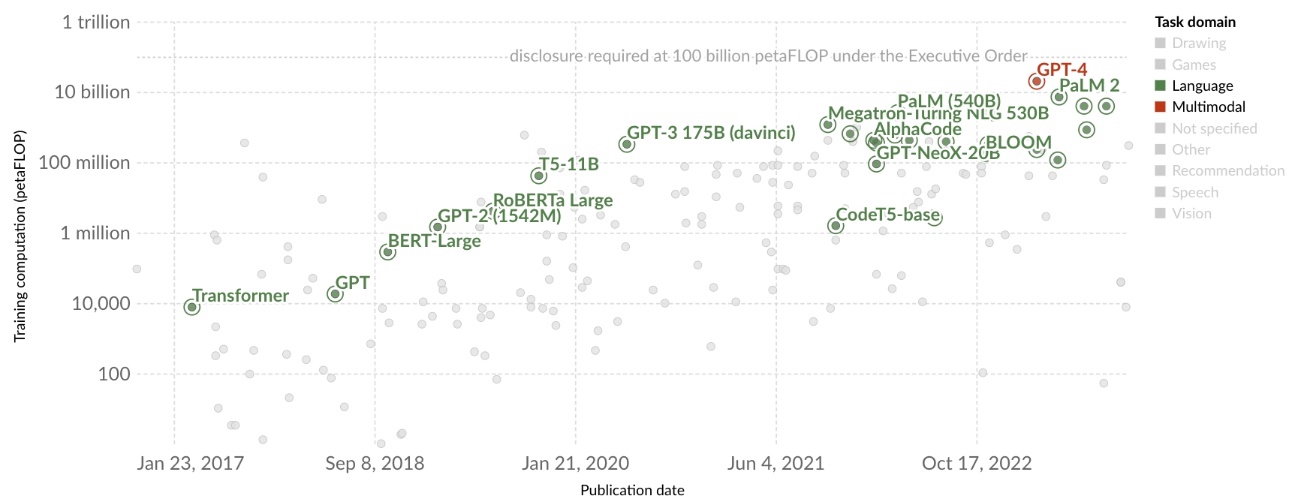
**LLM’s and Text Generation**

# Intro to LLMs



## Encoder vs Decoder Models

The first LLM to gain broad adoption was [BERT(opens in a new tab)](https://arxiv.org/abs/1810.04805) (Bidirectional Encoder Representations from Transformers), an encoder-only model. Encoder-only models are most commonly used as base models for subsequent fine-tuning with a distinct objective, e.g. for the inference-time task of binary classification of movie reviews.

However, before BERT was released, the first [GPT(opens in a new tab)](https://openai.com/research/language-unsupervised) (Generative Pre-Trained Transformer) model, a decoder-only model, was released by OpenAI. Decoder-only models are most commonly used for the inference-time task of text generation. In distinction to encoder-only models, the Transformer's pre-training objective of next token prediction is very similar to the decoder-only model's inference-time task of text generation.

## Decoding Parameters

Rather than the single next token, an LLM’s output is actually a probability distribution across tokens. In order to choose the single next token to generate, a decoding mechanism must be specified. Many LLM inference APIs expose the same decoding parameters:

* temperature
* top\_p or top\_k
* \*\_penalty (often repetition\_ penalty, or frequency\_penalty and presence\_penalty)

Many of the decoding parameters serve to augment the LLM's next token probability distribution. For example, increasing the temperature flattens the probability distribution, making it far more likely to sample a token that is not at the very top of the distribution and resulting in more "creative" LLM responses.

However, to ensure a more reproducible and deterministic response, you likely want to set temperature=0 and possibly top\_p=1 (depending on the inference implementation) to request greedy decoding.

## What is a propmt

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Together AI Account

<https://api.together.xyz/settings/api-keys>

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Anatomy of a prompt

Escala de tiempo

Descripción generada automáticamente

\*Importante el trigger.

## OPEN AI API Key

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

# NLP Fundamentals

## Natural Languages vs. Structured Languages

A natural language is a language that evolved naturally through human communication, such as Spanish, Mandarin, or American Sign Language.

A structured language is an invented or constructed language, such as a computer programming language.

NLP reveals structure and meaning from human language to computers and its importance has grown in the modern age.

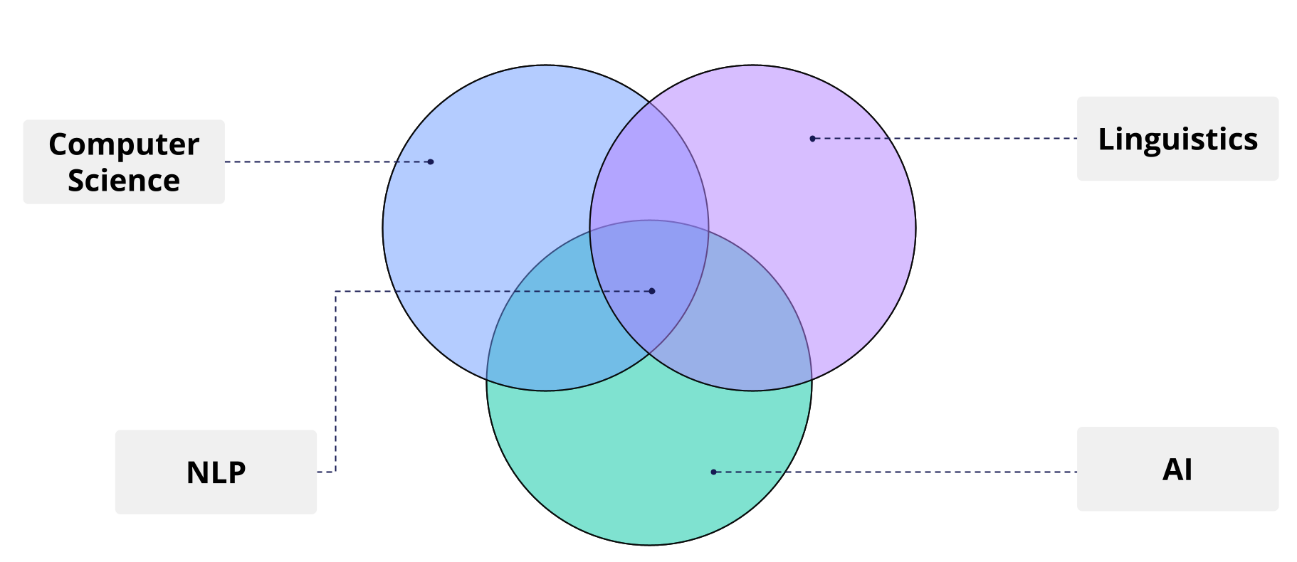
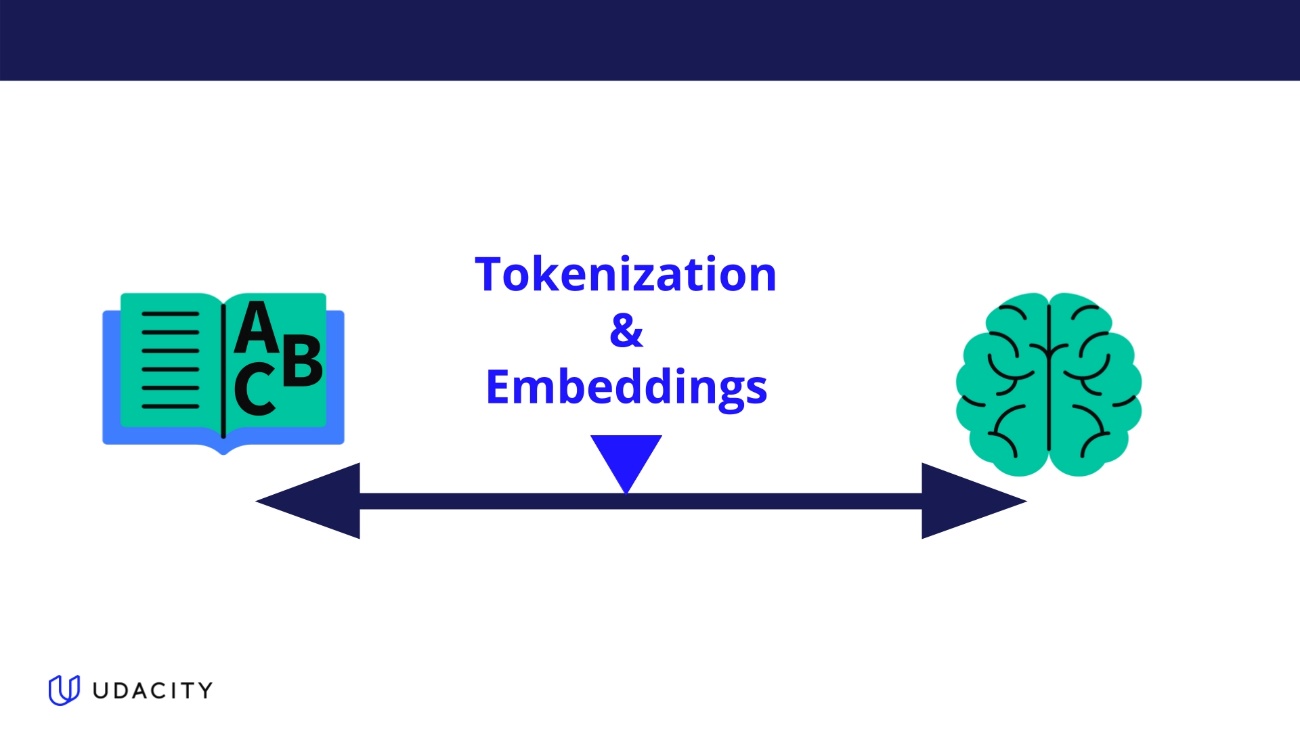


Imagen que contiene Tabla

Descripción generada automáticamente

Computers don't understand raw language and need to translate text into computer-usable format



Tokenization breaks text into chunks called "tokens" and encodes these as numerical representations

Embeddings encode context into a vector representation.

## Tokenization

* Normalization cleans text for consistency by removing complexity.
* Pretokenization breaks the text into smaller "words" and will be the base of what tokens will be.
* Tokenization breaks text into smaller parts called "tokens" to create meaningful building blocks.
* Postprocessing applies additional transformations, such as adding tags at the beginning and end of sentences.

Existen varios métodos de tokenización:

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

## Embeddings

Classic methods for vectorization (such as bag-of-words, one-hot encoding, and TF-IDF) can lack contextual relationships.

Embeddings can encode context by vectorizing text/tokens into representational vectors.

Gráfico

Descripción generada automáticamente con confianza mediaDiagrama

Descripción generada automáticamente con confianza baja

Imagen que contiene Diagrama

Descripción generada automáticamente

## Sequences

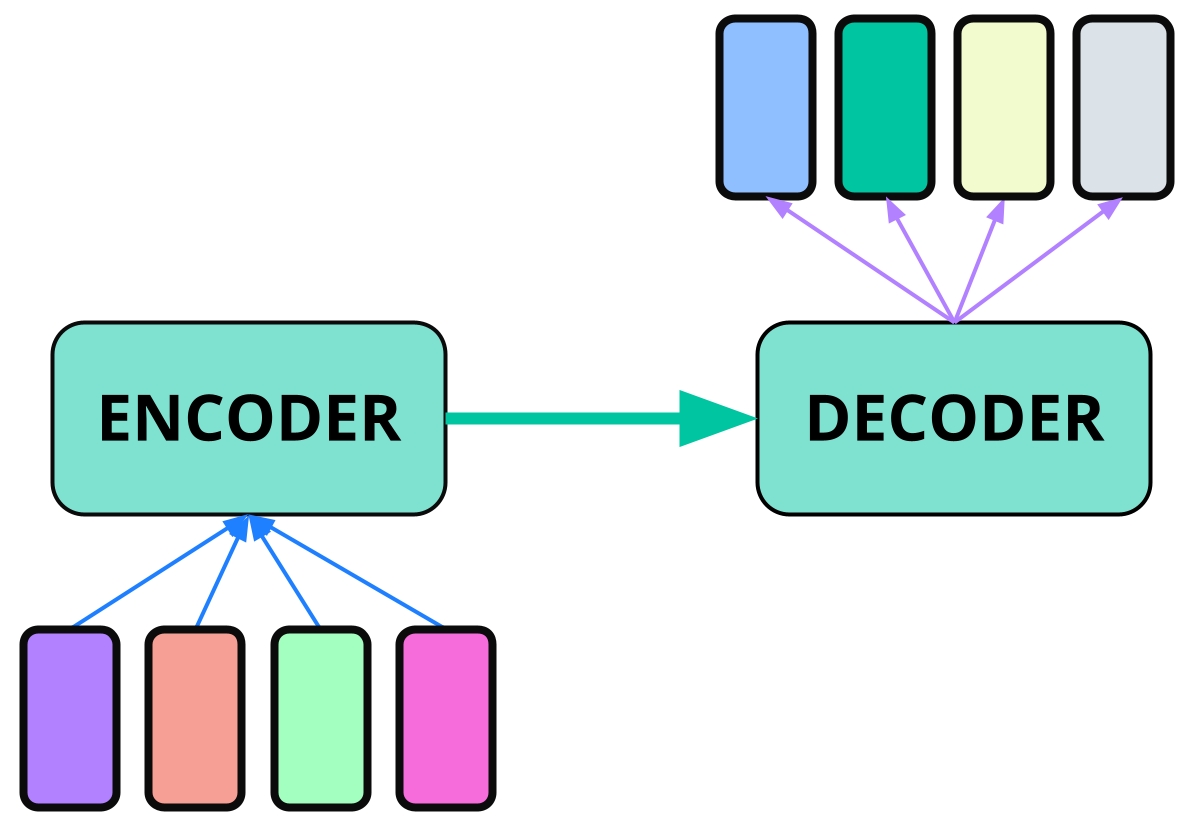
Text can be treated as a sequence of characters, words, tokens, or embeddings.

Depending on the task, there are different model architectures to handle the sequence.

Icono

Descripción generada automáticamente

For an **encoder-decoder** model, the ***encoder*** encodes the input sequence into a representation of context while the ***decoder*** decodes this representation to generate an output sequence.



## Models to to that

Diagrama

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente con confianza media

Autoregressive models tend to repeat the same tokens in its output sequence, so sampling methods are frequently used for choosing the next token where the probability of a particular token is chosen.

* Temperature: adjusts the randomness in choosing the next token
* Top-k sampling: samples from only the k most likely tokens
* Nucleus or top-p sampling: uses a dynamic cutoff for sampling the most likely tokens (cumulative probability is under p)
* Beam search: considers the likelihood of strings of multiple tokens instead of just a single next token

## Semantic search

Means that instead of simply looking for the exact keywords from the question, we are looking for text that is most similar to the question across a number of dimensions.

If semantic search means finding the most similar text, how do we measure that similarity? We need to define a measurement of "distance", and whichever piece of text has the shortest distance is the most similar.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Diagrama, Dibujo de ingeniería

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente