Using pretrained foundation models: prompt engineering and fine-tuning

Generative AI Foundations on AWS

Emily Webber, Principal ML Specialist SA at AWS

Lesson 3 – Level 300

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

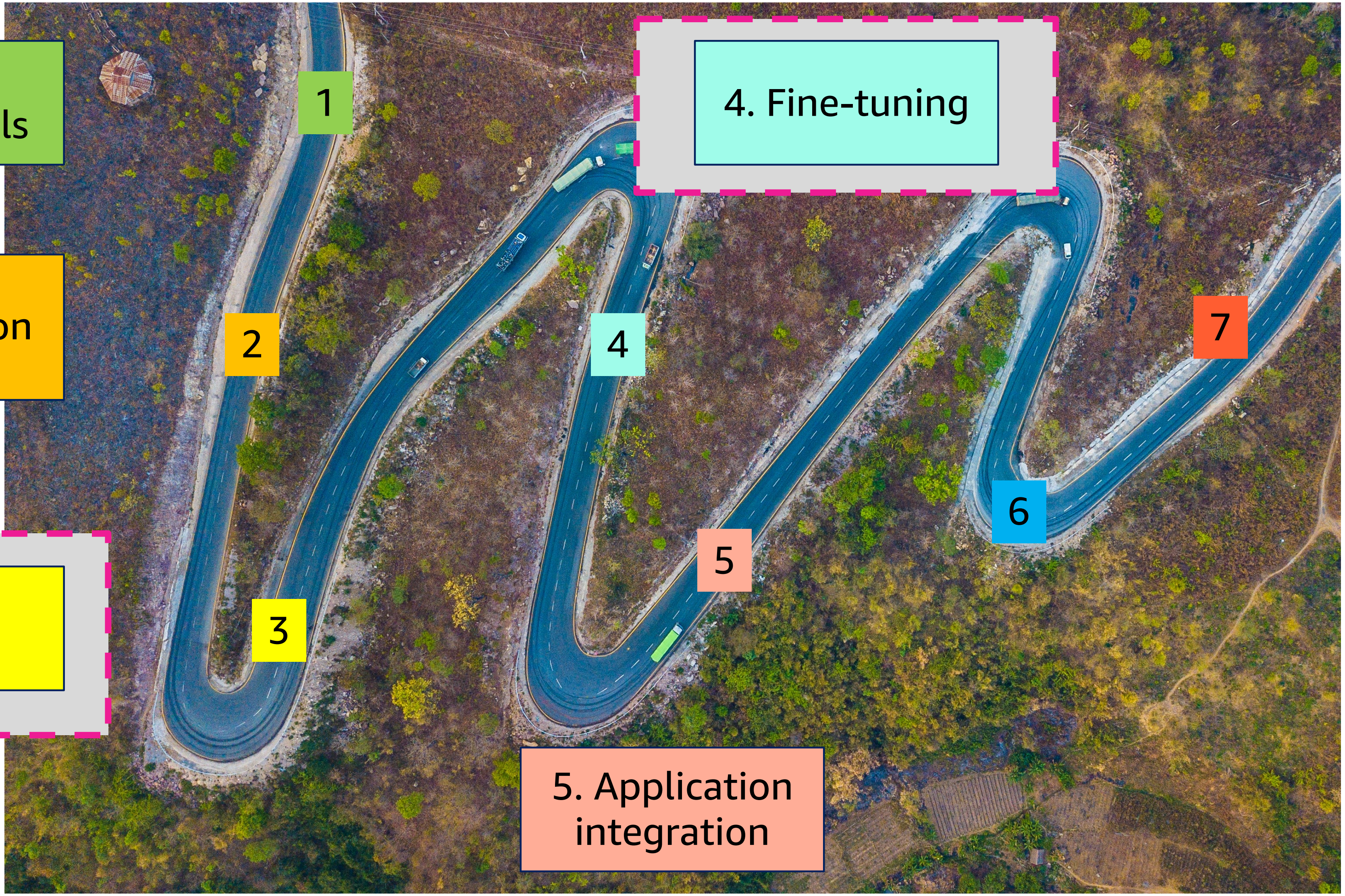
Today’s activities

- Prompt engineering: zero-shot, single-shot, few-shot
- Instruction examples: Summarization, classification, translation
- Fine-tuning: classic, parameter efficient, controlled
- Hands-on walk through: SageMaker JumpStart

Reminder – everything we discuss today
is possible on AWS and SageMaker!

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

So you found your foundation model, now what?



1. Primer on foundation models

2. Picked your starting foundation model

3. Prompt engineering

4. Fine-tuning

5. Application integration

6. Human feedback

7. Pretraining

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is a prompt?

- Tell me a story about a little girl riding her bicycle.
- Why was six afraid of seven?
- What is airspeed velocity of an unladed swallow?
- “Big is small,” by Georgia O’Keefe, high resolution, orange flowers, trending on art station, futuristic. **Not:** blurry, fuzzy, low quality

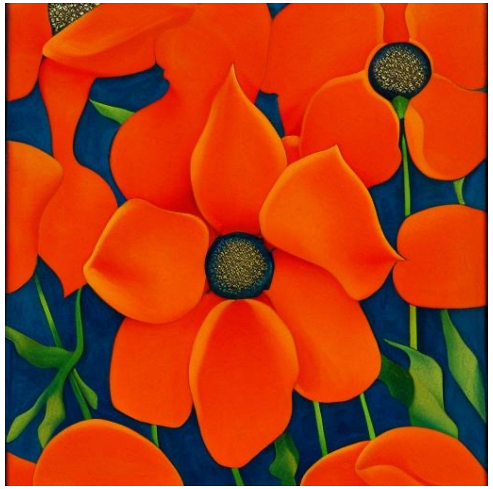
Your input to the model (aka “prompt”)

→

The foundation model

→

- There once was a girl named Susie ….
- Because seven ate nine!
- African or European swallow?

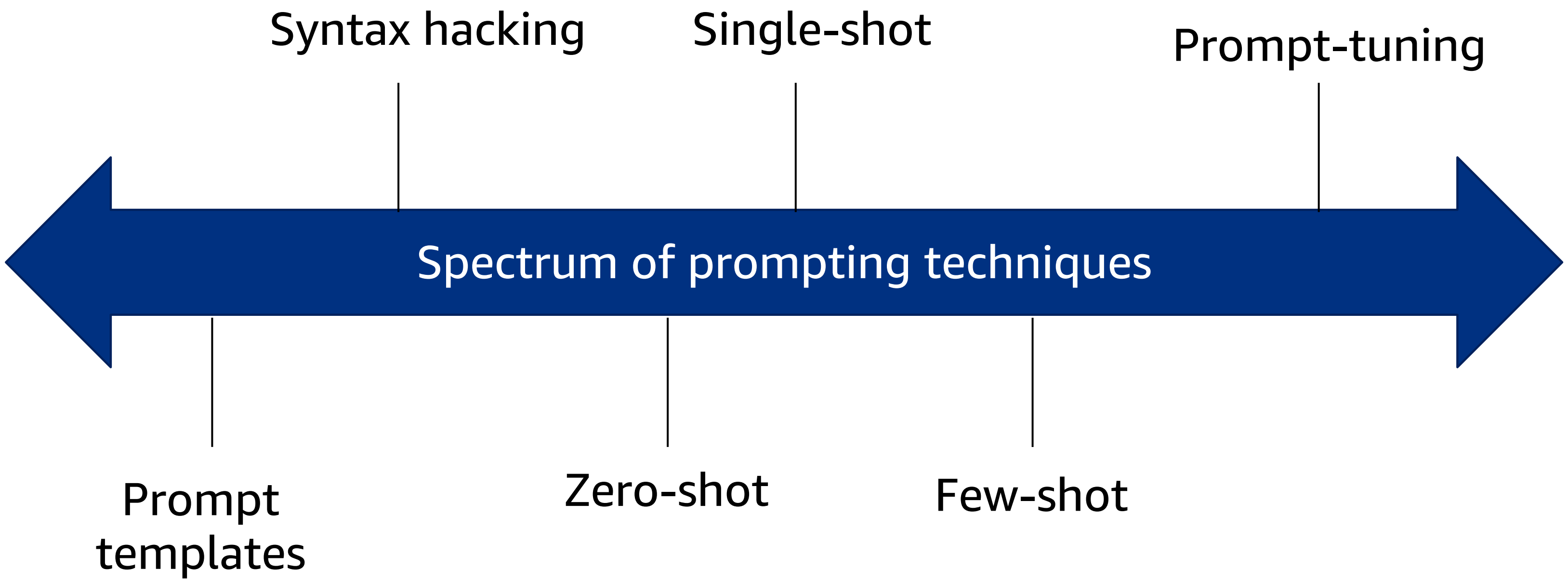


The response from the model

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is prompt engineering?

Getting a perfect answer from the model on your first try



Spectrum of prompting techniques

Syntax hacking

Single-shot

Prompt-tuning

Prompt templates

Zero-shot

Few-shot

Testing every combination until it works

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Instruction-tuned models make life easier

Prompt: Why was six afraid of seven?

Answer: Did they think so? That was the point. Six times out of eight was a good thing. And no one can say that if people believed so, they did. But if you look at the data, in

Answer: The origin of the phrase "Six afraid of seven" is unclear, but it is likely that it is a mnemonic device used to aid in the memorization of the order of the numbers from 1 to 10. The expression "Nine on Ten" is a mnemonic device used in a similar way.

Prompt: What is the difference between a mimosa and a samosa?

Answer: What we find is a mix of traits and characteristics that might be confused. Let's call them mimicry. In the mimosa you're not mimulating a color in the

Answer: A mimosa is a drink made of champagne and orange juice while a samosa is a type of savory Indian pastry often filled with spiced vegetables or meat. While both items have the word "samosa" in them, they are unrelated.

Before: GPT-2

After: AI21 Jurassic-2 Jumbo Instruct

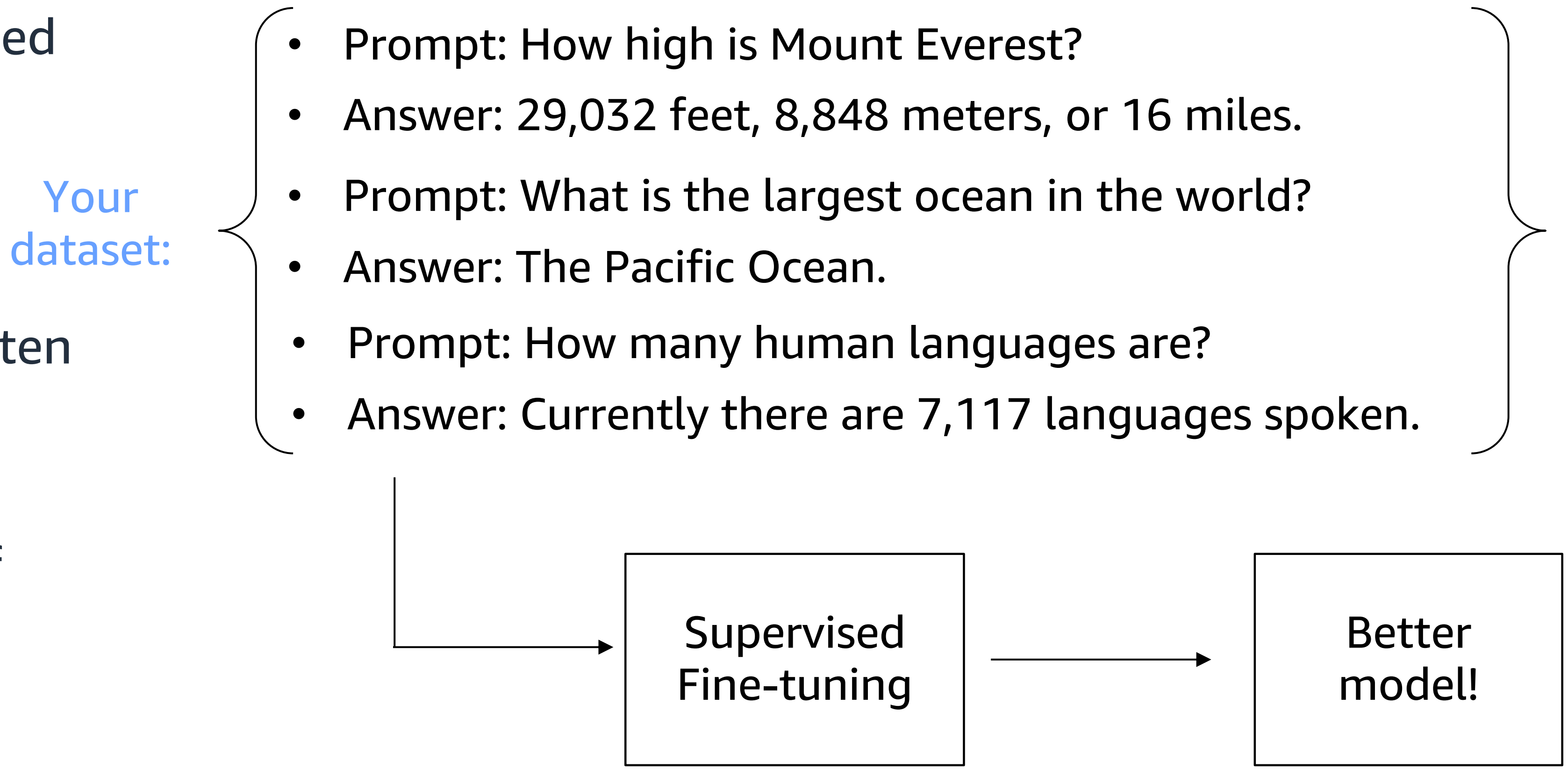
Hallucinated word!

Close, but not quite.

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Try to use instruction-tuned models when possible

- Instruction tuning uses supervised learning to adapt the model’s behavior
- Takes prompts and human-written responses to these prompts
- Ensures the model is capable of *following instructions*
- Unlocks instructions like “translate,” “summarize,” “extract” …



Your dataset:

- Prompt: How high is Mount Everest?
- Answer: 29,032 feet, 8,848 meters, or 16 miles.
- Prompt: What is the largest ocean in the world?
- Answer: The Pacific Ocean.
- Prompt: How many human languages are?
- Answer: Currently there are 7,117 languages spoken.

Pro tip: if your prompt isn’t working, make sure your model has been instruction-tuned.

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best case scenario – zero-shot prompting

In a perfect world you send any arbitrary question to a model and get back a perfect answer on your first try.

Stable diffusion: a delicious plate of nachos

Prompt: What are the basic ingredients of nachos?

Answer: Nachos are a delicious food, originally developed in Mexico City in the 1940s, and later popularized across the United States.

The basic ingredients of nachos are tortilla chips, cheese, and jalepeno peppers. In the US it is especially common to also include guacamole, sour cream, salsa, and ground beef in the dish.

Nachos are usually an appetizer, coming before the main entrée, but can be served as the primary meal in some settings, like for Super Bowl watch parties!

Pro tip: we call this "zero-shot" because you don't need to send any previous examples in the prompt.

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Alternative – single-shot prompting

If zero-shot fails, try adding an example to the prompt of what you want the model to produce.

The "single shot"

Prompt:
Appetizer: samosa
Main course: saag paneer

Appetizer: spinach dip
Main course:

Answer: burgers

Stable diffusion: a delicious bowl of spinach dip, appealing, award winning, high resolution
Not: disgusting, messy, weird

Pro tip: we call this "single-shot" because we provide *exactly one* example to the model in the prompt

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Alternative – few-shot prompting

If single-shot fails, add a few more examples to the prompt.

Stable diffusion: a cute panda bear, in the wild, closeup, photorealism, realistic
Not: cartoon, animation

A "few shots"

Prompt:
Data analysis: pandas

Statistical machine learning: sklearn

Deep learning: PyTorch

Front-end development:

Answer: Django

Pro tip: we call this "few" because we provide *multiple examples* to the model in the prompt

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using prompt engineering to solve summarization

The easiest way to do this is just to *paste the entire document into the prompt*. Then include a single instruction at the bottom of the prompt to summarize.

Prompt: The Pug is a breed of dog originally from China, with physically distinctive features of a wrinkly, short-muzzled face, and curled tail. The breed has a fine, glossy coat that comes in a variety of colors, most often fawn (light brown) or black, and a compact, square body with well developed and thick muscles all over the body.

Pugs were brought from China to Europe in the sixteenth century and were popularized in Western Europe by the House of Orange of the Netherlands, and the House of Stuart.[2] In the United Kingdom, in the nineteenth century, Queen Victoria developed a passion for pugs which she passed on to other members of the Royal Family.

Pugs are known for being sociable and gentle companion dogs.[3] The American Kennel Club describes the breed's personality as "even-tempered and charming".[4] Pugs remain popular into the twenty-first century, with some famous celebrity owners.
Summarize:

Answer: The Pug is a distinctive breed of dog originally from China, known for its wrinkly face, curled tail, and compact body. They were brought to Europe in the sixteenth century and gained popularity in Western Europe, including among the British Royal Family. Pugs are sociable and gentle companion dogs, described as "even-tempered and charming" by the American Kennel Club. They continue to be popular today, with some notable celebrity owners.

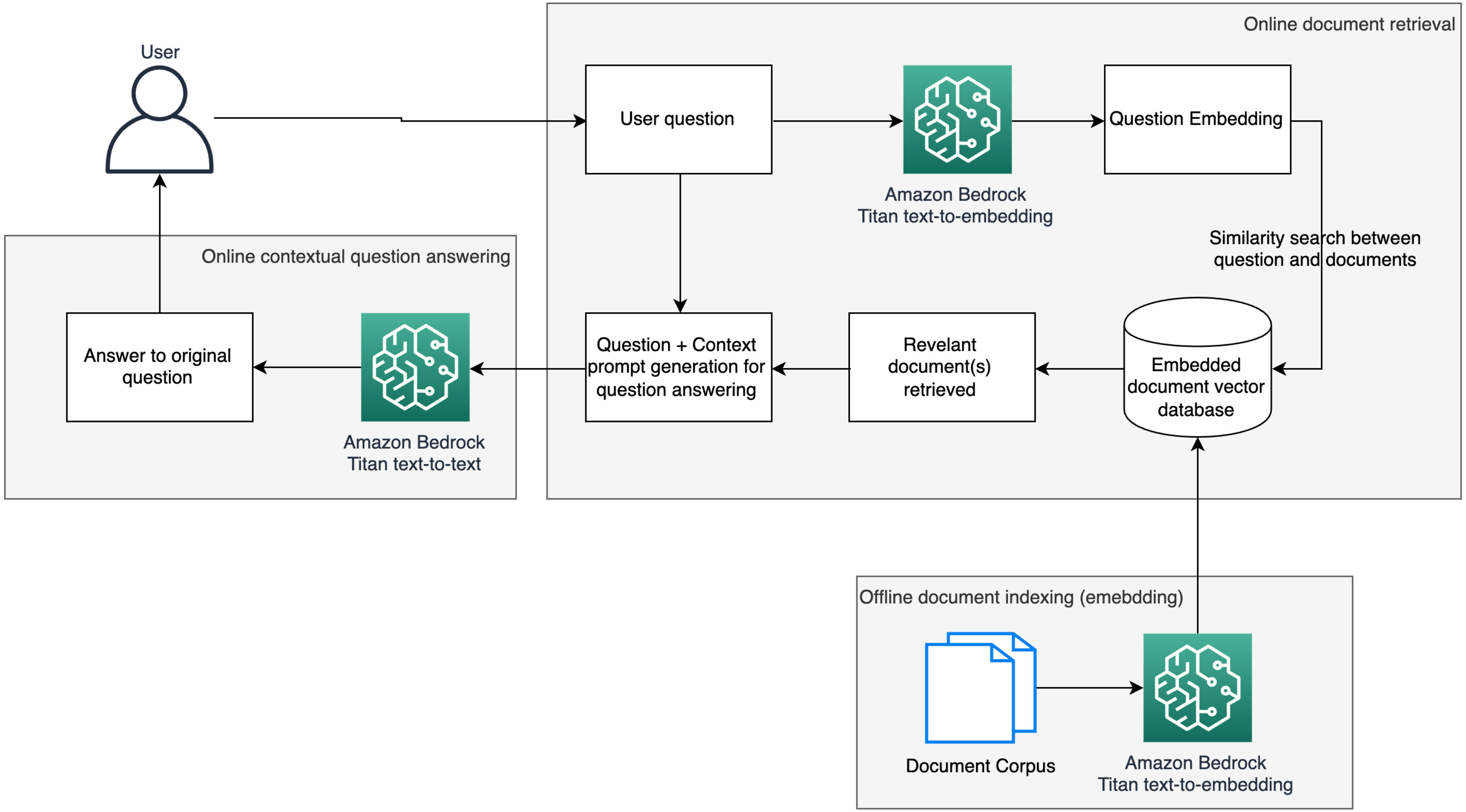
©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval augmented generation (RAG)

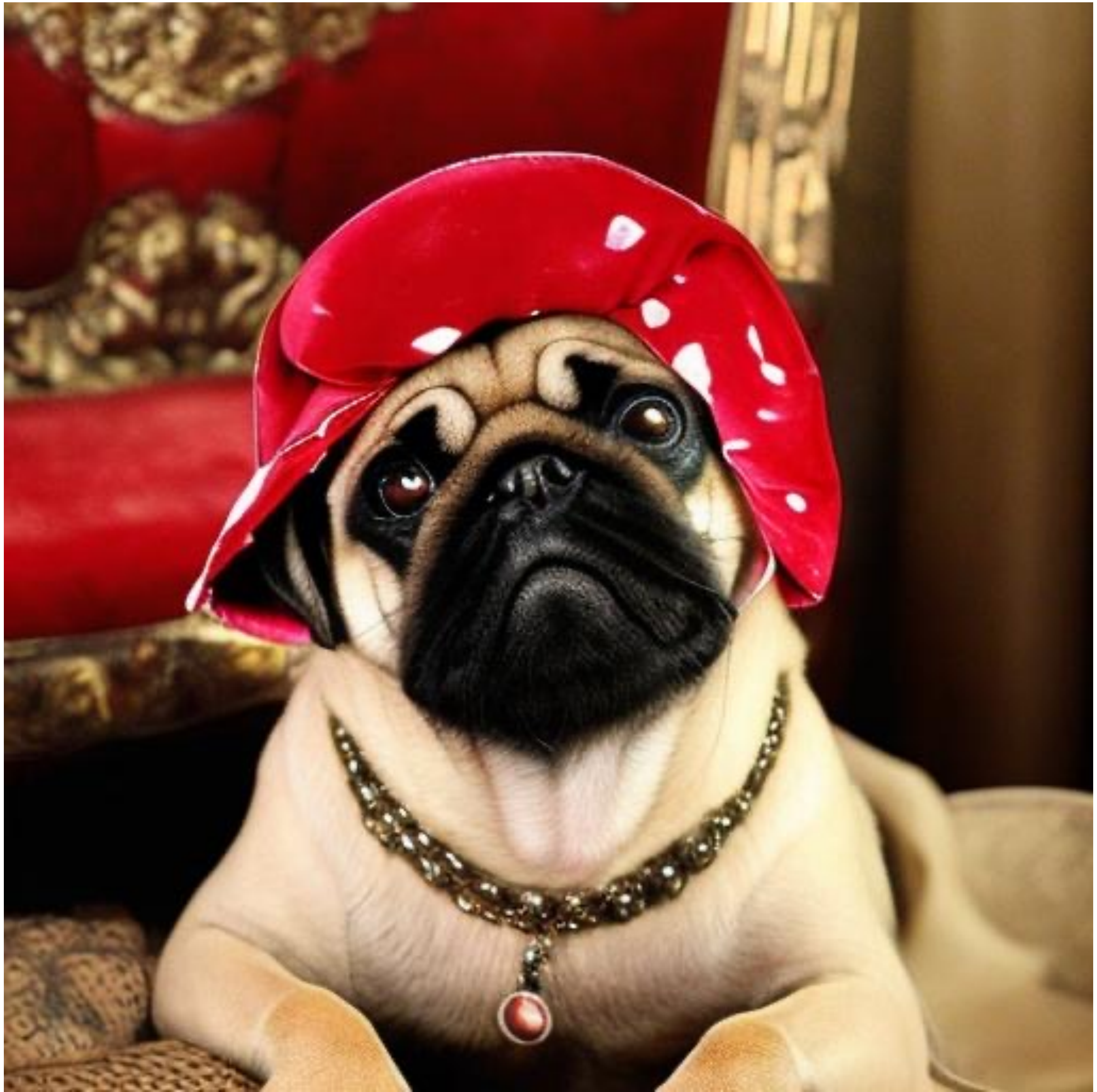
- Lookup documents in a searchable datasets
- Documents are stored as embeddings
- Use a nearest neighbor to find the nearest document
- Put that into the prompt to extract the answer and respond to the customer
- Higher accuracy assurance
- Strong performance without fine-tuning, but room for growth

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval augmented generation (RAG)



©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Stable diffusion: an adorable pug dog, with queen victoria, #puglife

Using prompt engineering to solve classification

Same as last time, we'll paste the relevant information into the prompt, and provide the model our instruction.

Stable diffusion: koalas climbing a tree

Prompt: This bear has a large, round head with fluffy, rounded ears. They have a broad, flat nose and small, round eyes. They have a unique adaptation for their diet, as their teeth are specially designed for chewing and grinding eucalyptus leaves, which make up the majority of their diet. They have sharp incisors and molars that help them strip and chew the tough, fibrous leaves.

These bears are known for their ability to climb trees. They have strong, clawed limbs with two opposable thumbs on each front paw and sharp claws on their hind limbs. These adaptations allow them to grip tree branches securely and move with agility among the eucalyptus trees, which are their primary habitat.

These bears are generally solitary animals and spend most of their time resting and sleeping. They are known for their sedentary lifestyle, spending around 18 to 20 hours a day sleeping or resting in the safety of tree branches. They are primarily nocturnal, being more active during the night, but can be seen moving slowly and leisurely during the day as well.

What type of bear is this?

Answer: Koalas.

eb Services, Inc. or its affiliates. All rights reserved.

Using prompt engineering to solve translation

You guessed it! Paste in your text, and provide an instruction to translate it.

Stable diffusion: a british flag waving in the wind, london

Prompt:
I parked my car in the driveway, took out the trash, and put my umbrella in the trunk.

Translate this into British English.

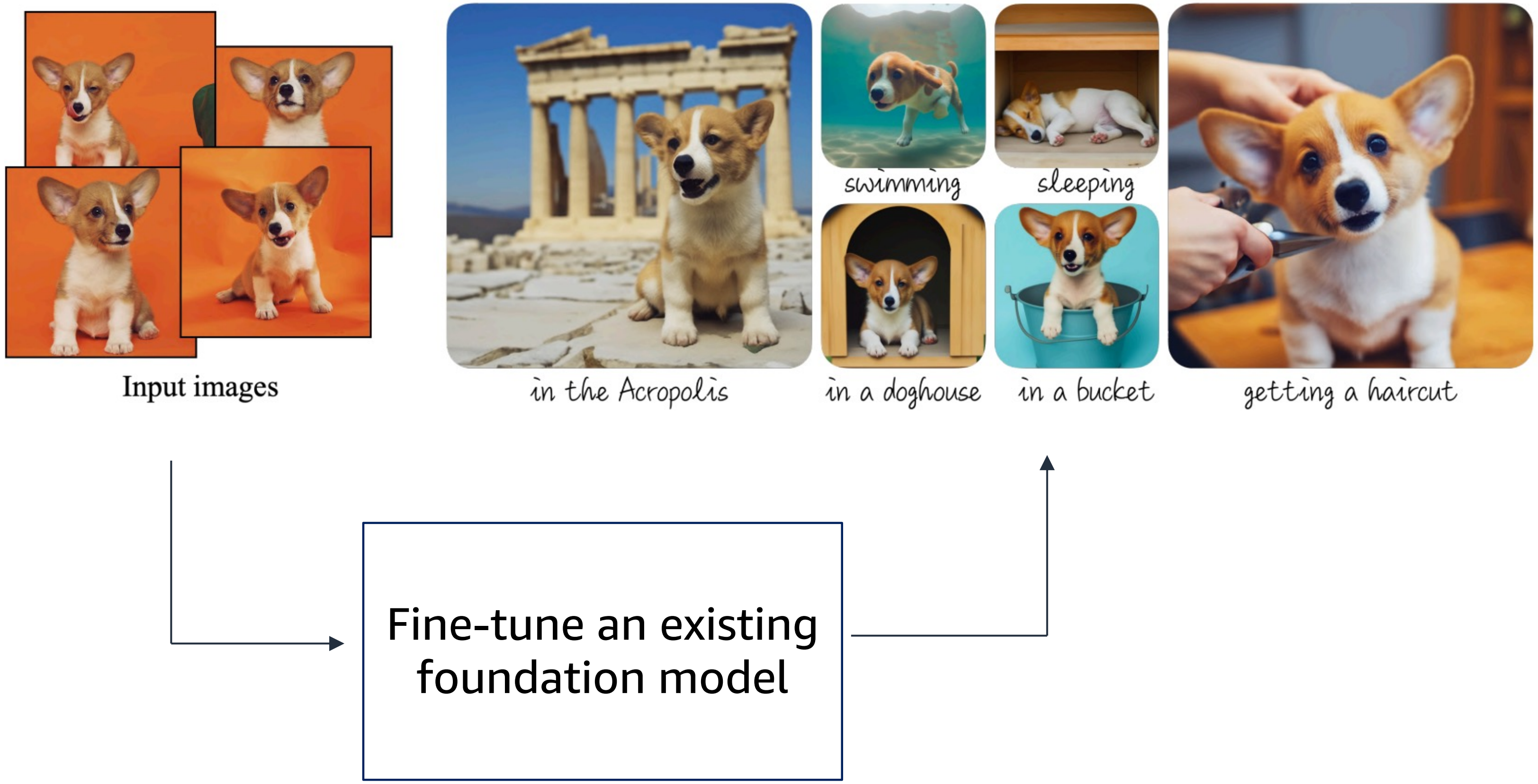
Answer: "I parked my car on the drive, took out the rubbish, and put my brolly in the boot."

Pro tip: you can also use the “translate” concept to change styles, adapt syntax, and even rewrite software in other programming languages!

eb Services, Inc. or its affiliates. All rights reserved.

When all else fails: fine-tuning

- Prompt engineering with a good foundation model base should do *something* for you.
- Only you know what *enough* looks like for your use case
- You can create a new model artifact with your own customized datasets



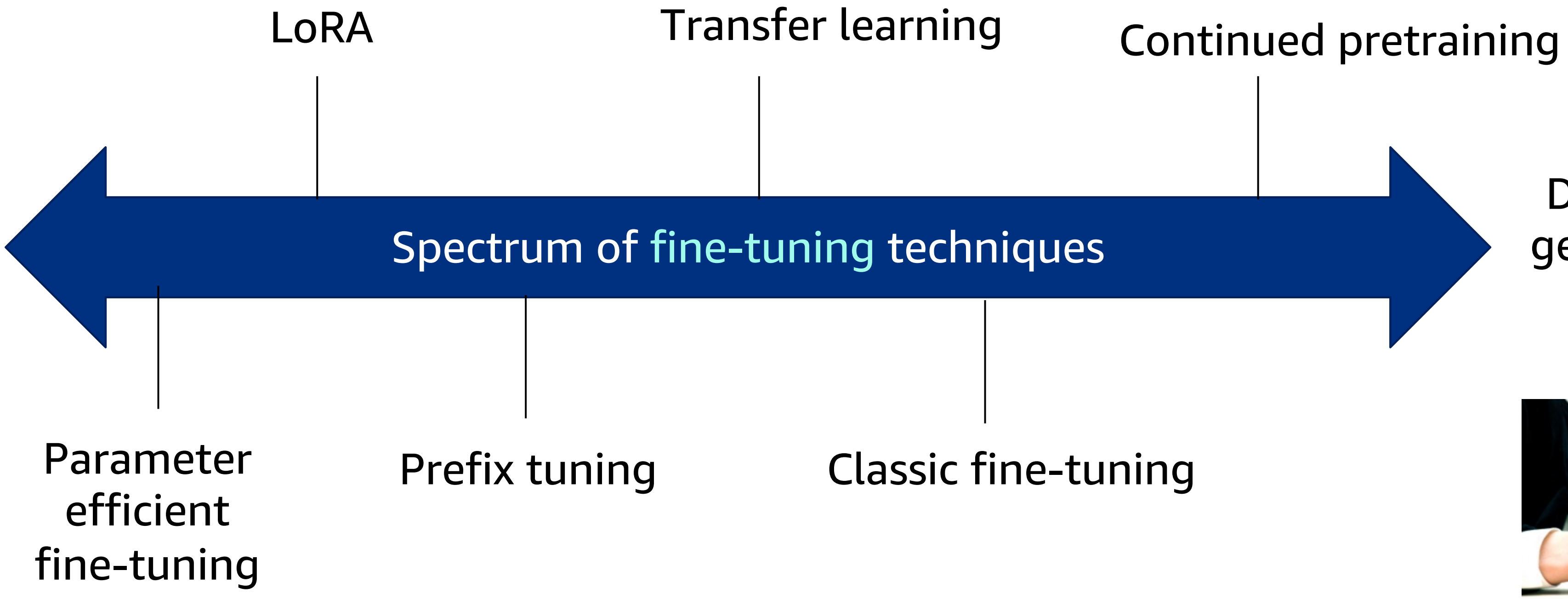
Fine-tune an existing foundation model

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation
Ruiz et. al, 2023
https://dreambooth.github.io/

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is fine-tuning?

Creating a net new model that solves every problem



Spectrum of fine-tuning techniques

LoRA

Transfer learning

Continued pretraining

Parameter efficient fine-tuning

Prefix tuning

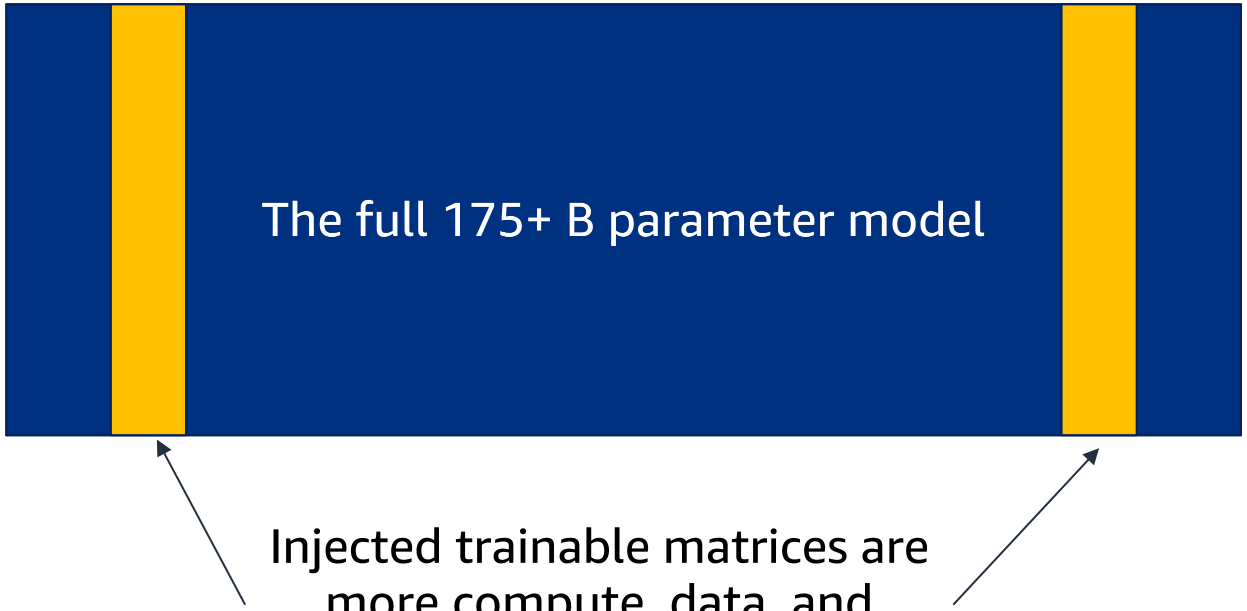
Classic fine-tuning

Doing a lot but getting nowhere

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Parameter-efficient fine-tuning

- A new library from Hugging Face to fine-tune *only a few parameters*
- Saves you from needing to train a massive model with hundreds of billions of parameters
- LoRA injects trainable weights into a frozen model. These are the ones you train.



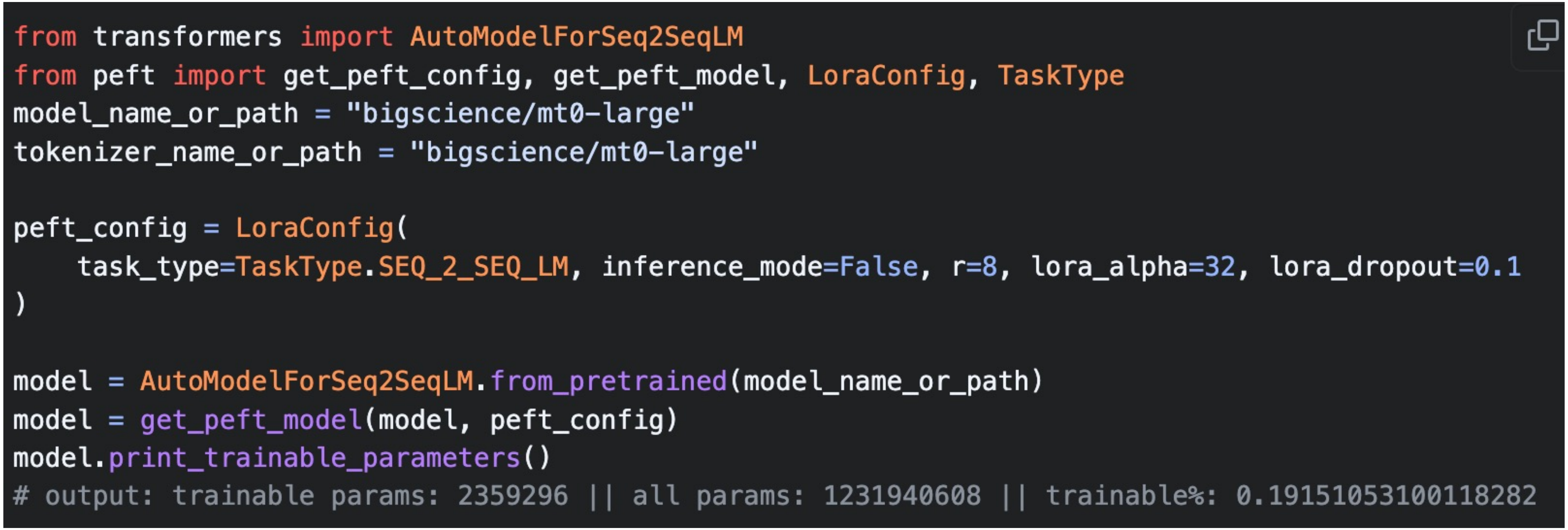
Injected trainable matrices are more compute, data, and storage efficient

🟦 Frozen weights from the original model

🟨 Blocks injected by LoRA for fine-tuning

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Parameter-efficient fine-tuning 🤗



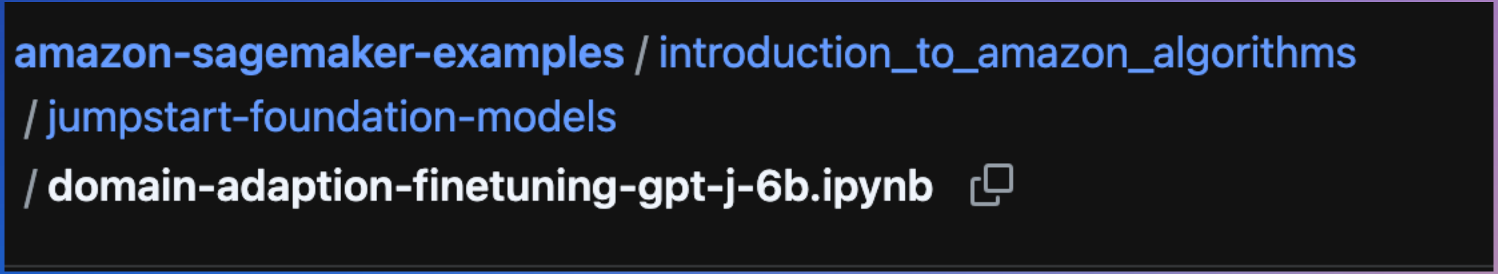
```python
from transformers import AutoModelForSeq2SeqLM
from peft import get_peft_config, get_peft_model, LoraConfig, TaskType
model_name_or_path = "bigscience/mt0-large"
tokenizer_name_or_path = "bigscience/mt0-large"

peft_config = LoraConfig(
    task_type=TaskType.SEQ_2_SEQ_LM, inference_mode=False, r=8, lora_alpha=32, lora_dropout=0.1
)

model = AutoModelForSeq2SeqLM.from_pretrained(model_name_or_path)
model = get_peft_model(model, peft_config)
model.print_trainable_parameters()
# output: trainable params: 2359296 || all params: 1231940608 || trainable%: 0.19151053100118282
```

https://github.com/huggingface/peft

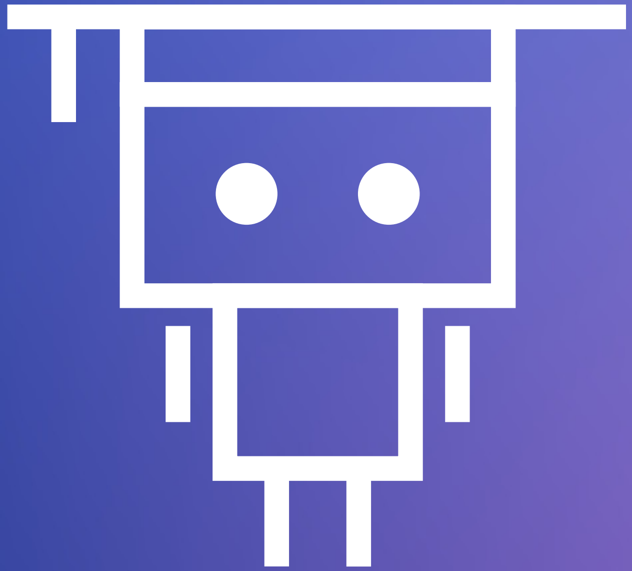
©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://bit.ly/sm-nb-3

Hands-on demo


amazon-sagemaker-examples / introduction_to_amazon_algorithms / jumpstart-foundation-models / **domain-adaption-finetuning-gpt-j-6b.ipynb**

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Thank you!

Type: Corrections, feedback, or other questions?
Contact us at https://support.awsamazon.com/#/contacts/aws-academy.
All trademarks are the property of their owners.

©2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.