

Open in app ↗

Sign up

Sign In



Search Medium



# Two minutes NLP — Learn the ROUGE metric by examples

ROUGE-N, ROUGE-L, ROUGE-S, pros and cons, and ROUGE vs BLEU



Fabio Chiusano · [Follow](#)

Published in NLPlanet

5 min read · Jan 19, 2022



Listen



Share

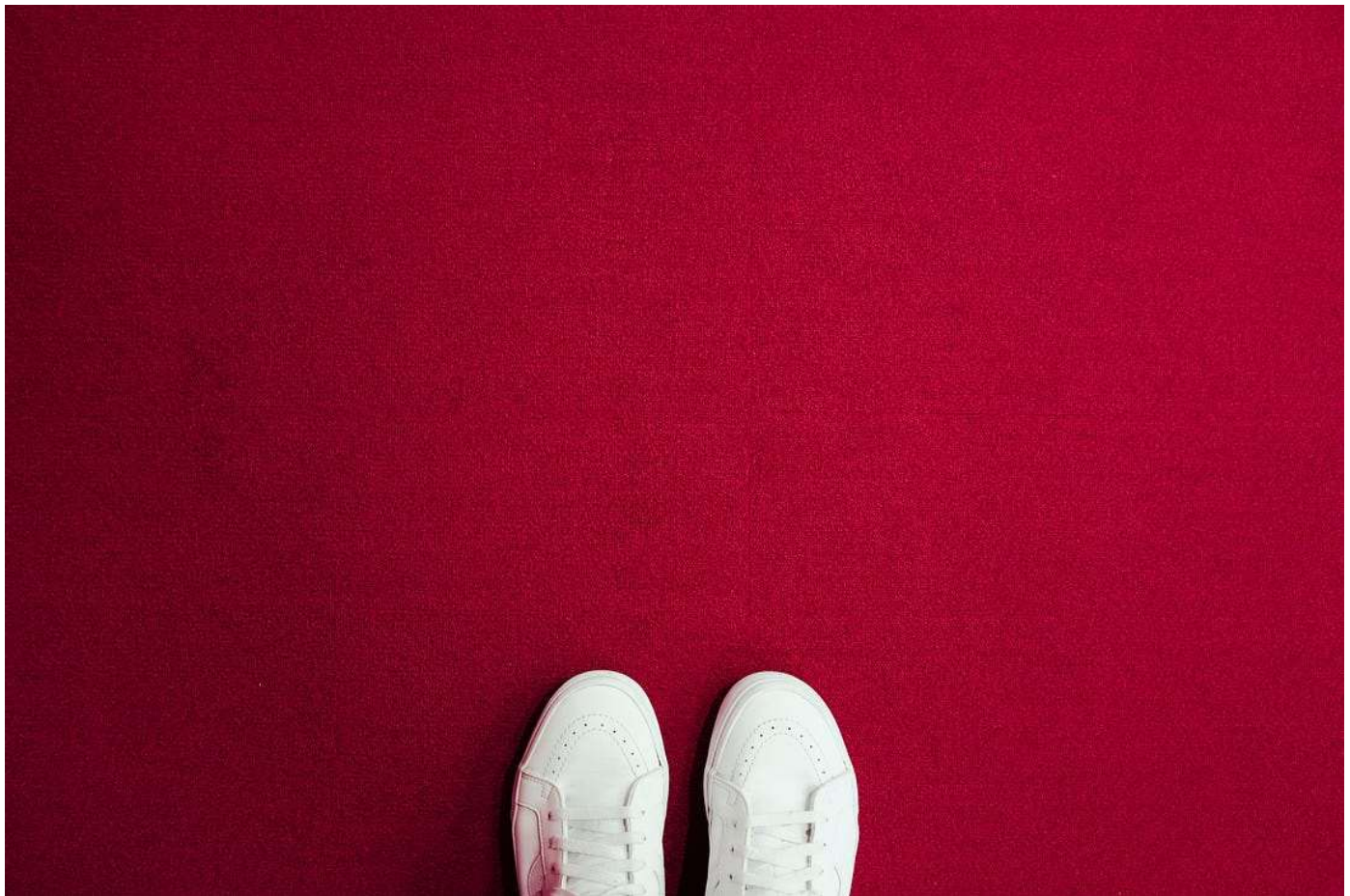


Photo by [Christian Chen](#) on [Unsplash](#)

**ROUGE** (Recall-Oriented Understudy for Gisting Evaluation), is a set of metrics and a software package specifically designed for evaluating automatic summarization, but that can be also used for machine translation. The metrics compare an automatically produced summary or translation against reference (high-quality and human-produced) summaries or translations.

In this article, we cover the main metrics used in the ROUGE package.

## ROUGE-N

ROUGE-N measures the number of matching n-grams between the model-generated text and a human-produced reference.

Consider the reference  $R$  and the candidate summary  $C$ :

- $R$ : The cat is on the mat.
- $C$ : The cat and the dog.

## ROUGE-1

Using  $R$  and  $C$ , we are going to compute the precision, recall, and F1-score of the matching n-grams. Let's start computing ROUGE-1 by considering 1-grams only.

ROUGE-1 precision can be computed as the ratio of the number of unigrams in  $C$  that appear also in  $R$  (that are the words “the”, “cat”, and “the”), over the number of unigrams in  $C$ .

$$\text{ROUGE-1 precision} = 3/5 = 0.6$$

ROUGE-1 recall can be computed as the ratio of the number of unigrams in  $R$  that appear also in  $C$  (that are the words “the”, “cat”, and “the”), over the number of unigrams in  $R$ .

$$\text{ROUGE-1 recall} = 3/6 = 0.5$$

Then, ROUGE-1 F1-score can be directly obtained from the ROUGE-1 precision and recall using the standard F1-score formula.

$$\text{ROUGE-1 F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) = 0.36$$

## ROUGE-2

Let's try computing the ROUGE-2 considering 2-grams.

Remember our reference  $R$  and candidate summary  $C$ :

- $R$ : The cat is on the mat.
- $C$ : The cat and the dog.

ROUGE-2 precision is the ratio of the number of 2-grams in  $C$  that appear also in  $R$  (only the 2-gram “the cat”), over the number of 2-grams in  $C$ .

$$\text{ROUGE-2 precision} = 1/4 = 0.25$$

ROUGE-1 recall is the ratio of the number of 2-grams in  $R$  that appear also in  $C$  (only the 2-gram “the cat”), over the number of 2-grams in  $R$ .

$$\text{ROUGE-2 recall} = 1/5 = 0.20$$

Therefore, the F1-score is:

$$\text{ROUGE-2 F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) = 0.22$$

## ROUGE-L

ROUGE-L is based on the longest common subsequence (LCS) between our model output and reference, i.e. the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both. A longer shared sequence should indicate more similarity between the two sequences.

We can compute ROUGE-L recall, precision, and F1-score just like we did with ROUGE-N, but this time we replace each n-gram match with the LCS.

Remember our reference  $R$  and candidate summary  $C$ :

- $R$ : The cat is on the mat.
- $C$ : The cat and the dog.

The LCS is the 3-gram “the cat the” (remember that the words are not necessarily consecutive), which appears in both  $R$  and  $C$ .

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in  $C$ .

$$\text{ROUGE-L precision} = 3/5 = 0.6$$

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in  $R$ .

$$\text{ROUGE-L recall} = 3/6 = 0.5$$

Therefore, the F1-score is:

$$\text{ROUGE-L F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) = 0.55$$

## ROUGE-S

ROUGE-S allows us to add a degree of leniency to the n-gram matching performed with ROUGE-N and ROUGE-L. ROUGE-S is a skip-gram concurrence metric: this allows to search for consecutive words from the reference text that appear in the model output but are separated by one-or-more other words.

Consider the new reference  $R$  and candidate summary  $C$ :

- $R$ : The cat is on the mat.
- $C$ : The gray cat and the dog.

If we consider the 2-gram “the cat”, the ROUGE-2 metric would match it only if it appears in *C* exactly, but this is not the case since *C* contains “the gray cat”. However, using ROUGE-S with unigram skipping, “the cat” would match “the gray cat” too.

We can compute ROUGE-S precision, recall, and F1-score in the same way as the other ROUGE metrics.

## Pros and Cons of ROUGE

This is the tradeoff to take into account when using ROUGE.

- *Pros*: it correlates positively with human evaluation, it's inexpensive to compute and language-independent.
- *Cons*: ROUGE does not manage different words that have the same meaning, as it measures syntactical matches rather than semantics.

## ROUGE vs BLEU

In case you don't know the BLEU metric already, I suggest that you read the companion article [Learn the BLEU metric by examples](#) to get a grasp on it.

In general:

- BLEU focuses on precision: how much the words (and/or n-grams) in the candidate model outputs appear in the human reference.
- ROUGE focuses on recall: how much the words (and/or n-grams) in the human references appear in the candidate model outputs.

These results are complementing, as is often the case in the precision-recall tradeoff.

## Computing ROUGE with Python

Implementing the ROUGE metrics in Python is easy thanks to the [Python rouge library](#), where you can find ROUGE-1, ROUGE-2, and ROUGE-L. Although present in the rouge paper, ROUGE-S would seem that over time it has been used less and less.

Thank you for reading! If you are interested in learning more about NLP, remember to follow NLPlanet on [Medium](#), [LinkedIn](#), and [Twitter](#)!

## NLPlanet related posts

### Two minutes NLP — Learn the BLEU metric by examples

BLEU, n-grams, geometric mean, and brevity penalty

medium.com

### Awesome NLP — 18 High-Quality Resources for studying NLP

Tutorials, code examples, video courses, course notes, and articles

medium.com

### Two minutes NLP — Gopher Language Model performance in a nutshell

Gopher, GPT-3, Jurassic-1, and Megatron-Turing NLG

medium.com

NLP

Naturallanguageprocessing

Data Science

Machine Learning

Artificial Intelligence



Follow

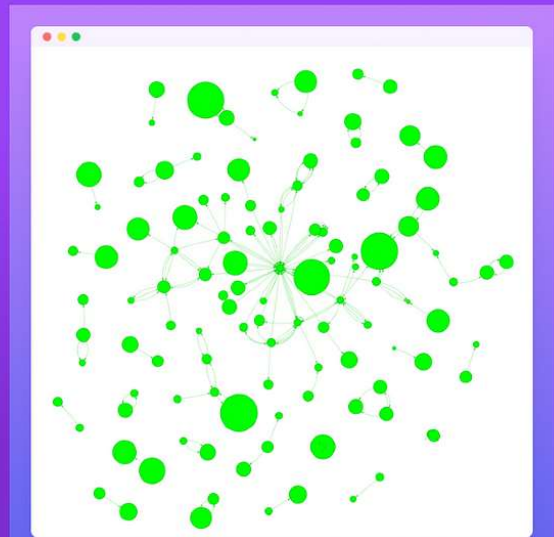
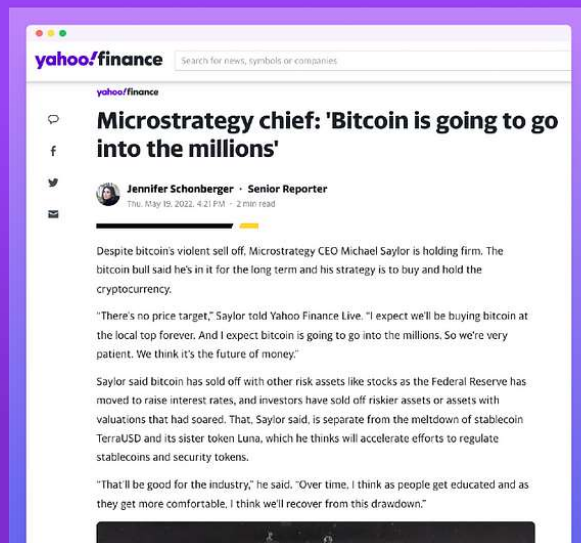
## Written by Fabio Chiusano

3.2K Followers · Editor for NLPlanet

Freelance data scientist — Top Medium writer in Artificial Intelligence

### More from Fabio Chiusano and NLPlanet

# Building a Knowledge Base from Texts



Fabio Chiusano in NLPlanet

## Building a Knowledge Base from Texts: a Full Practical Example

Implementing a pipeline for extracting a Knowledge Base from texts or online articles

12 min read · May 24, 2022



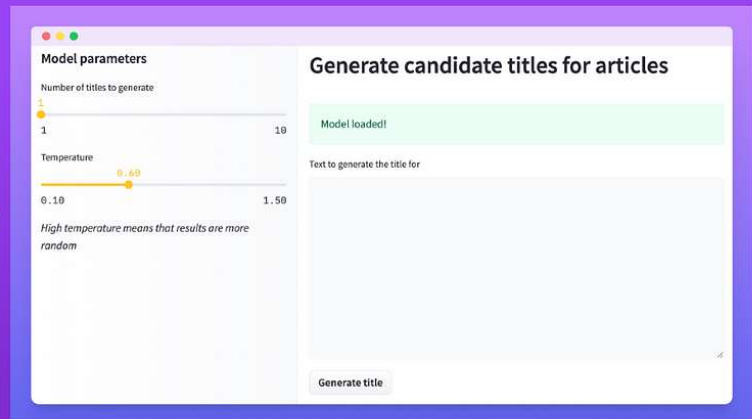
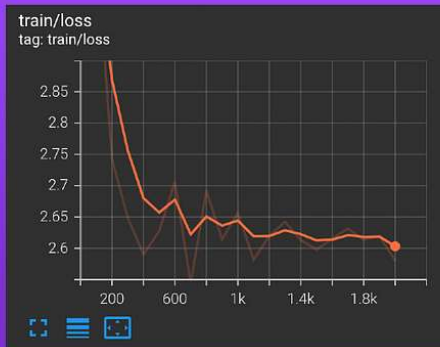
1K



8



# Finetuning T5 + Streamlit



Fabio Chiusano in NLPlanet

## A Full Guide to Finetuning T5 for Text2Text and Building a Demo with Streamlit

All you need to know to build a full demo: Hugging Face Hub, Tensorboard, Streamlit, and Hugging Face Spaces

15 min read · May 17, 2022



361



6







Fabio Chiusano in NLPlanet

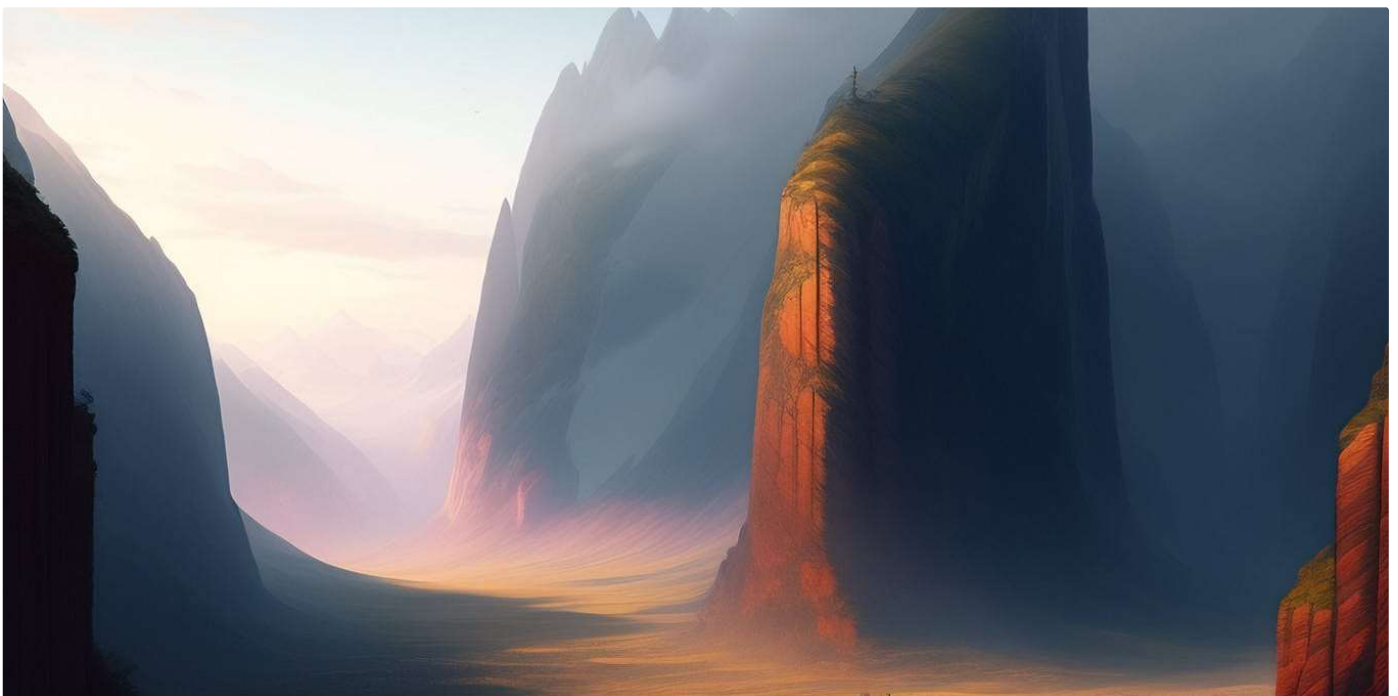
## Weekly AI and NLP News—July 3rd 2023


DeepMind's Gemini model, ElevenLabs Voice Library, and more funding

5 min read · Jul 3



30



 Fabio Chiusano in NLPlanet

Weekly AI and NLP News— July 10th 2023

GPT-4 general availability, best GPUs for deep learning, and scaling Transformers to 1B tokens

5 min read · Jul 10

 28





See all from Fabio Chiusano

See all from NLPlanet

Recommended from Medium



 Ruben Winastwan in Towards Data Science



## Semantic Textual Similarity with BERT

How to use BERT to calculate the semantic similarity between two texts

★ • 11 min read • Feb 15

👏 183



Is there a  
subscription fee  
for ChatGPT  
Plus?

**Extractive:** \$20

**Abstractive:** Yes, there is  
a \$20 subscription fee for  
ChatGPT Plus.

OpenAI is launching a premium and paid-for version of ChatGPT. The free app will remain available. But it is liable to go offline during busy periods – and, during those, the people who have paid its monthly fee will have priority access. That is just one of the perks offered in return for the \$20 subscription to “ChatGPT Plus”.



Skanda Vivek in Towards Data Science

## Extractive vs Generative Q&A — Which is better for your business?

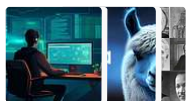
The arrival of ChatGPT hints at a new era of search engines, this tutorial dives into the 2 basic types of AI based question answering

★ • 6 min read • Feb 6

👏 59

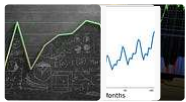


### Lists



#### Natural Language Processing

444 stories • 82 saves



## Predictive Modeling w/ Python

18 stories · 176 saves



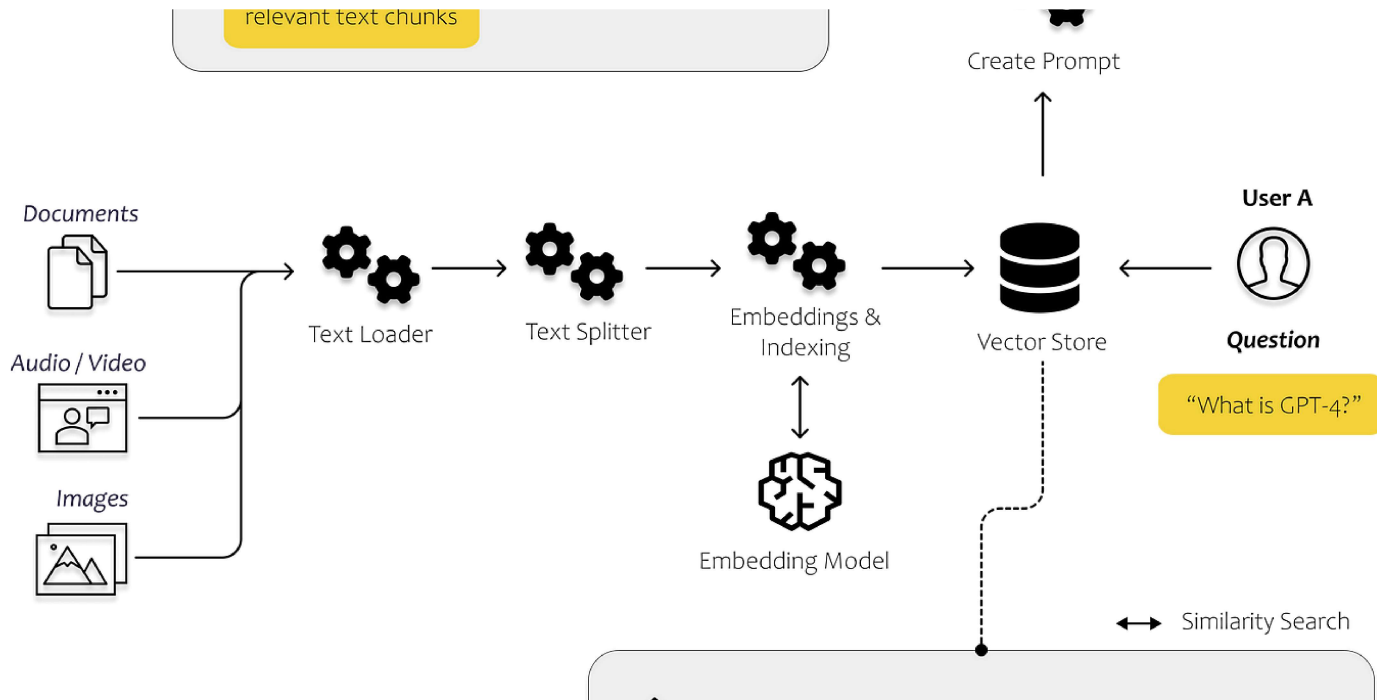
## Practical Guides to Machine Learning

10 stories · 194 saves



## ChatGPT prompts

22 stories · 170 saves



Dominik Polzer in Towards Data Science

## All You Need to Know to Build Your First LLM App

A step-by-step tutorial to document loaders, embeddings, vector stores and prompt templates

★ · 26 min read · Jun 21

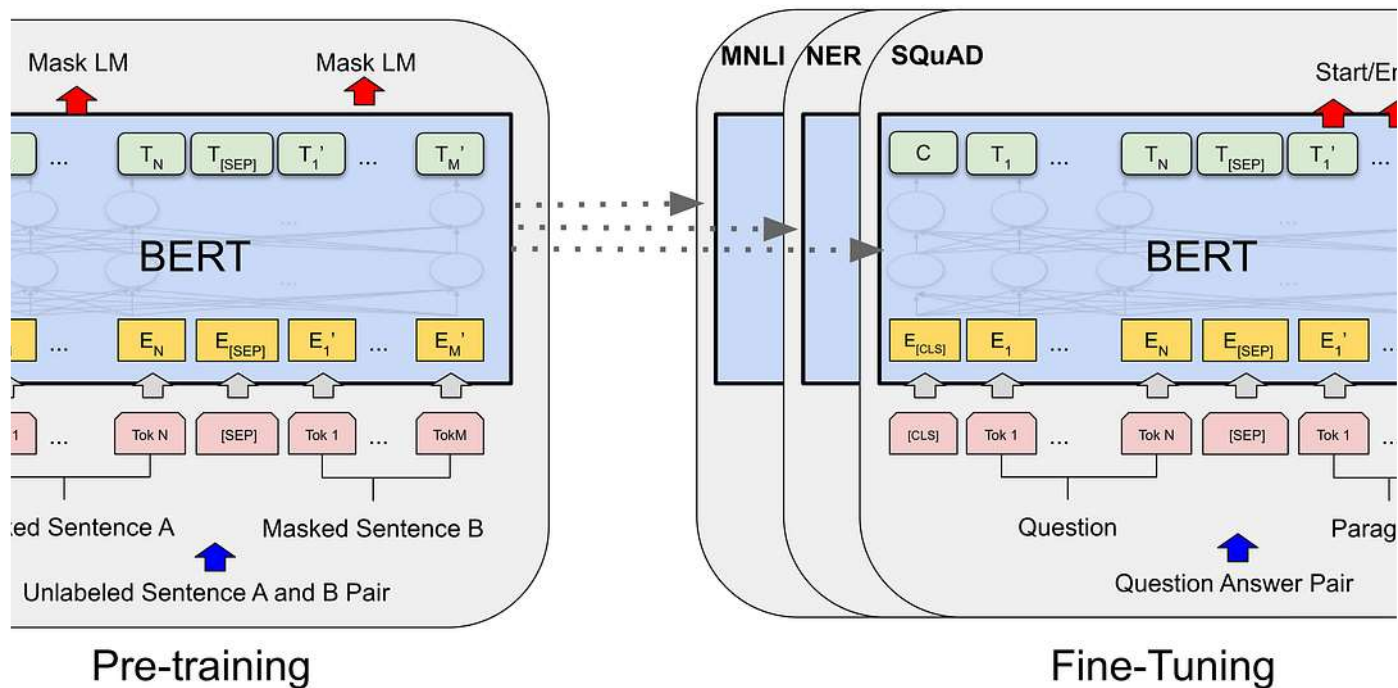


3.4K



30





Cassius in Cassandra

## Introducing the BERT Language Model

A Gentle Introduction to BERT and natural language processing (NLP) models for Beginners.



4 min read · Mar 21



42



1



Tell me how ChatGPT works.



ChatGPT is a large language model that uses deep learning techniques to generate human-like text. It is based on the GPT (Generative Pre-trained Transformer) architecture, which uses a transformer neural network to process and generate text. The model is pre-trained on a massive dataset of text, such as books, articles, and websites, so it can understand the patterns and structure of natural language. When given a prompt or a starting point, the model uses this pre-trained knowledge to generate text that continues the given input in a coherent and natural way.





Molly Ruby in Towards Data Science

## How ChatGPT Works: The Models Behind The Bot

A brief introduction to the intuition and methodology behind the chat bot you can't stop hearing about.

🌟 • 8 min read • Jan 30



7.6K



127



Leonie Monigatti in Towards Data Science

## Getting Started with LangChain: A Beginner's Guide to Building LLM-Powered Applications

A LangChain tutorial to build anything with large language models in Python

🌟 • 12 min read • Apr 25



3.6K



23

[See more recommendations](#)