- The two major architectures for language models. The first is encoders, and the second, decoders. These architectures largely correspond to two different tasks/capabilities.

- The first capability is embedding and the second is text generation.

- Embedding text means the process of converting a sequence of words into a single vector or a sequence of vectors.

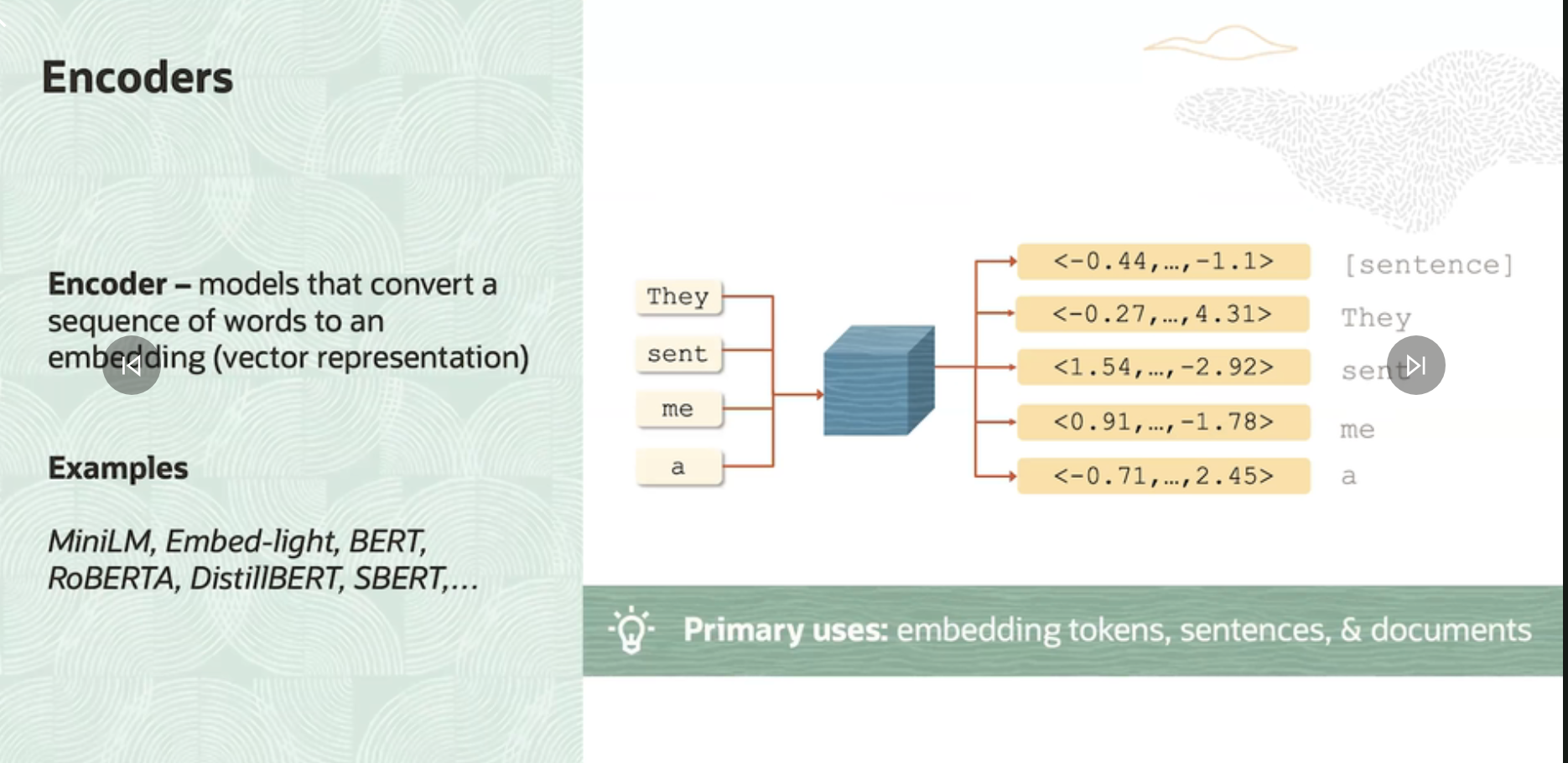
- Encoder models are designed to encode text, that is produce embeddings, and decoder models are designed to decode or generate text.

- Three different buckets corresponding to architecture-- encoders, decoders, and a combination model called an encoder-decoder.

- Decoder models tend to be pretty large, especially compared to the relatively smaller encoders.

- when models are too small, they tend to be poor text generators.

- Encoders are designed to embed text, that means taking a sequence of words and converting it to vectors.

-

- Let's say you want to take an input text snippet and retrieve a similar document from a corpus.

To accomplish this, you could encode or synonymously embed each document in the corpus and store them in an index. When you get the input snippet, you encode that, too, and check the similarity of the encoded input against the similarity of each document in the corpus. And then you return the most similar.

-Decoder models take a sequence of tokens and emit the next token in the sequence, based on the probability of the vocabulary which they compute.

- A decoder only produces a single token at a time.

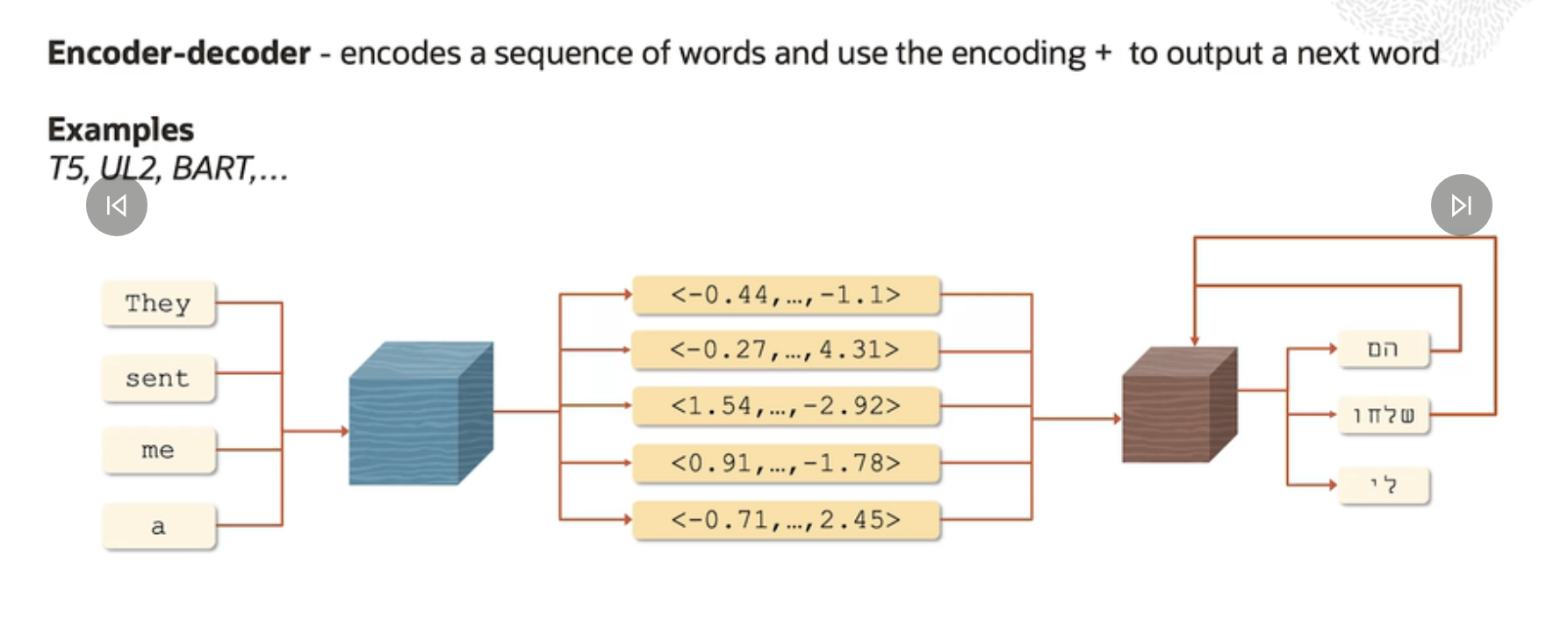
- We can always invoke a decoder over and over to generate as many new tokens as we want.

- To generate a sequence of new tokens with a decoder, what we need to do is, first, feed in a sequence of tokens and invoke the model to produce the next token.

Then append the generated token to the input sequence and feed it back to the model so that it can produce the second token. But, this is computationally very expensive. Typically, you would not use a decoder model for embedding. Instead, you'd use an encoder.

- Decoders are really in vogue now because they've shown tremendous capability for generating fluent text. Moreover, they've been shown capable of doing things like answering questions, participating dialogue, and more.

- Encoder-decoder models- Encoder-decoder models have primarily been utilized for sequence-to-sequence tasks like translation.

- 

- How this works? We send the English tokens to the model. They get passed to the encoder, which will embed all the tokens in addition to the sentence. And then the embeddings get passed to the decoder, which then decodes words one at a time.

