



# Unveiling the Power of GenAI: Extracting Insights from Chats

Witold Chmielowiec

Małgorzata Wachulec

Marta Janas

# Meet Our Team

Data Science IT Team at Equinix



**Małgorzata Wachulec**



**Marta Janas**



**Witold Chmielowiec**

# Problem statement

Hey there!  
How can I help?

I would like to add  
collaborators to my  
project. How can I do it?



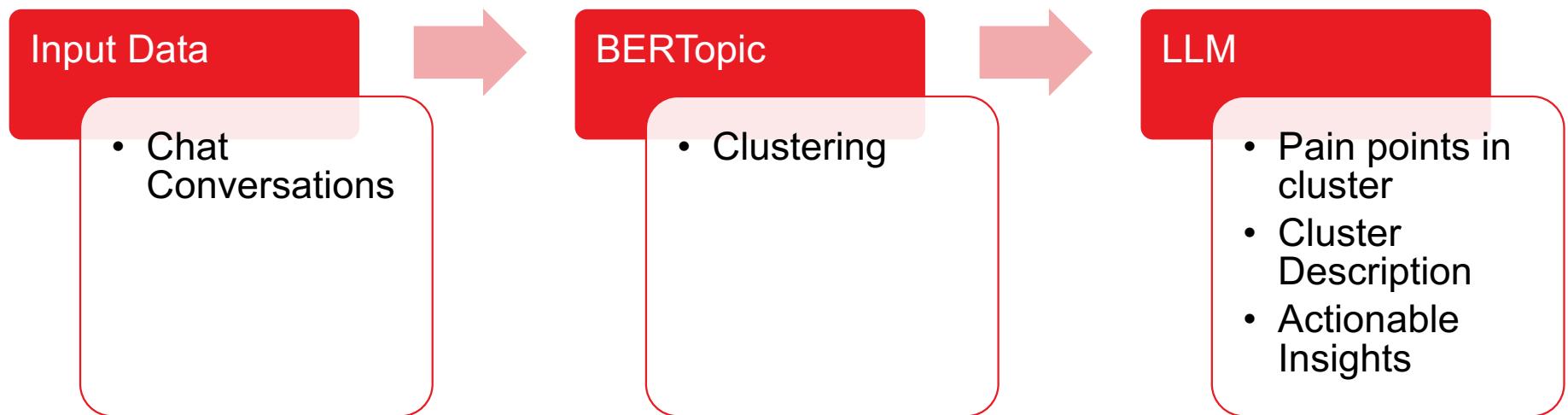


## Why?

- **Analyzing root causes of the pain points**
- **Opportunity to improve systems / processes to address the issues**
- **Identifying unmet needs**
- **Reducing number of issues that customers encounter**
- **Reducing time required to solve the tickets and respond to the chats**
- **Improving customer satisfaction**



# How?



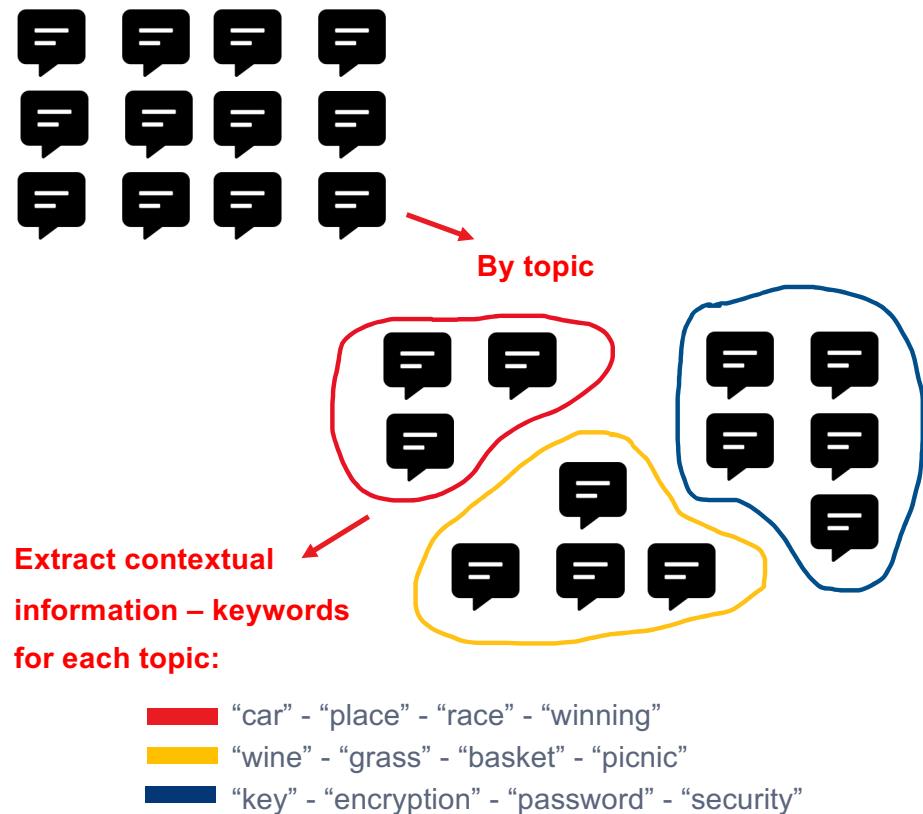
# Preparing data & EDA

# BERTopic

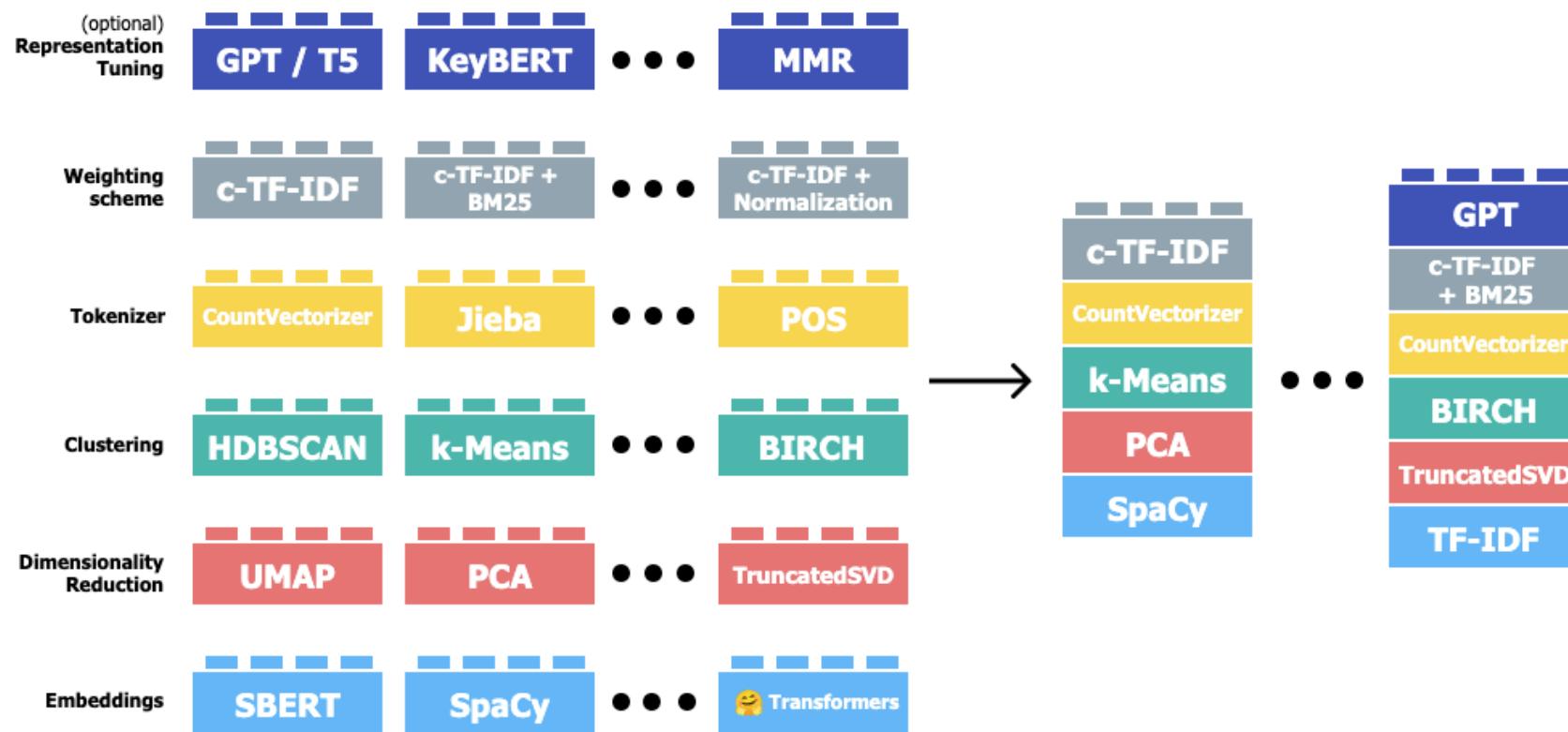
# What is topic modelling?

Making sense of a bunch of documents

Grouping documents

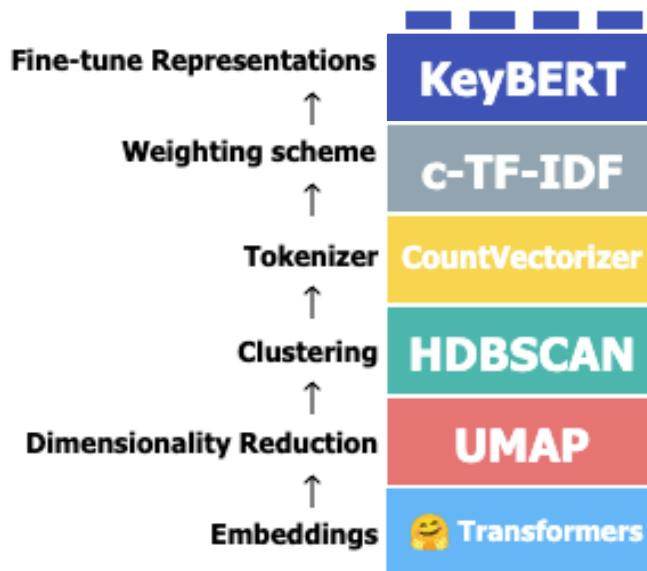


# Building the topic modelling solution with BERTopic



Source: <https://maartengr.github.io/BERTopic/algorithm/algorithm.html>

# BERTopic steps explained



- Step 6 - (Optional) Fine-tune topic representations
- Step 5 - Weight tokens, create topic representation
- Step 4 - Tokenization of topics
- Step 3 - Clustering reduced embeddings into topics
- Step 2 - Reducing dimensionality of embeddings
- Step 1 - Embedding documents

Source: <https://maartengr.github.io/BERTopic/algorithm/algorithm.html>

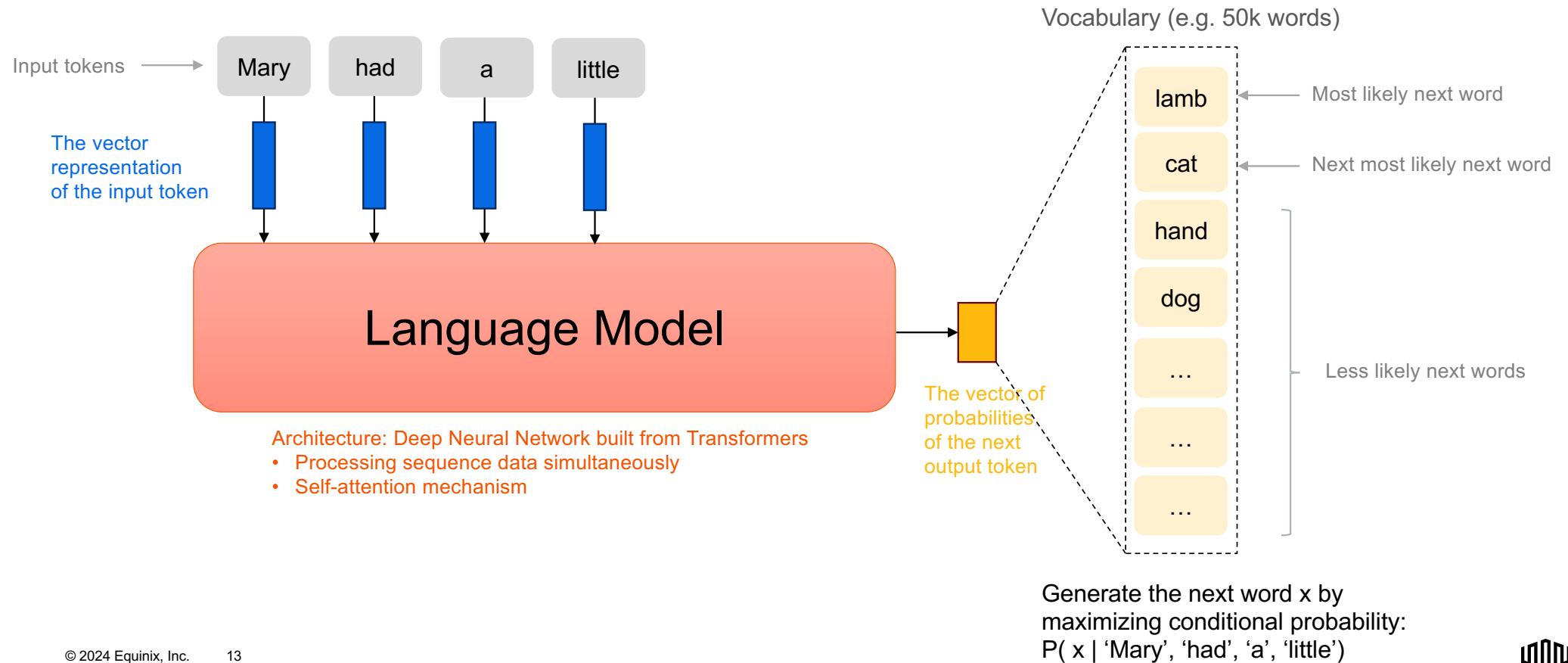
# Large Language Models

# **What is a Large Language Model (LLM)?**

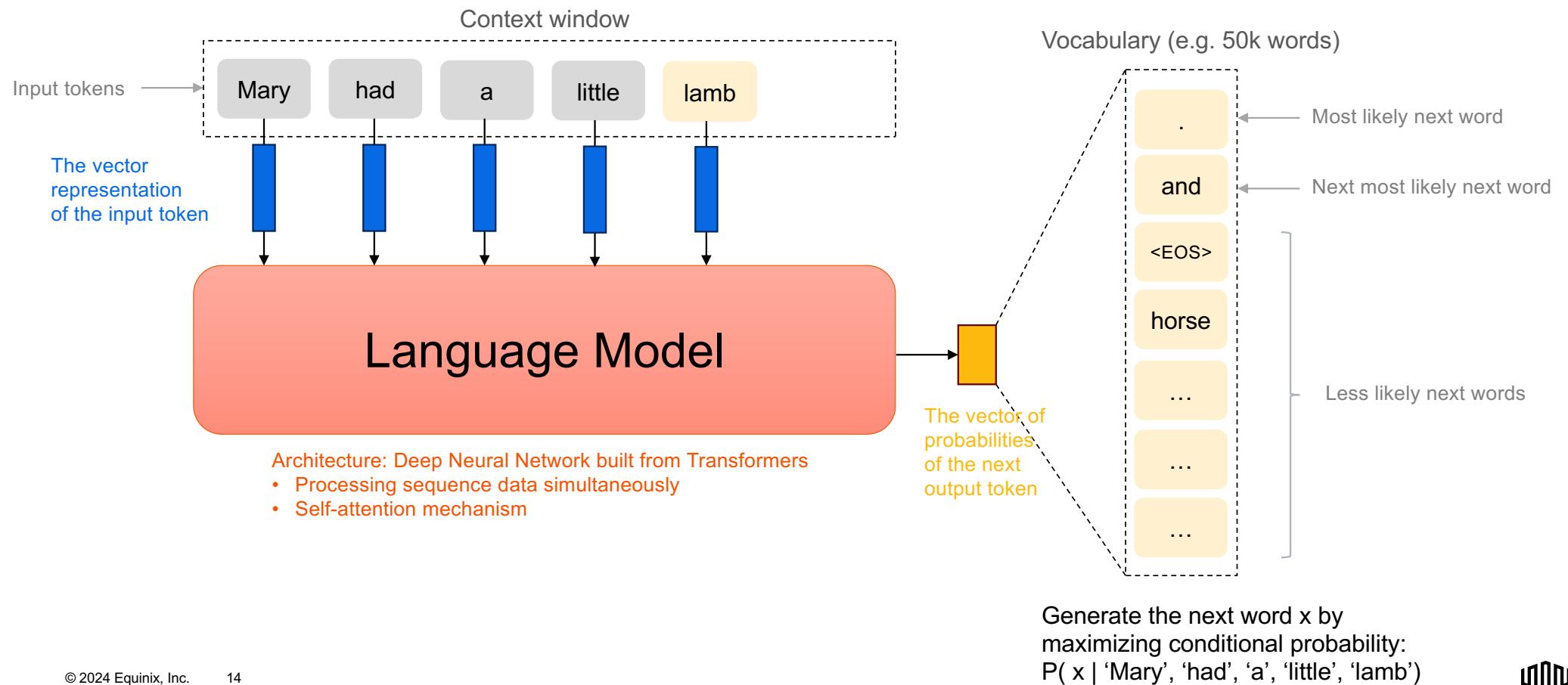
What is a  
Language Model?

Why is the model  
'Large'?

# What is a Language Model?



# What is a Language Model?



# Why is the model ‘Large’?

## 1. Large number of model parameters

- Learnable elements of the model that allow it to understand and generate natural language
- Hundreds of gigabytes (GB) and more of storage space for the model (each float32 parameter needs 4 bytes)

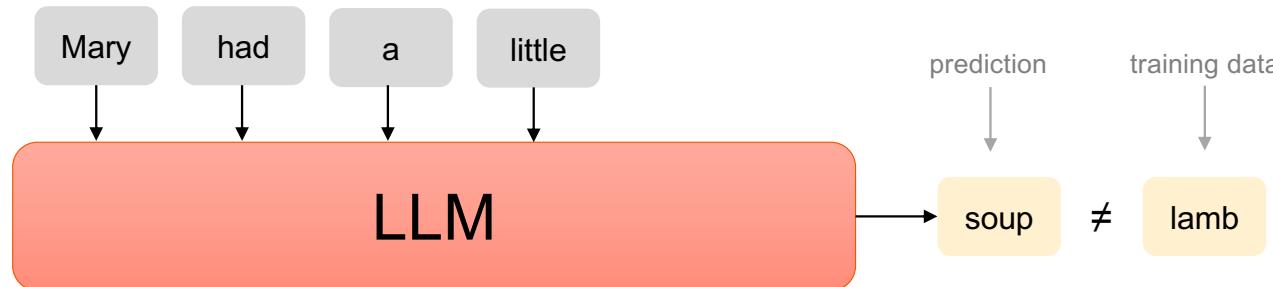
## 2. Large training dataset

- Training on massive corpora of text: web pages, books, news articles, Wikipedia, social media
- Range from tens of gigabytes (GB) to petabytes (PB)

Name	Developer	Release	Number of parameters	Corpus size
GPT-3	OpenAI	May 2020	175 billion (~700 GB)	300 billion tokens
PaLM 2	Google	May 2023	540 billion (~2.1 TB)	3.6 trillion tokens
Llama 3	Meta AI	April 2024	70 billion (~280 GB)	15 trillion tokens

# LLM training

- **Training process (self-supervised learning)**



1. Model predicts the next word in a sentence based on the context provided by the preceding words (*forward propagation*).
  2. Model adjust internal parameters to reduce the prediction error (*backward propagation + gradient decent*).
- **Model learns grammar, semantics and statistical patterns between words and phrases.**
  - **It also acquires general knowledge from training examples.**

# Prompt & Response (Gemini 1.0 Pro)

## Example 1

**Prompt:** How many moons does Jupiter have?

**Response:** Jupiter has 79 known moons, making it the planet with the most moons in our solar system. This number includes both large, spherical moons like Ganymede and Callisto, as well as smaller, irregularly shaped moons.

## Example 2

**Prompt:** What is the weather like in Warsaw today?

**Response:** I do not have real-time capabilities and my knowledge cutoff is April 2023. Therefore, I cannot provide you with the current weather conditions in Warsaw. For the most up-to-date and accurate weather information, I recommend checking a reputable weather app or website, such as the Weather Channel or AccuWeather.

The screenshot shows the Gemini 1.0 Pro interface within the Google Cloud Platform. The top navigation bar includes 'Google Cloud', a dropdown for 'helix-ds-metal-dev', a search bar, and various icons. The main workspace is titled 'Untitled prompt' and contains two sections: 'Prompt' and 'Response'. The 'Prompt' section contains the question 'What is the weather like in Warsaw today?'. The 'Response' section provides a disclaimer about real-time capabilities and recommends checking other sources for the most up-to-date weather information. On the right side of the interface, there are several configuration panels: 'Model' set to 'gemini-1.0-pro-001', 'Region' set to 'us-central1 (Iowa)', 'Temperature' slider at 0.9, 'Output token limit' set to 8192, and a 'Settings' panel with options for 'Functions', 'Temperature', 'Output token limit', and 'Safety Settings'. A 'RESET PARAMETERS' button is also visible.



# Summary

# Key takeaways from using LLM

- **LLM is a great tool for extracting information from text**
  - Generates labels, issues, insights, etc.
- **LLMs are rapidly and constantly evolving, enabling more tasks and better results**
  - Stay up-to-date in this field
- **Remember about the context window size**
  - Prepare appropriate prompt size
- **Each time LLM generates a slightly different output for the same input**
  - Not reproducible results
- **Not easy to obtain a parsable output (in JSON)**
  - Prompt engineering

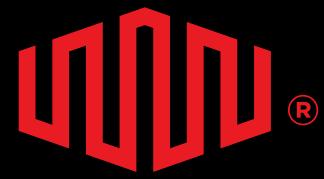
# Summary & alternative approaches

## Summary

- We used clustering method (BERTopic) to group similar chats, and then applied LLM to extract relevant information from each group
- This is just an example how to approach to such analysis, but everything depends on:
  - your business case and requirements
  - number of chats to analysis and LLM context window size
  - one-time analysis vs. analysis of incremental data (e.g. once in a month)

## Alternative approaches

- Using only LLM to group and extract issues and insights
- First extracting issues using LLM than clustering them, and finally using LLM again to extract relevant info from the issues
- And many others...



EQUINIX

