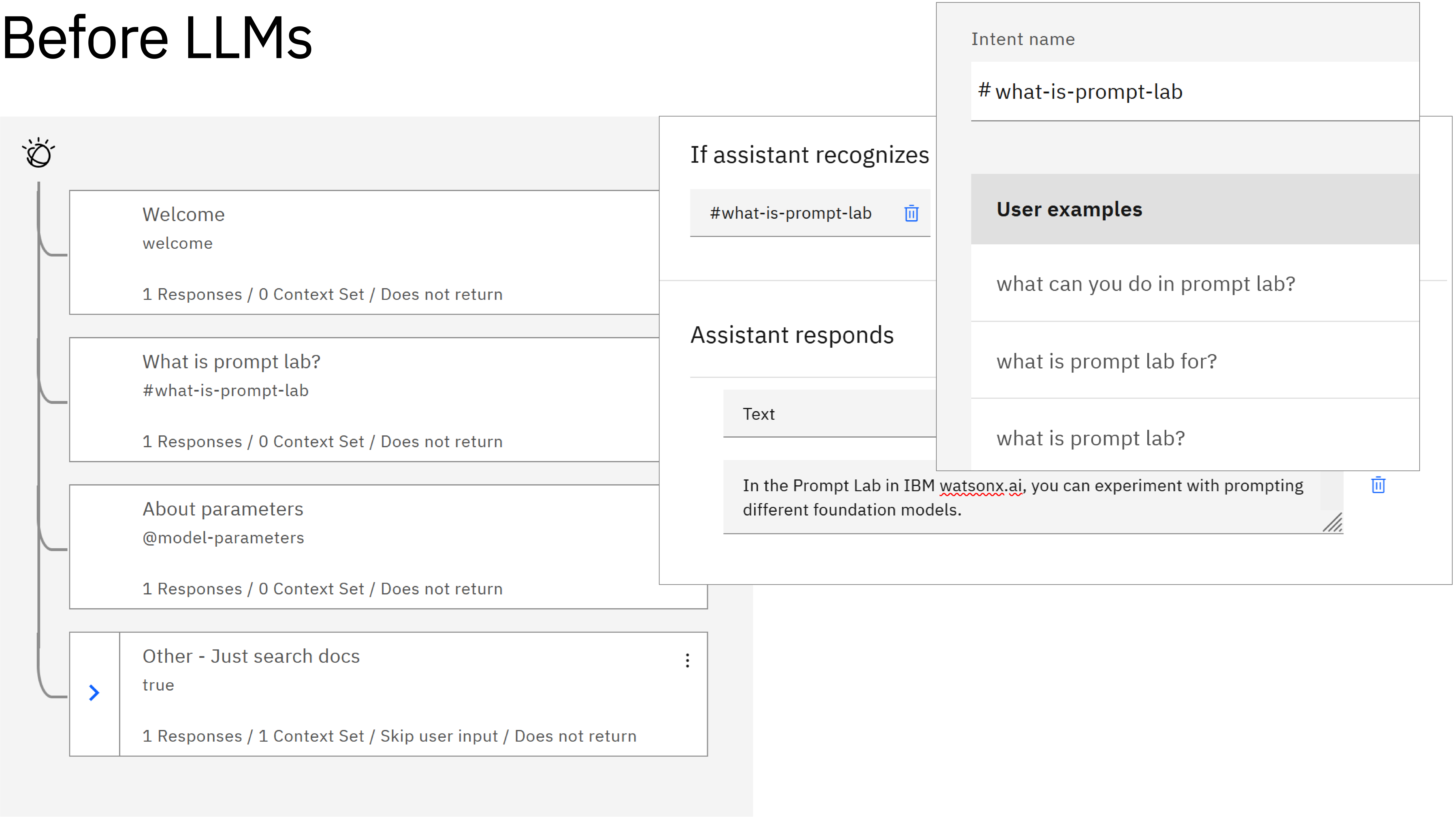
#what-is-prompt-lab

Assistant responds

Text

In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different foundation models.

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

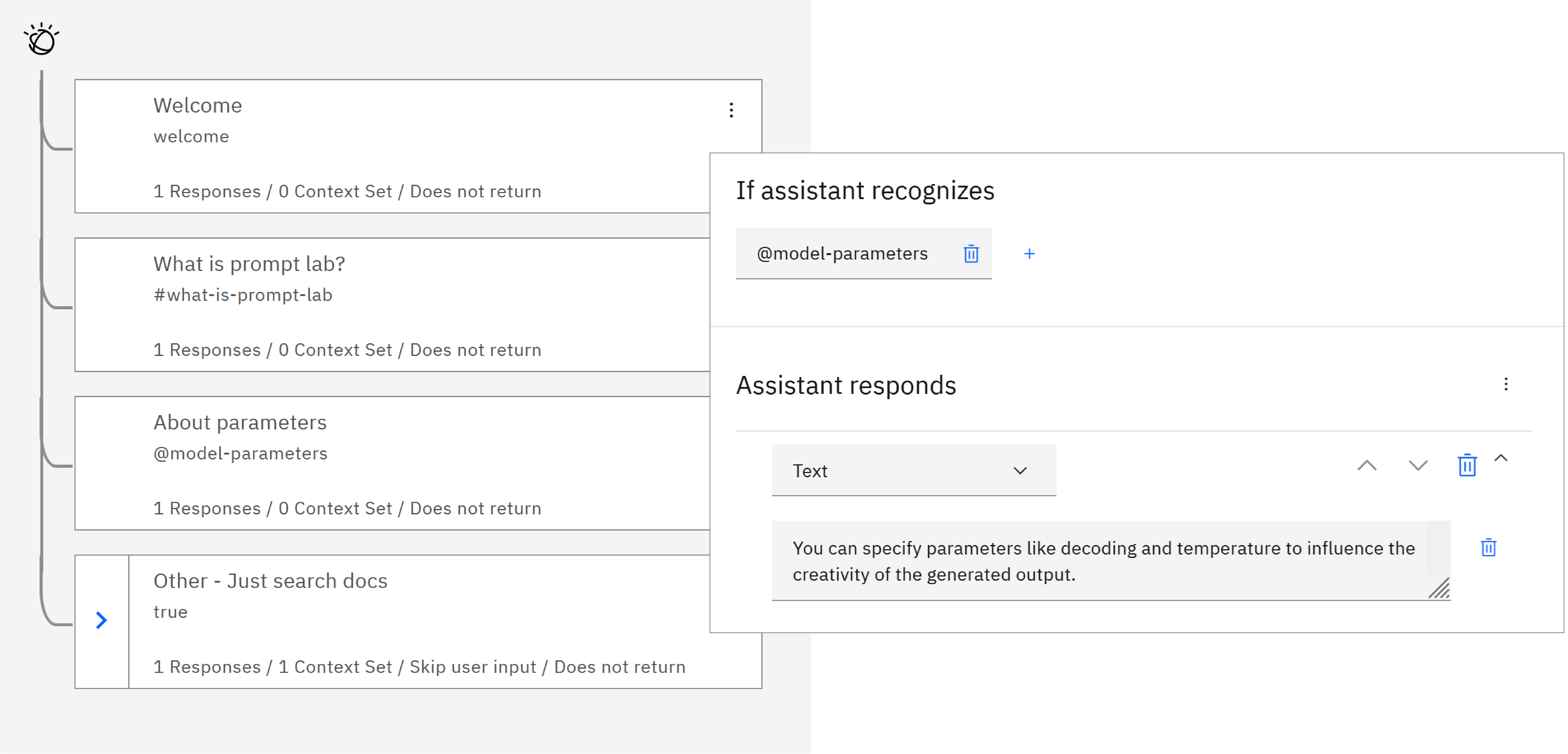
If assistant recognizes
#what-is-prompt-lab

Assistant responds
Text
In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different foundation models.

Intent name
# what-is-prompt-lab

User examples
what can you do in prompt lab?
what is prompt lab for?
what is prompt lab?

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

If assistant recognizes

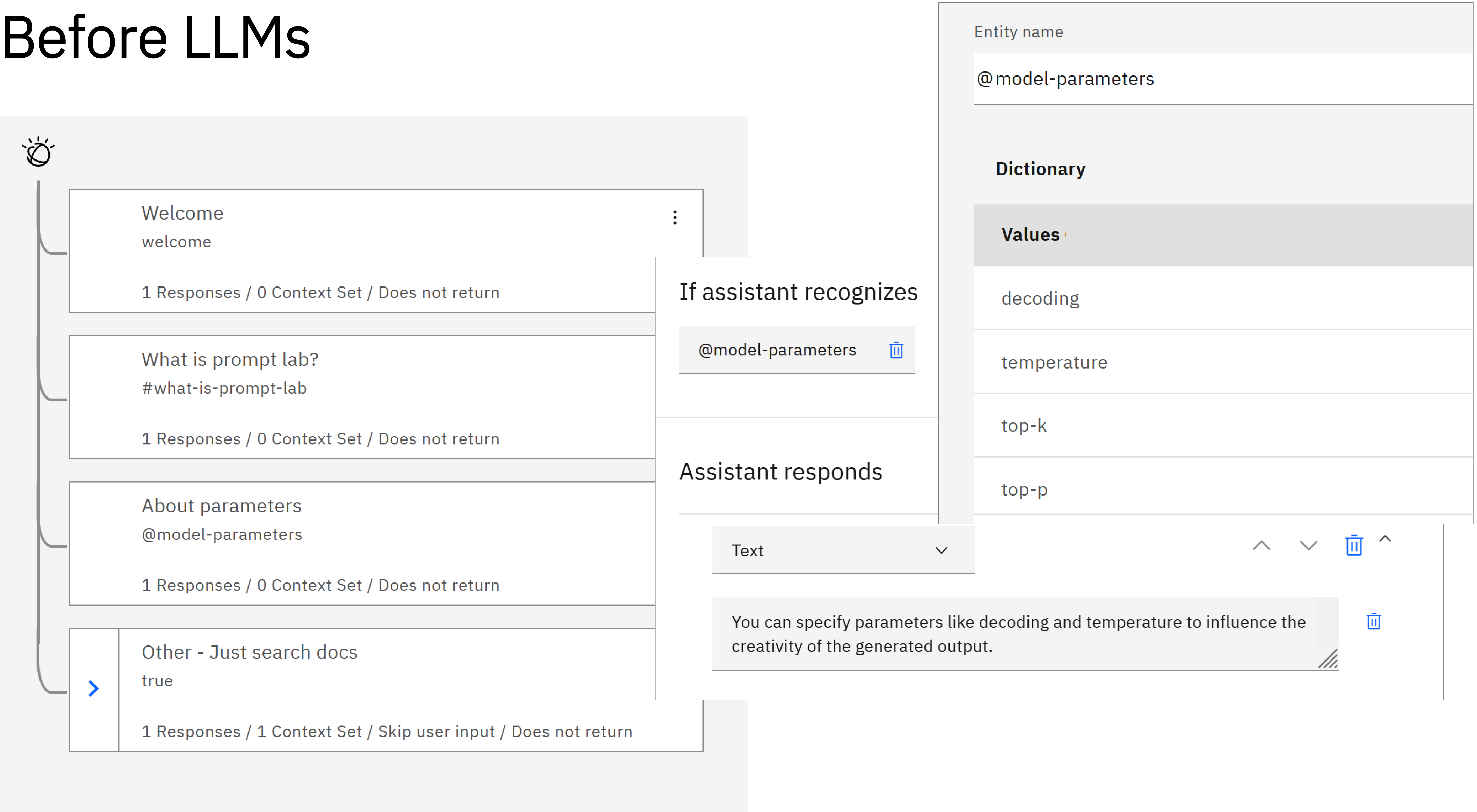
@model-parameters

Assistant responds

Text

You can specify parameters like decoding and temperature to influence the creativity of the generated output.

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not return

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not return

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

If assistant recognizes

@model-parameters

Assistant responds

Text

You can specify parameters like decoding and temperature to influence the creativity of the generated output.

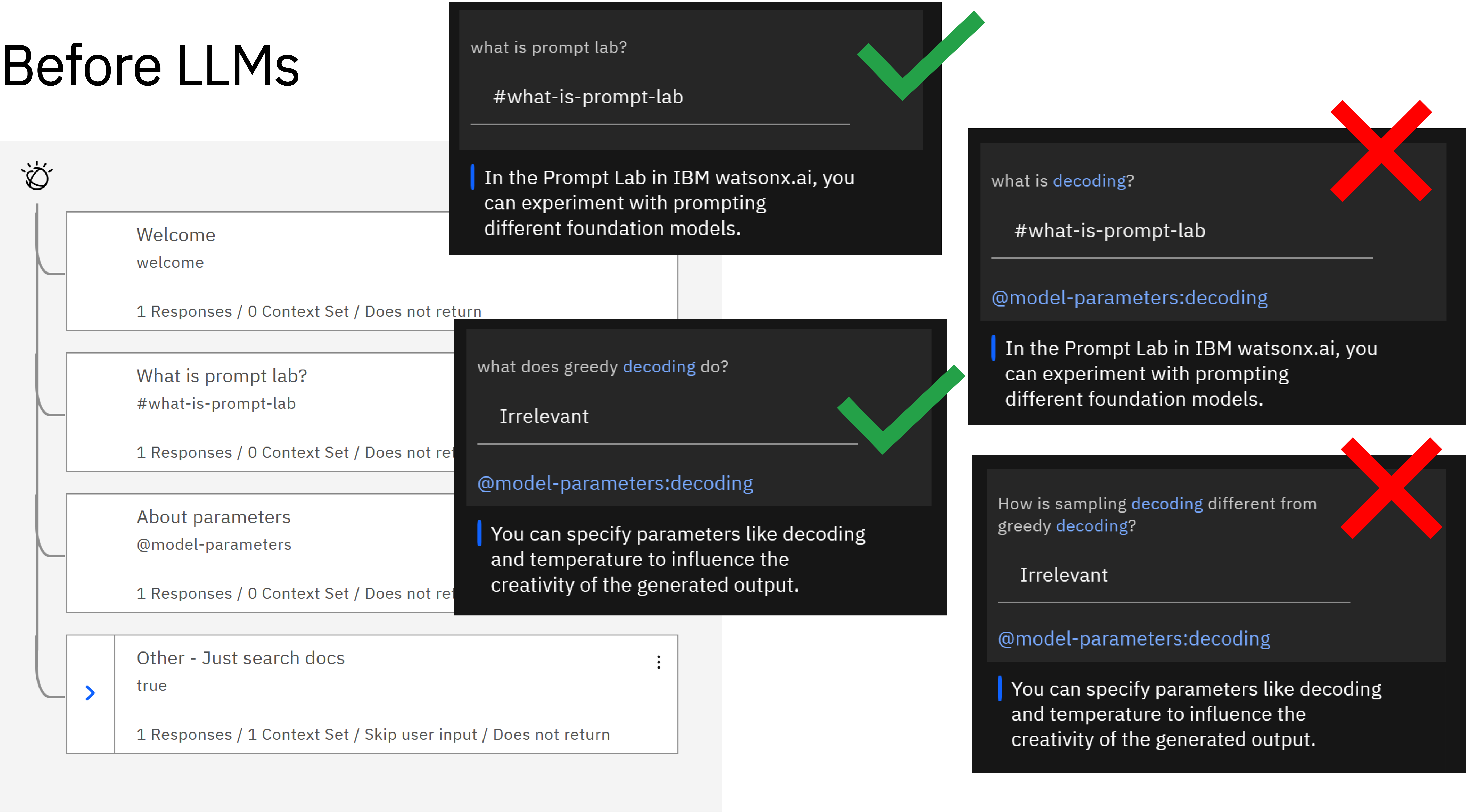
Entity name

@model-parameters

Dictionary

Values
decoding
temperature
top-k
top-p

Before LLMs



Welcome
welcome
1 Responses / 0 Context Set / Does not return

What is prompt lab?
#what-is-prompt-lab
1 Responses / 0 Context Set / Does not ret

About parameters
@model-parameters
1 Responses / 0 Context Set / Does not ret

Other - Just search docs
true
1 Responses / 1 Context Set / Skip user input / Does not return

what is prompt lab?
#what-is-prompt-lab
In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different foundation models.
✓

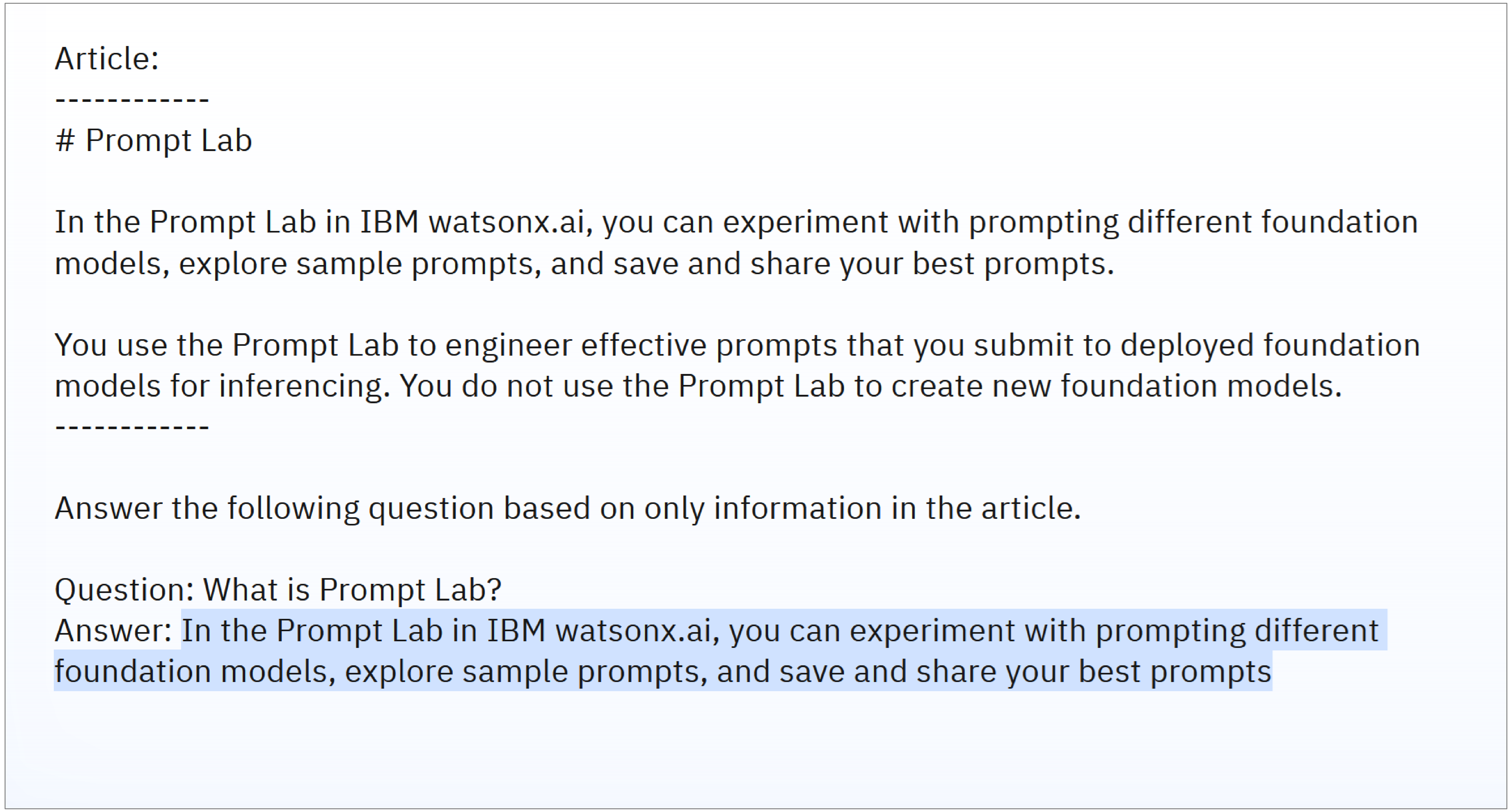
what does greedy decoding do?
Irrelevant
@model-parameters:decoding
You can specify parameters like decoding and temperature to influence the creativity of the generated output.
✓

what is decoding?
#what-is-prompt-lab
@model-parameters:decoding
In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different foundation models.
✗

How is sampling decoding different from greedy decoding?
Irrelevant
@model-parameters:decoding
You can specify parameters like decoding and temperature to influence the creativity of the generated output.
✗

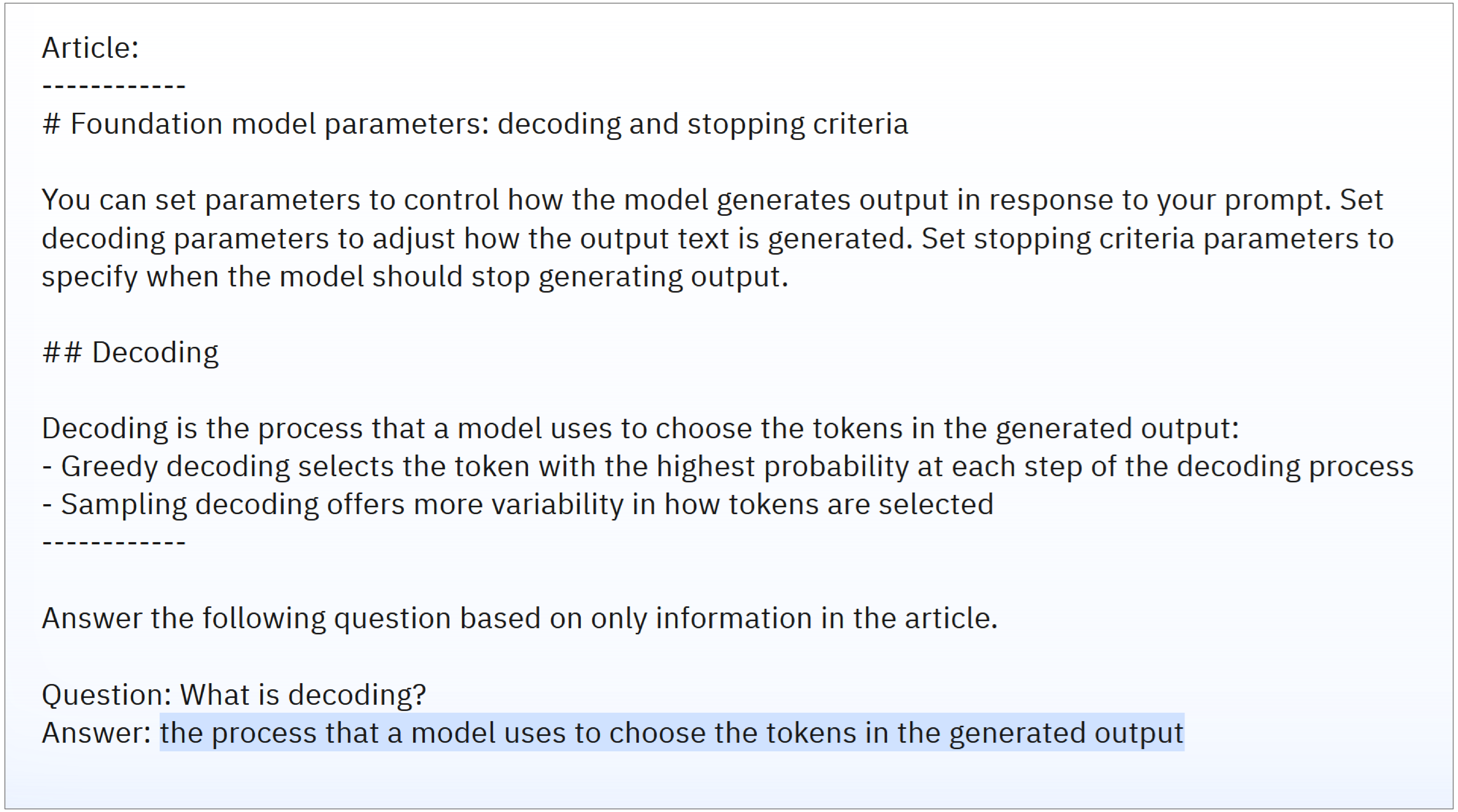
After LLMs

With appropriate prompting, and techniques like RAG, an LLM can handle most input without special handling.



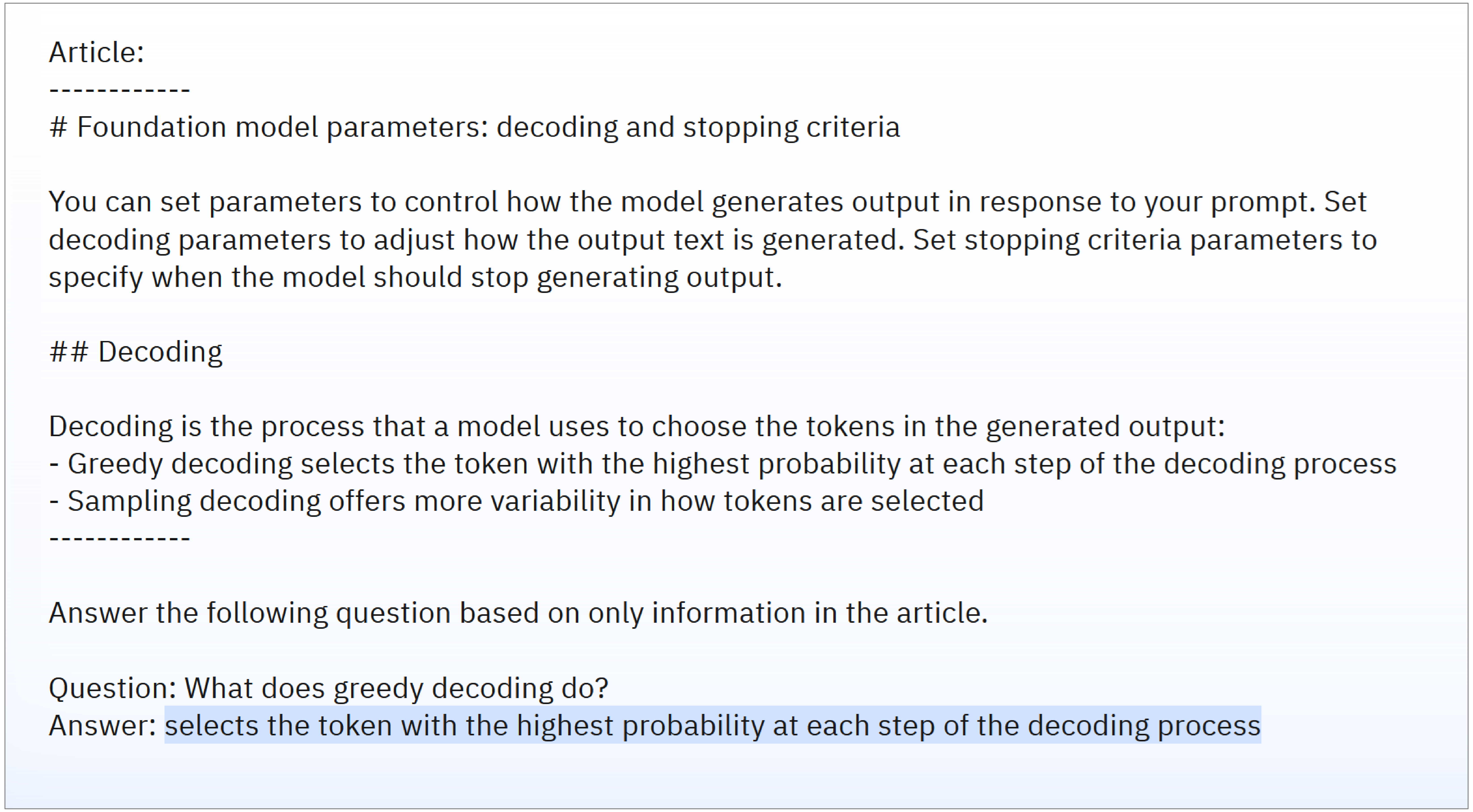
```
Article:
------------
# Prompt Lab

In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different foundation
models, explore sample prompts, and save and share your best prompts.

You use the Prompt Lab to engineer effective prompts that you submit to deployed foundation
models for inferencing. You do not use the Prompt Lab to create new foundation models.
------------

Answer the following question based on only information in the article.

Question: What is Prompt Lab?
Answer: In the Prompt Lab in IBM watsonx.ai, you can experiment with prompting different
foundation models, explore sample prompts, and save and share your best prompts
```

After LLMs



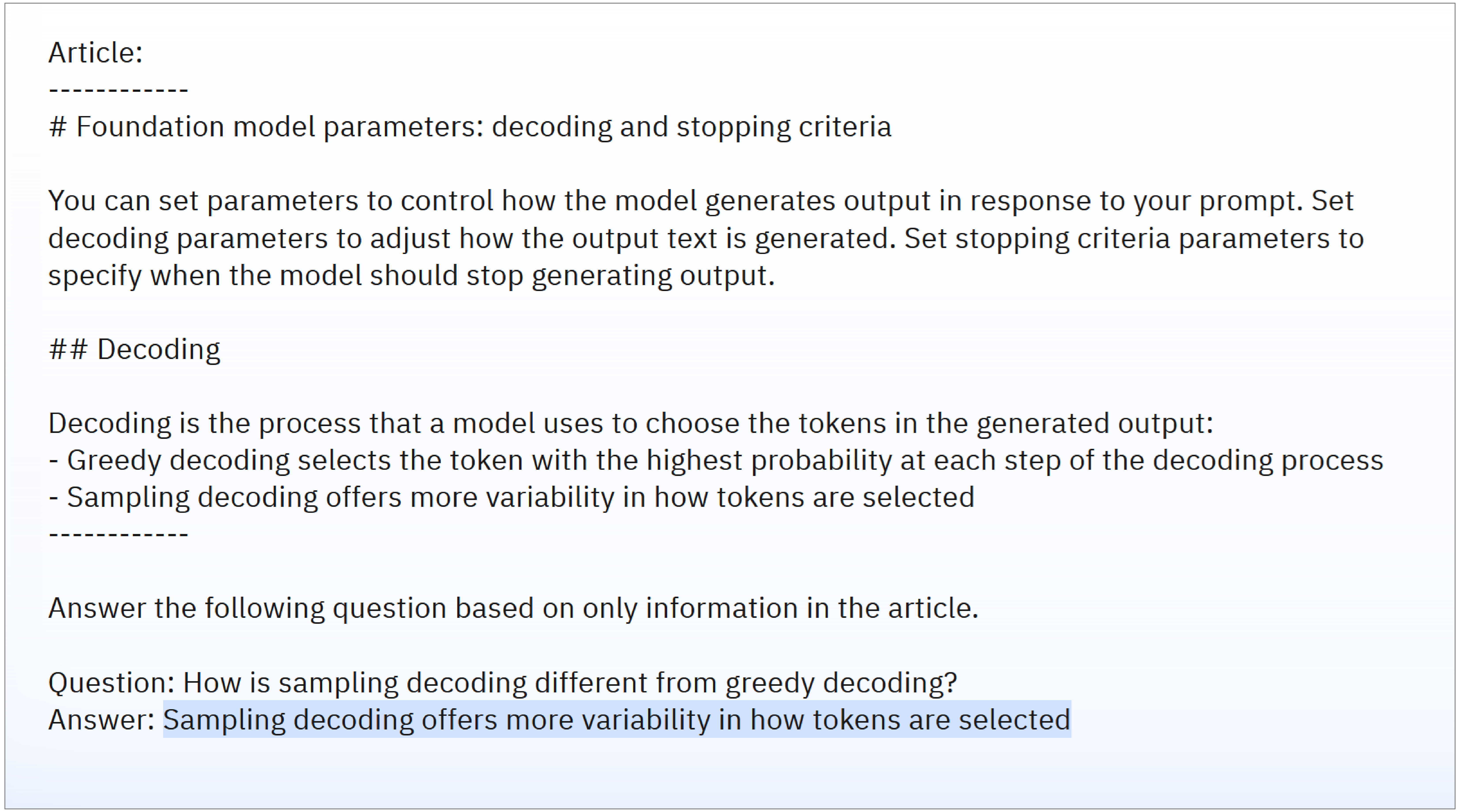
```
Article:
------------
# Foundation model parameters: decoding and stopping criteria

You can set parameters to control how the model generates output in response to your prompt. Set
decoding parameters to adjust how the output text is generated. Set stopping criteria parameters to
specify when the model should stop generating output.

## Decoding

Decoding is the process that a model uses to choose the tokens in the generated output:
- Greedy decoding selects the token with the highest probability at each step of the decoding process
- Sampling decoding offers more variability in how tokens are selected
------------

Answer the following question based on only information in the article.

Question: What is decoding?
Answer: the process that a model uses to choose the tokens in the generated output
```

After LLMs



```
Article:
------------
# Foundation model parameters: decoding and stopping criteria

You can set parameters to control how the model generates output in response to your prompt. Set
decoding parameters to adjust how the output text is generated. Set stopping criteria parameters to
specify when the model should stop generating output.

## Decoding

Decoding is the process that a model uses to choose the tokens in the generated output:
- Greedy decoding selects the token with the highest probability at each step of the decoding process
- Sampling decoding offers more variability in how tokens are selected
------------

Answer the following question based on only information in the article.

Question: What does greedy decoding do?
Answer: selects the token with the highest probability at each step of the decoding process
```

After LLMs



```
Article:
------------
# Foundation model parameters: decoding and stopping criteria

You can set parameters to control how the model generates output in response to your prompt. Set
decoding parameters to adjust how the output text is generated. Set stopping criteria parameters to
specify when the model should stop generating output.

## Decoding

Decoding is the process that a model uses to choose the tokens in the generated output:
- Greedy decoding selects the token with the highest probability at each step of the decoding process
- Sampling decoding offers more variability in how tokens are selected
------------

Answer the following question based on only information in the article.

Question: How is sampling decoding different from greedy decoding?
Answer: Sampling decoding offers more variability in how tokens are selected
```