**Transformers: Encoder-Only, Decoder-Only, and Encoder-Decoder Models**

**1. Encoder-Only Models (e.g., BERT, RoBERTa, DeBERTa)**

* **Architecture:** Uses only the encoder part of a transformer.
* **How It Works:** Processes the entire input at once and captures relationships between words.
* **Pre-Training:**
  + **Masked Language Modeling (MLM):** Random words are masked, and the model predicts them.
  + **Next Sentence Prediction (NSP) (for BERT):** Determines if one sentence follows another logically.
* **Best Suited For:**
  + Text classification
  + Sentiment analysis
  + Named Entity Recognition (NER)
  + Question answering
  + Semantic search
* **Why?** Captures rich contextual meaning but doesn’t generate new text.

**2. Decoder-Only Models (e.g., GPT, GPT-2, GPT-3, LLaMA)**

* **Architecture:** Uses only the decoder part of a transformer.
* **How It Works:** Predicts words sequentially, generating text in an auto-regressive manner.
* **Pre-Training:**
  + **Causal Language Modeling (CLM):** Predicts the next word based only on previous words.
* **Best Suited For:**
  + Text generation (e.g., chatbots, stories, articles)
  + Code generation (e.g., GitHub Copilot)
  + Autocomplete and text completion
* **Why?** Designed for sequential word prediction but lacks deep understanding of full input sequences.

**3. Encoder-Decoder (Seq2Seq) Models (e.g., T5, BART, mT5)**

* **Architecture:** Uses both encoder and decoder components.
* **How It Works:**
  + The encoder processes input text to understand its meaning.
  + The decoder generates new text based on the encoded representation.
* **Pre-Training:**
  + **Text Infilling (for BART):** Random parts of the input text are deleted or corrupted and must be reconstructed.
  + **Span Corruption (for T5):** Random phrases are replaced with a single [MASK] token.
  + **Sentence Reordering:** Scrambled sentences are reordered correctly.
* **Best Suited For:**
  + Machine translation
  + Text summarization
  + Structured text generation
  + Question answering (where both input understanding and response generation are needed)
* **Why?** Combines understanding and controlled generation, making it ideal for structured NLP tasks.

**Comparison Summary**

| **Model Type** | **Architecture** | **Pre-Training Task** | **Best For** |
| --- | --- | --- | --- |
| **Encoder-Only** | BERT, RoBERTa | Masked Language Modeling (MLM) | Understanding tasks (classification, search, QA) |
| **Decoder-Only** | GPT, LLaMA | Causal Language Modeling (CLM) | Text generation (chatbots, stories, code) |
| **Encoder-Decoder** | T5, BART | Span Corruption, Text Infilling | Structured generation (translation, summarization) |

**Key Takeaways**

* **Use an encoder-only model** (BERT) for **text understanding** tasks.
* **Use a decoder-only model** (GPT) for **text generation** tasks.
* **Use an encoder-decoder model** (T5, BART) for **structured text generation** (translation, summarization, etc.).