# Understanding Tokenizers in Hugging Face Transformers

This document explains the relationship between various models and their respective tokenizers in the Hugging Face Transformers library. Each model has a specific tokenizer type that ensures text is preprocessed correctly before being fed into the model.

## Model vs Tokenizer – In Detail

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Tokenizer | Tokenization Type | Example Output | Notes |
| bert-base-uncased | BertTokenizer | WordPiece | ['working', 'on', 'res', '##umes'] | Breaks words into subwords using ## to indicate continuation. Must use with BERT-based models. |
| t5-small | T5Tokenizer | SentencePiece | ['▁Summarize', '▁this', '▁res', 'ume', '.'] | Uses special character '▁' for word boundaries. Best for summarization and translation tasks. |
| facebook/bart-base | BartTokenizer | Byte-Pair Encoding (BPE) | Splits frequent character patterns efficiently. | Encoder-decoder architecture. Ideal for text generation. |
| gpt2 | GPT2Tokenizer | Byte-Level BPE | Handles emojis, special symbols, code-like text. | Does not use padding by default. Great for natural text generation. |
| mistralai/Mistral-7B / LLaMA | LlamaTokenizer | Custom SentencePiece | Handles custom vocab and special tokens. | Must be used with LLaMA-based models due to custom rules. |

## How AutoTokenizer Helps

AutoTokenizer is a convenience class in Hugging Face that automatically selects the correct tokenizer based on the model name. This avoids manual imports and ensures the model receives properly tokenized inputs.