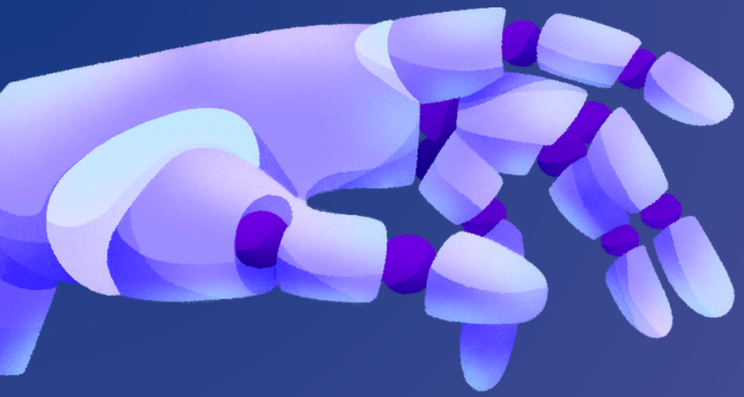


LARGE LANGUAGE MODELS API

INDEX

LLM FUNDAMENTALS

They are artificial intelligence models designed to process and generate human language. They are trained with large amounts of text, which allows them to learn patterns, syntax and context of the language. Some well-known examples of LLMs are GPT (like the model you are interacting with), BERT, and T5.

- **Training on large volumes of data.**
- **Transformer architecture:** The basis of LLMs, which uses the attention mechanism to process the context of words in a sequence, regardless of their distance in the text.
- **Word Embeddings (Word Embeddings):** Words are transformed into numeric vectors (Word Embeddings), representing their semantic meaning and allowing neural networks to process them.
- **Text generation:** LLMs are able to generate coherent text, completing sentences, answering questions or creating content from inputs.
- **Fine-tuning and Transfer Learning:** LLMs can adapt to specific tasks (such as classification or translation) by fine-tuning, taking advantage of general knowledge acquired during their training.

TRAINING PROCESS

1 PRE-TRAINING

- Unsupervised training
- Unlabelled text
- No specific output or input pairs

2 FINE-TUNING

- Downstream tasks (NLG, Classification etc)
- Input-output pairs
- Supervised training
- Labelled data

3 IN-PUT ENCODING

- Convert Text to numeric
- More tokenisation
- Positional encoding (tokens sequence and relations)
- Embedding (co-occurence of tokens in corpora)

4 AUTOREGRESSIVE GENERATION (GPT)

- Predicting tokens one at a time
- Uses seed promp

5 SAMPLING OR BEAM SEARCH

- Probabilistic decoding technique
- Sampling probability (creativity)
- Beam search (deterministic)

HOW DOES API WORK?

API (Application Programming Interface) is a tool that allows two applications to communicate with each other.

CLIENT ➔ REQUEST ➔ API ➔ SERVER ➔ RESPONSE ➔ CLIENT.

1. Request:
   - The client sends a request to the API.
   - The request includes a URL and sometimes additional data.
2. Processing:
   - The API reads the request and forwards it to the server.
3. Response:
   - The server processes the request and sends back the needed data (often in JSON or XML format).
4. Integration:
   - The client integrates the data into its interface for the user to see.

COMMON PARAMETERS

1 NUM_RETURN_SEQUENCES

Specifies how many different responses you want the model to generate in a single request.

2 MAX_LENGTH

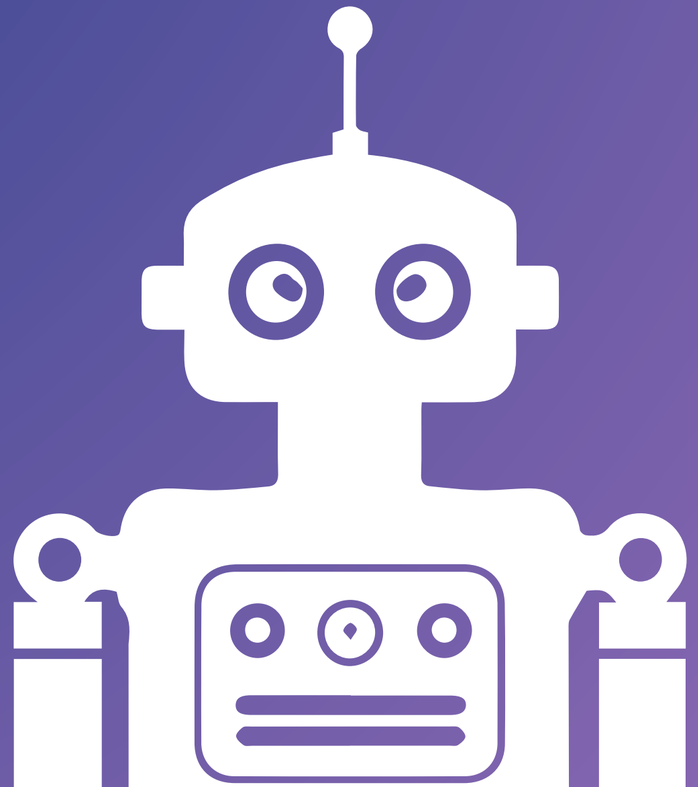
- Controls the maximum number of tokens the model can generate.
- Recommended value: Between 50 and 200, depending on the task.

3 TEMPERATURE

Adjusts the randomness of the responses:

- Low (0-0.3): Predictable and formal responses.
- Medium (0.7): Balanced responses.

High (1.0 or more): Creative but less consistent responses.



4 TOP_K

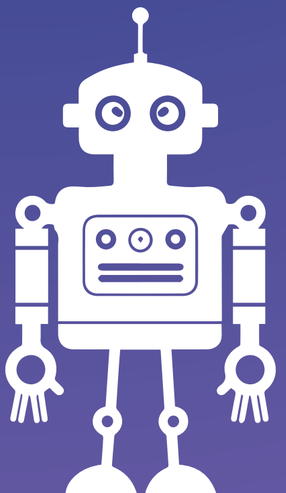
- Limits the number of possible words the model considers for the next token.
- Values:
- High (50-100): More creative responses.
- Low (10): More predictable responses.

5 TOP_P

- Also known as "nucleus sampling," it considers only words whose cumulative probability is below a certain threshold.
- Values:
- Low (0.3-0.5): More conservative responses.
- High (0.9): More diverse responses.

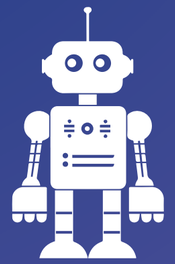
COMMON PARAMETERS - PRACTICAL EXAMPLE

```
payload = {
    "inputs": "What are the best places to visit in Paris?",
    "parameters": {
        "max_length": 150,
        "temperature": 0.7,
        "top_p": 0.9,
        "num_return_sequences": 2
    }
}
```



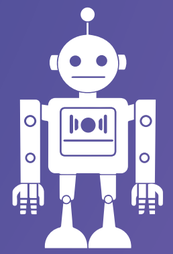
PARAMETER	PURPOSE	RECOMMENDED VALUE
max_length	Maximum response length	100-200
temperature	Creativity of the response	0.7
top_k	Limits the number of options	50-100
top_p	Filters based on probabilities	0.9
num_return_sequences	Number of responses per request	1-3

COMMON USES



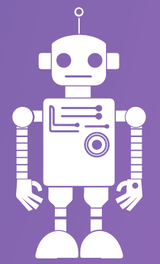
CHATBOTS

They power chatbots and virtual assistants to provide automated, conversational support for resolving customer queries, troubleshooting, and FAQs.



LANGUAGE TRANSLATIONS

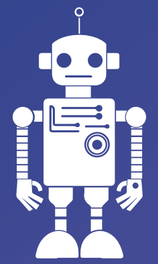
LLMs enable high-quality translation between multiple languages, helping bridge communication gaps and support international collaboration.



CONTENT GENERATION

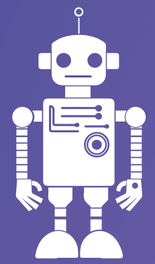
LLMs assist in writing articles, blogs, marketing copy, and creative texts like stories or poetry, helping content creators save time and enhance productivity.

COMMON USES



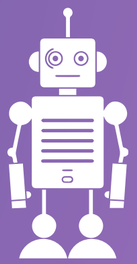
CODE ASSINTANCE

Developers use LLMs for code generation, debugging, and documentation, streamlining software development and reducing errors.



EDUCATION AND TUTORING

These models provide personalized learning experiences, explaining complex topics, answering questions, and creating study materials.



DATA ANALYSIS AND SUMMARIZATION

LLMs process and summarize large volumes of text, extracting insights and generating concise reports, aiding decision-making across various fields.

API EXAMPLES

API classes are sets of methods that an LLM system offers to f**acilitate different tasks** or services, and they are **organized according to their purpose** or type of use.

These classes allow developers to interact with the language model without having to deal with the internal details of the model

TYPES OF MODELS

TEXT

AUDIO

MULTIMODAL

VISION

VIDEO

REINFORCEMENT LEARNONG

TIME SERIES

GRAPH MODELS

API EXAMPLES

llama3

TRANSFORMER ARCHITECTURE

Bert

TWO-WAY TRAINING

ResNet

RESIDUAL CONNECTIONS

GRACIAS POR VUESTRA ATENCIÓN