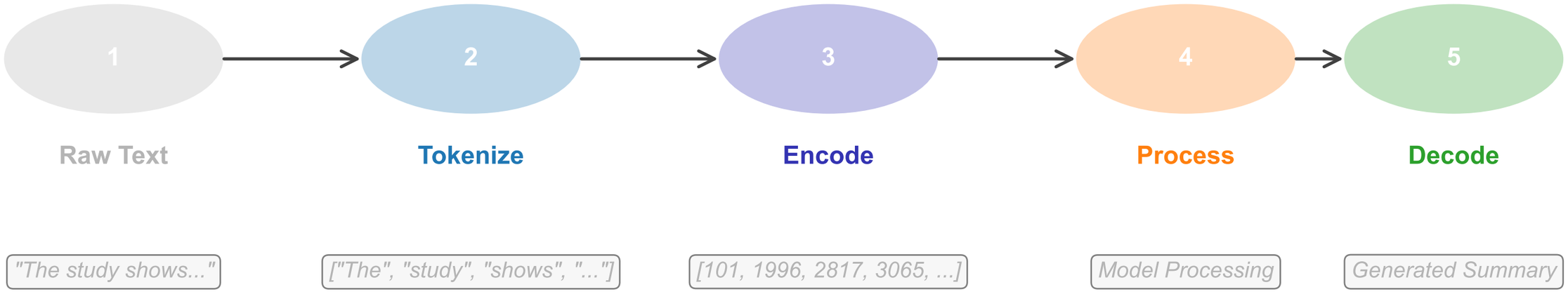
Tokenization and Encoding Pipeline



Special tokens: [CLS], [SEP], [PAD]

Vocabulary size: 30,000-50,000

Max length: 512-2048 tokens