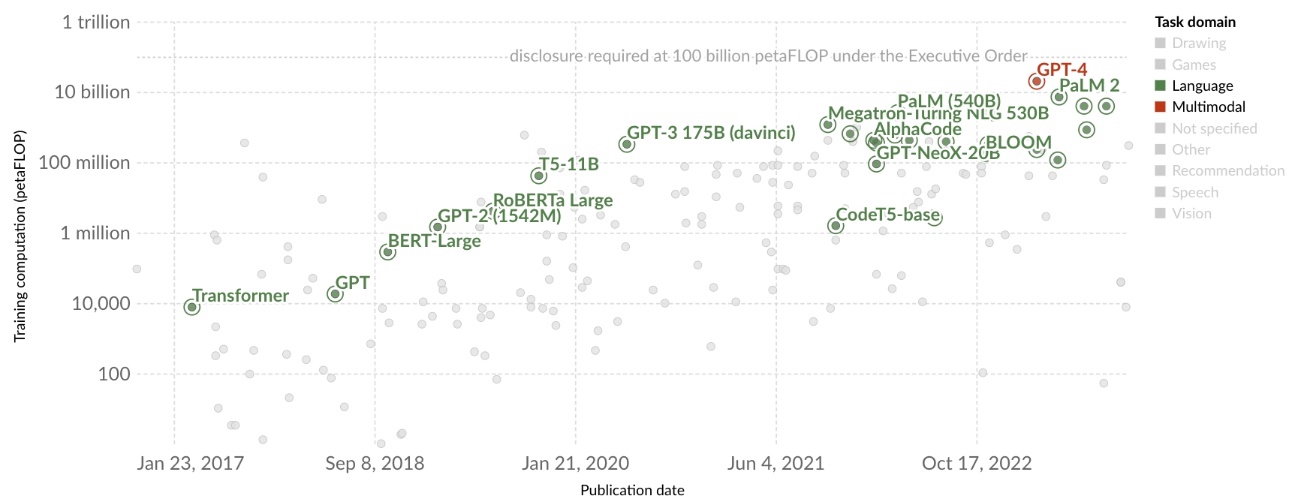
# LLM’s and Text Generation



## 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.