

Milano 17 GIUGNO 2024

Generative AI and Foundation Models, what's in it for us, developers?

François Bouteruche Senior Developer Advocate @ Amazon Web Services



improove



## >>AI CONF

17 GIUGNO 2024 4a EDIZIONE

adesso.it





















#### Introduction to generative Al

#### Image



Generated with Amazon Titan Image Generator G1

"A happy golden retriever wearing glasses and a hat in a portrait painting"

#### Text/Chat

> Question: What is generative AI?

Generative AI refers to a type of artificial intelligence that is capable of creating new and original content. It uses algorithms and machine learning to generate text, images, music, and other forms of media based on patterns and rules learned from existing data.

Generated with Amazon Titan Text G1 - Express



## What are Language Models?



#### Towards Large Language Models

Next word prediction with simple word embedding



? ? ? ? ? ; ; ; ; ; ; t ? lazy \_ person.



#### Towards Large Language Models

Next word prediction with 'memory' such as recurrent neural networks (RNN)

#### Sentence

?

?

fox jumps over the lazy \_

person.

rabbit.



# Transformers – Attention Is All You Need (2017)

#### **Attention Is All You Need**

#### "Attention is all you need"

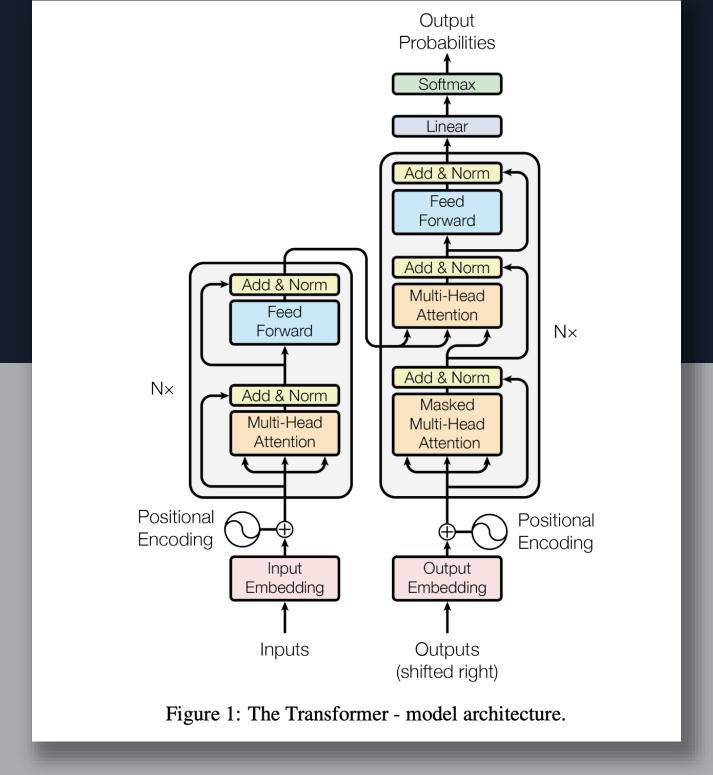
Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017)



# Transformers – Attention Is All You Need (2017)

#### The Transformer model architecture

- "Attention" enabled models to scale the understanding of relationships between words
- Efficiently use parallel computing



Ashish Vashwani et al., Attention Is All You Need.



#### Towards Large Language Models

Prediction using a transformer based network

Sentence

The quick brown fox jumps over the lazy \_



#### Towards Large Language Models

Prediction using a transformer based network

Sentence

The quick brown fox jumps over the lazy dog.

...

It appears to be referring to the classic English pangram "The quick brown fox jumps over the lazy dog".



# How does text generation work?

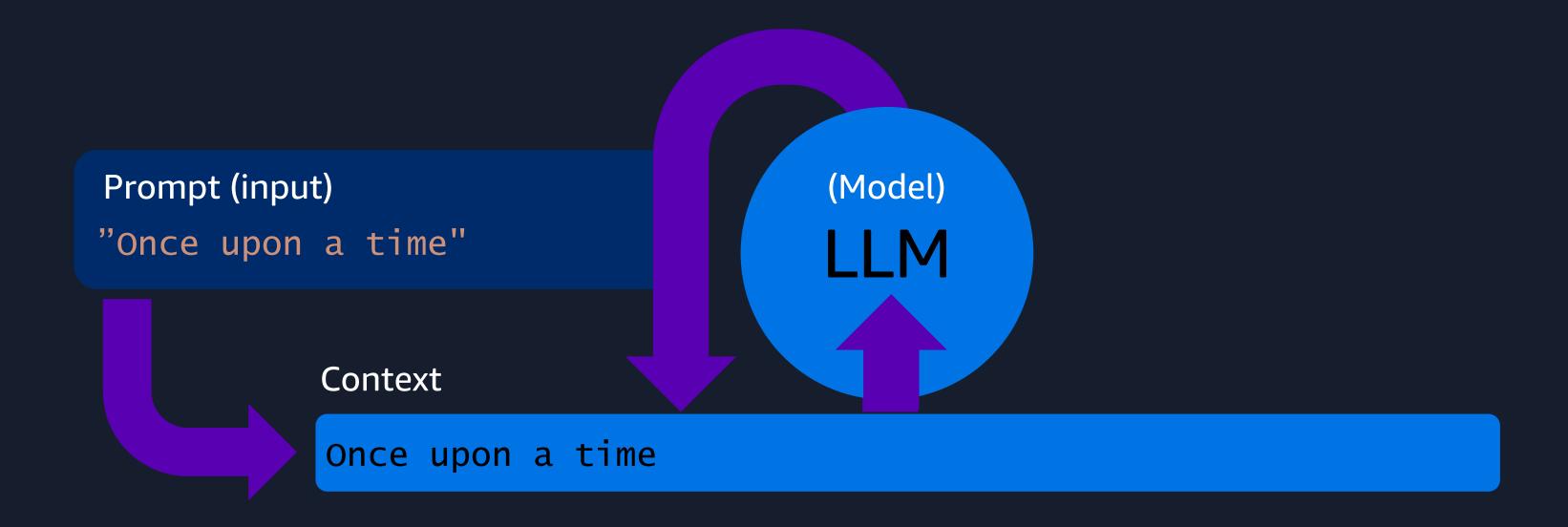


Prompt (input)

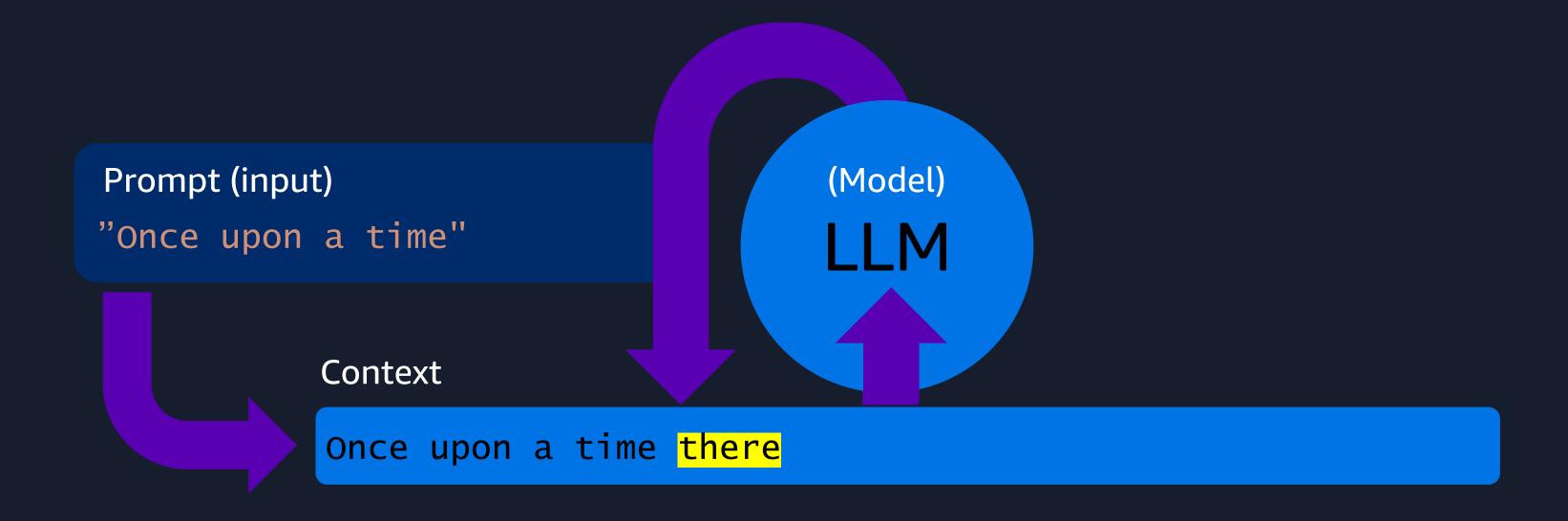
"Once upon a time"



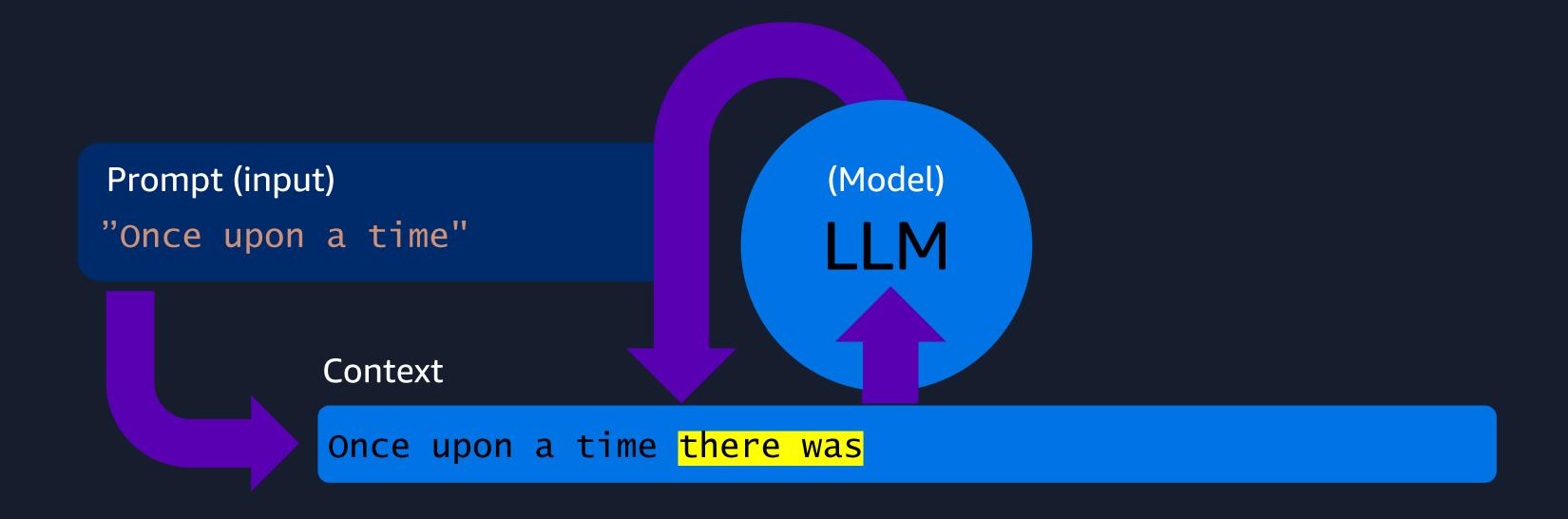




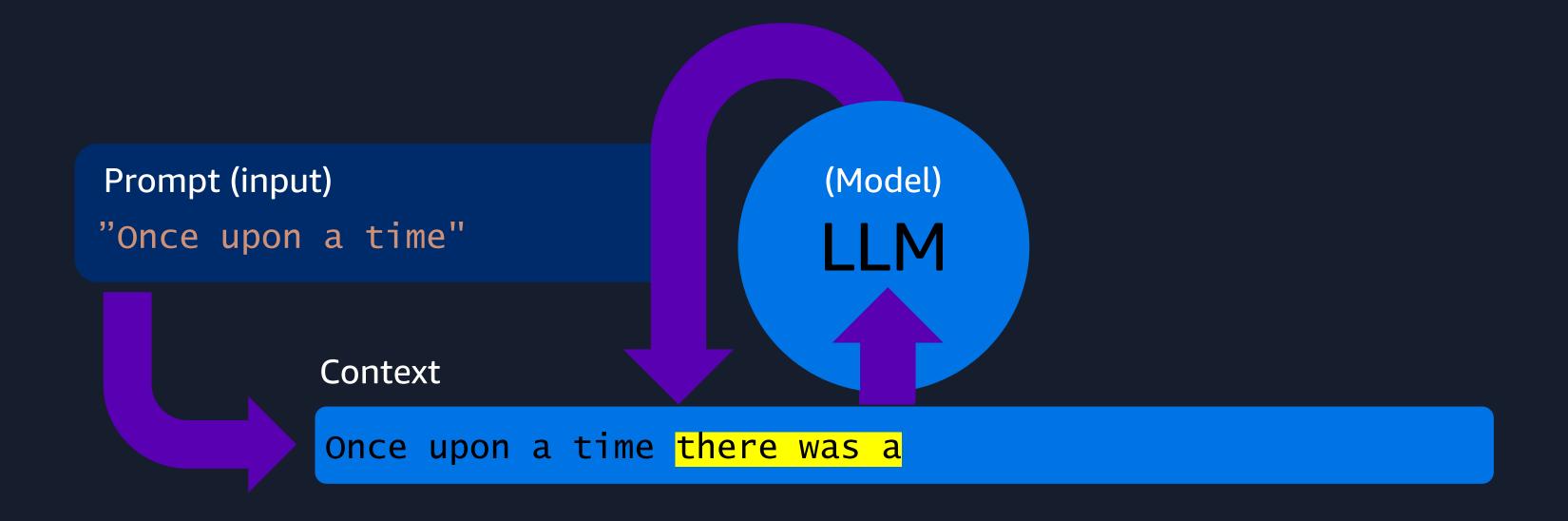




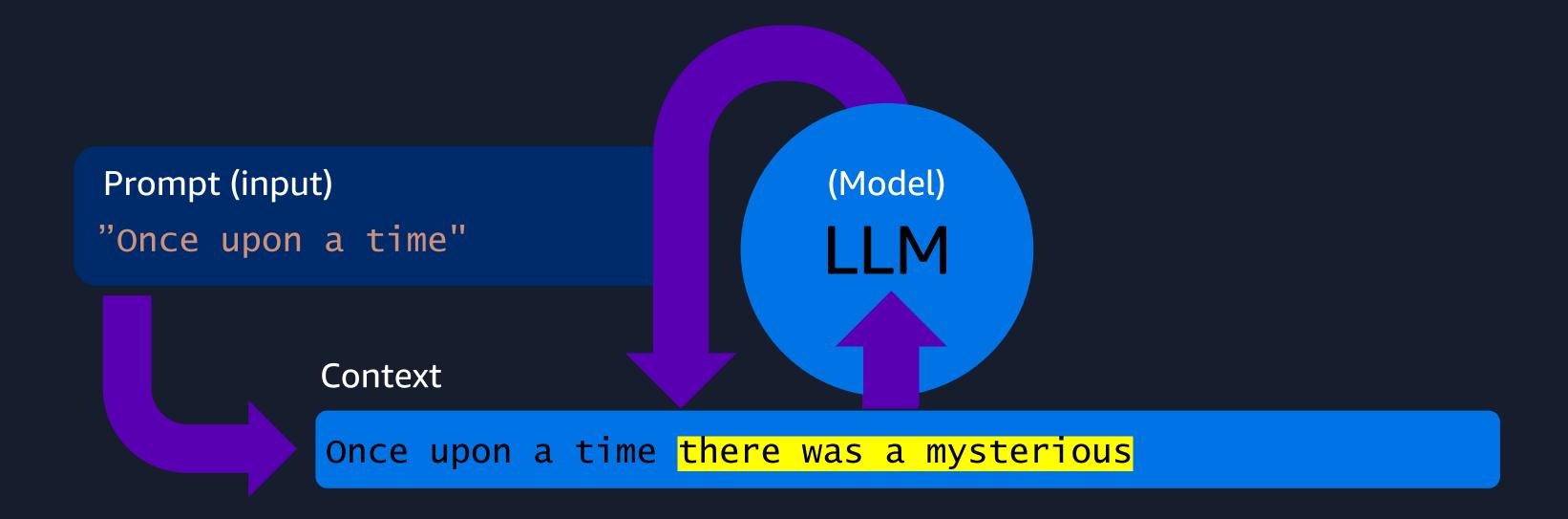




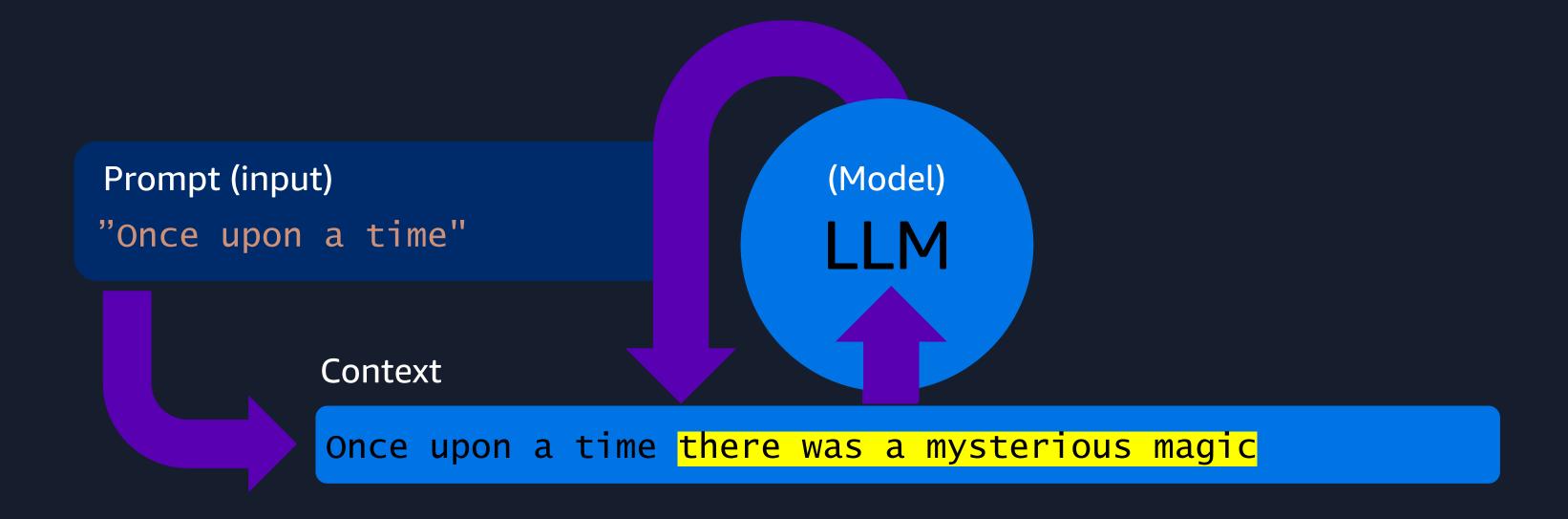




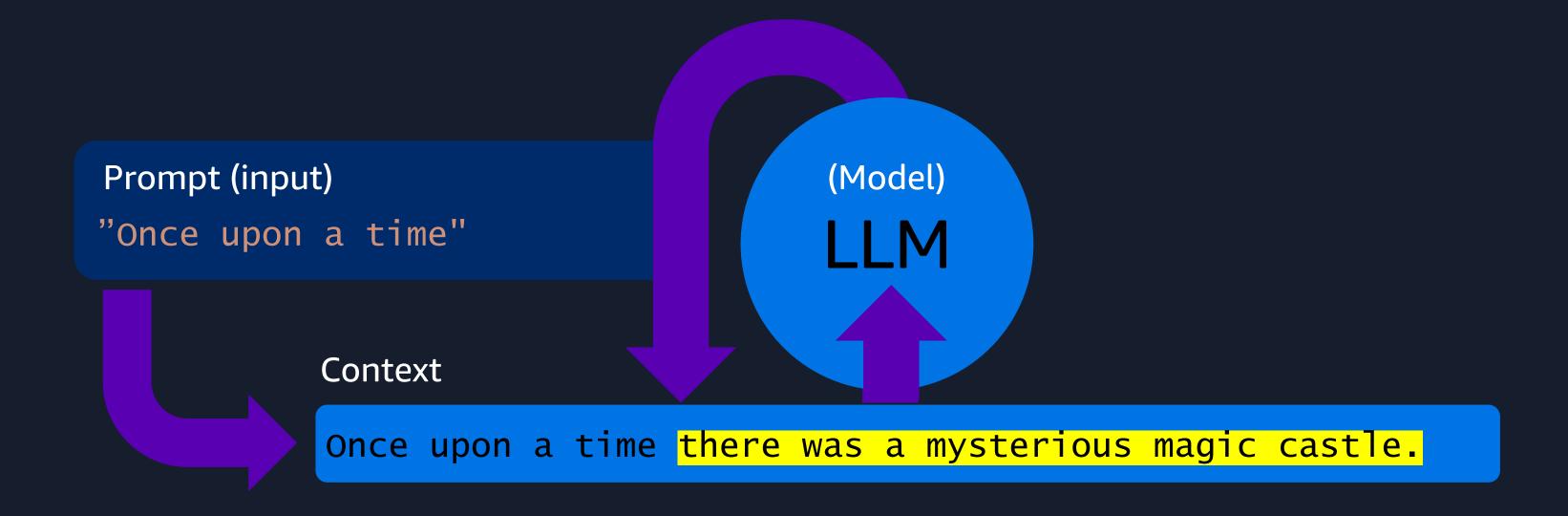




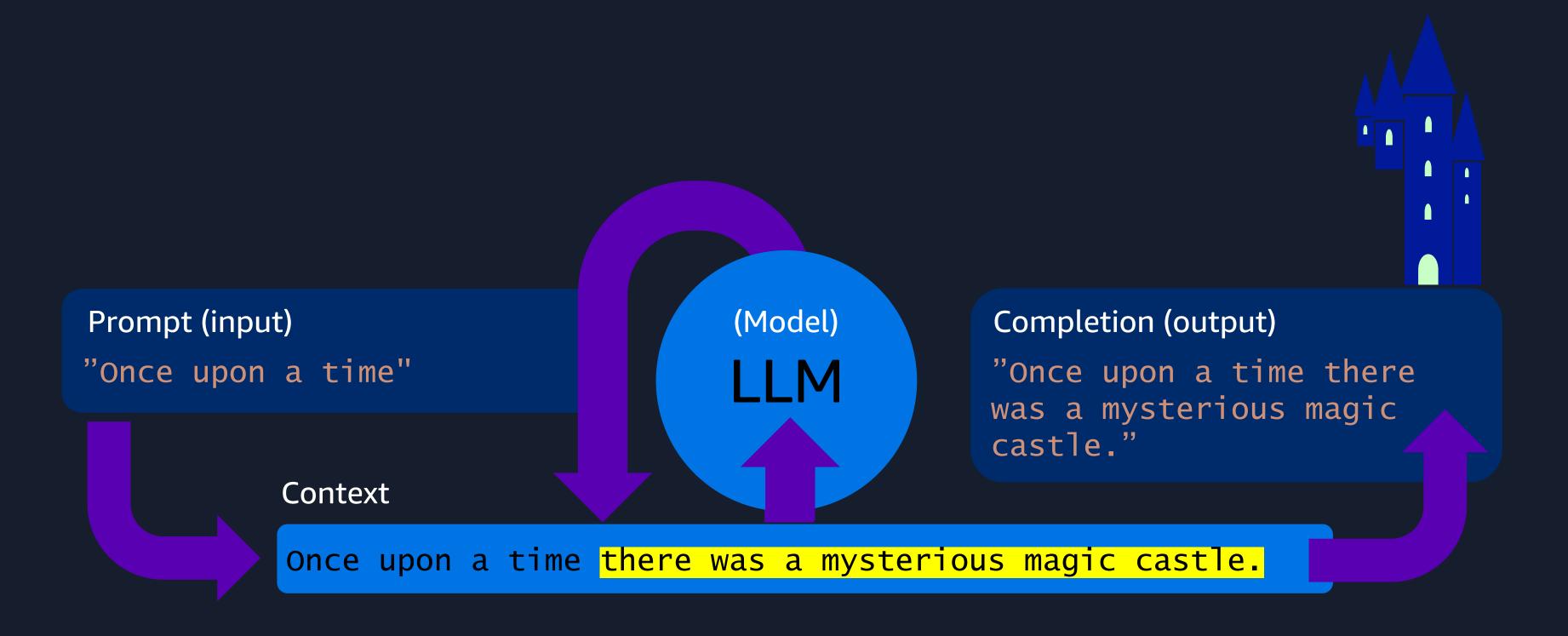








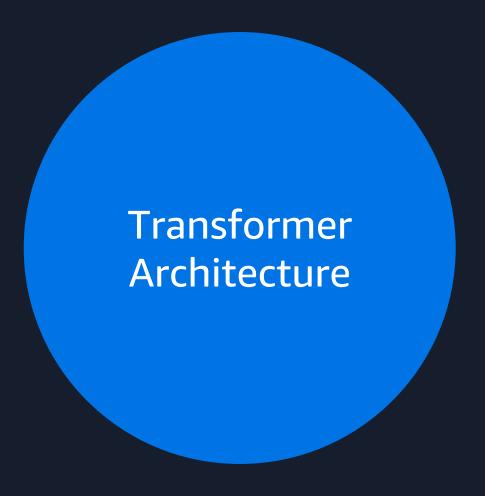




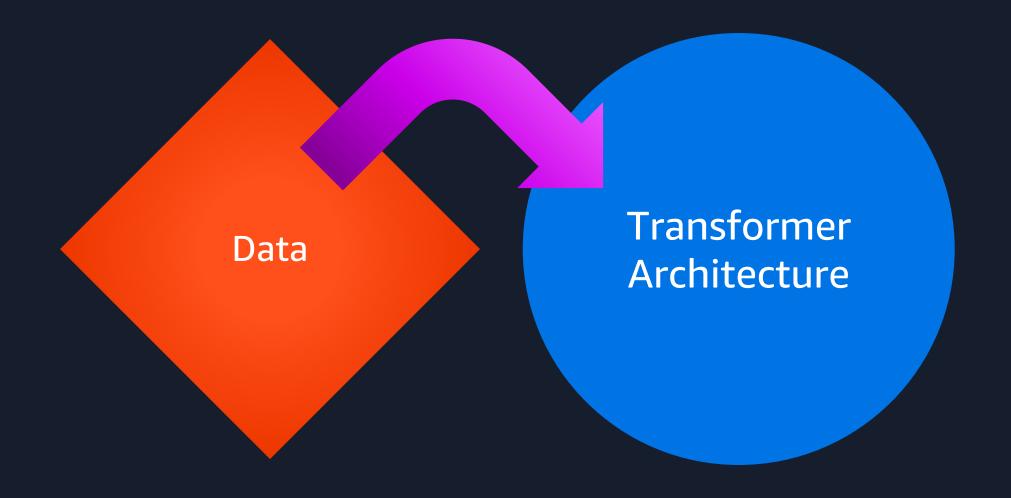


### What are Foundation Models?

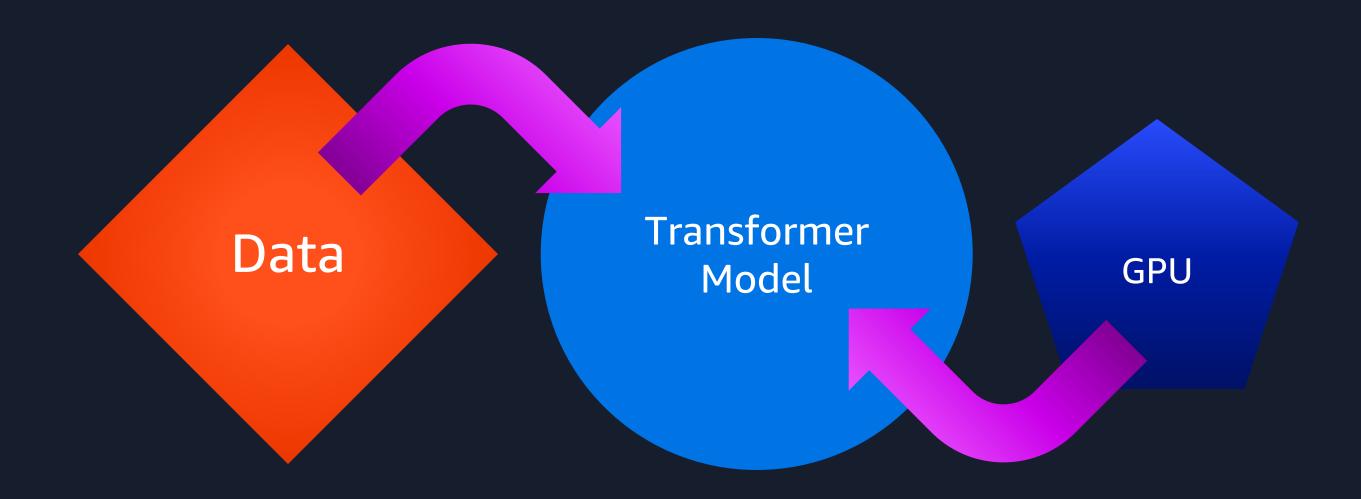




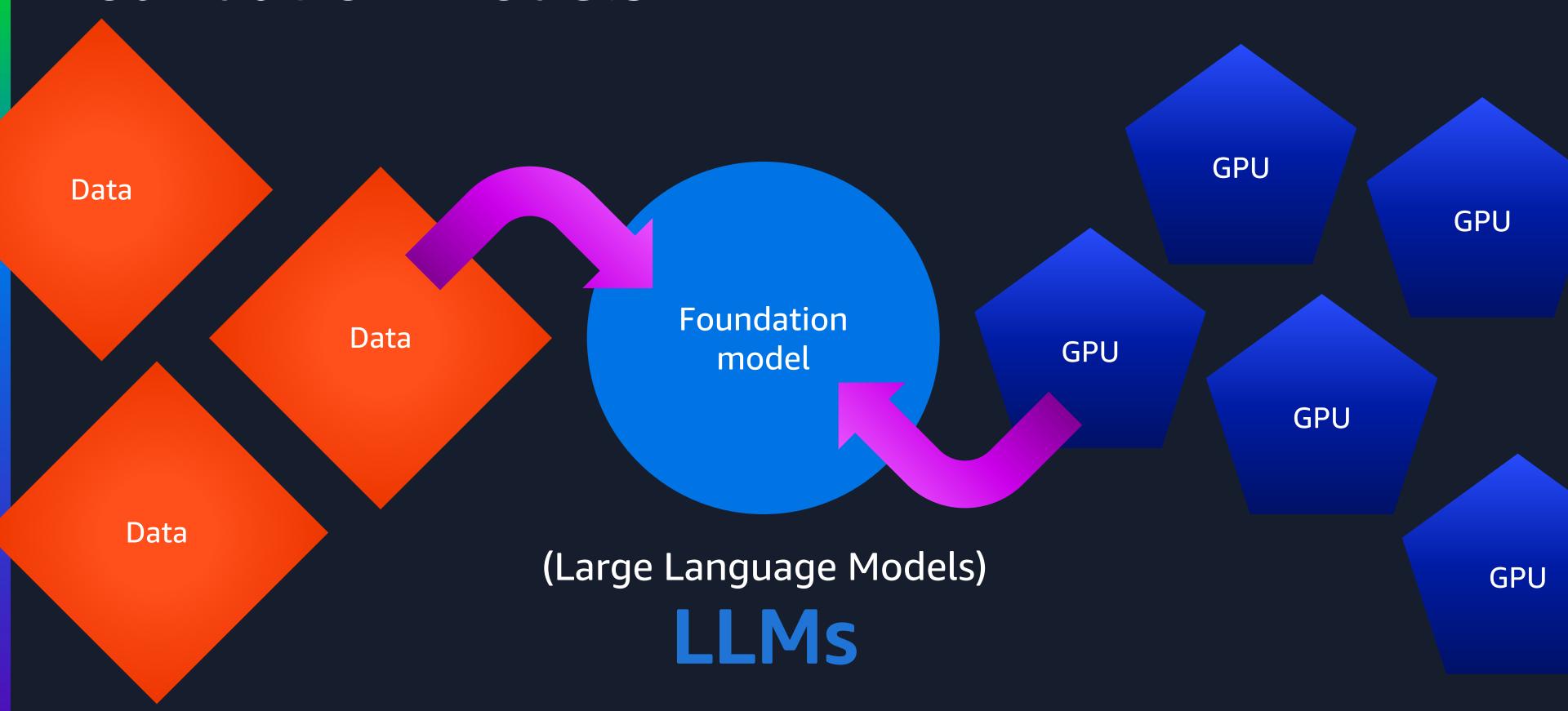












Al21 labs

amazon

ANTHROP\C

**S**cohere

**Meta** 

MISTRAL AI\_

stability.ai

Contextual answers, summarization, paraphrasing

Jurassic-2 Ultra
Jurassic-2 Mid

Text summarization, generation, Q&A, search, image generation

Amazon Titan Text Premier

Amazon Titan Text Lite

Amazon Titan Text Express

Amazon Titan Text Embeddings

Amazon Titan Text Embeddings V2

> Amazon Titan Multimodal Embeddings

Amazon Titan Image Generator Summarization, complex reasoning, writing, coding

Claude 3 Opus

Claude 3 Sonnet Claude 3 Haiku

Claude 2.1

Claude 2

Claude Instant

Text generation, search, classification

Command

Command Light

Embed English

Embed Multilingual

Command R+

Command R

Q&A and reading comprehension

Llama 3 8B

Llama 3 70B

Llama 2 13B

Llama 2 70B

Text summarization, text classification, text completion, code generation, Q&A

Mistral Large

Mistral 7B

Mixtral 8x7B

High-quality images and art

Stable Diffusion XL1.0

Stable Diffusion XL 0.8



## Let's discuss about real things



## Wrap up



### Key takeaways

WE pretend Foundation Models to be smart as we do

Foundation Models have no memory, they are stateless

Foundation Models may be optimized for a specific prompt format

Foundations Models share common parameters and have their own

YOU can build smart applications by mixing different AI services and FMs

YOU can build smart applications by creating AI Agents that are specialized



## Feedback is a gift 💛





## Thank you!

François Bouteruche

linkedin.com/in/francoisbouteruche





#### improove

>>AICONF

4a EDIZIONE