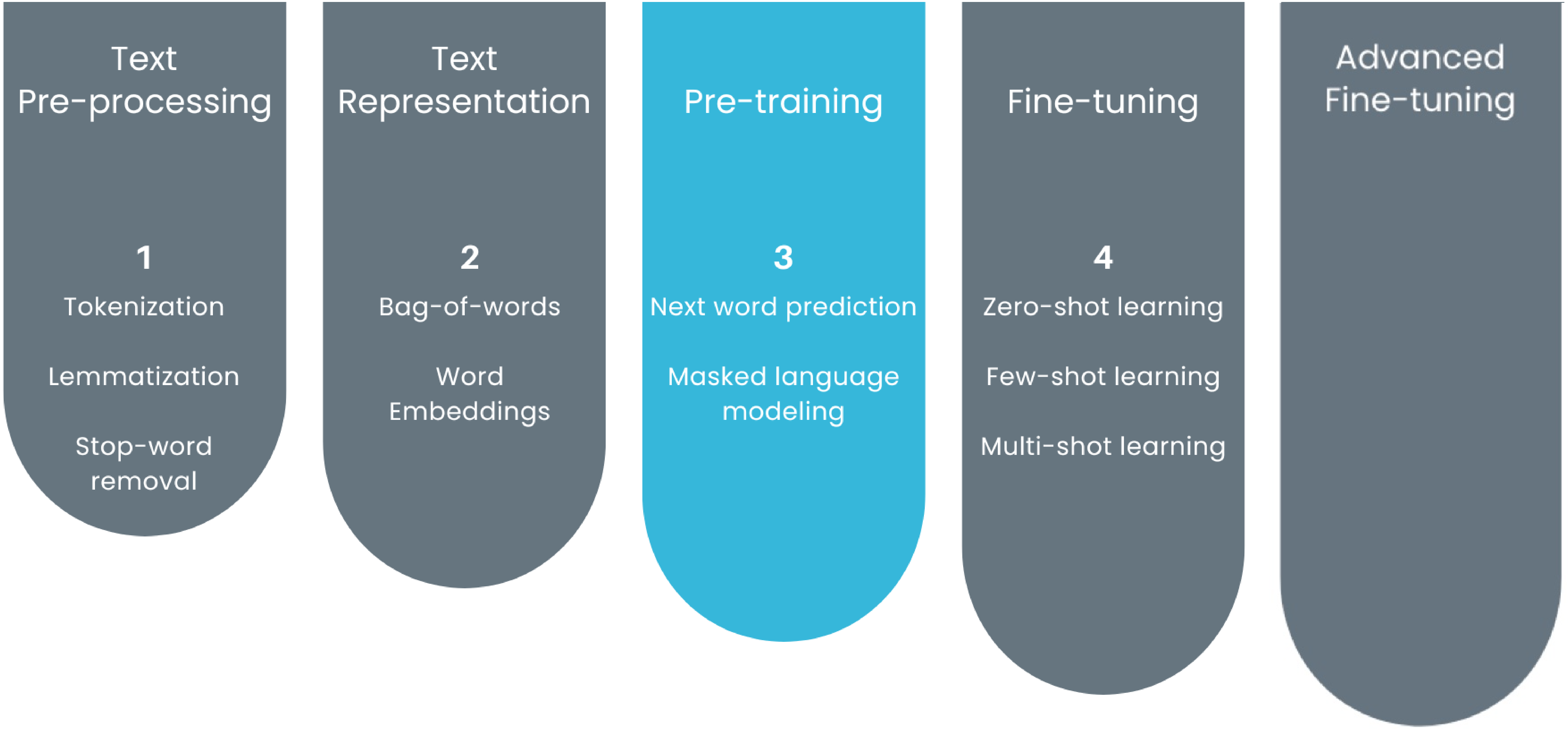
Building blocks to train LLMs

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

Vidhi Chugh
AI strategist and ethicist

datacamp

Where are we?



Text Pre-processing

1

Tokenization

Lemmatization

Stop-word removal

Text Representation

2

Bag-of-words

Word Embeddings

Pre-training

3

Next word prediction

Masked language modeling

Fine-tuning

4

Zero-shot learning

Few-shot learning

Multi-shot learning

Advanced Fine-tuning

Generative pre-training

- Trained using generative pre-training
  - Input data of text tokens
  - Trained to predict the tokens within the dataset

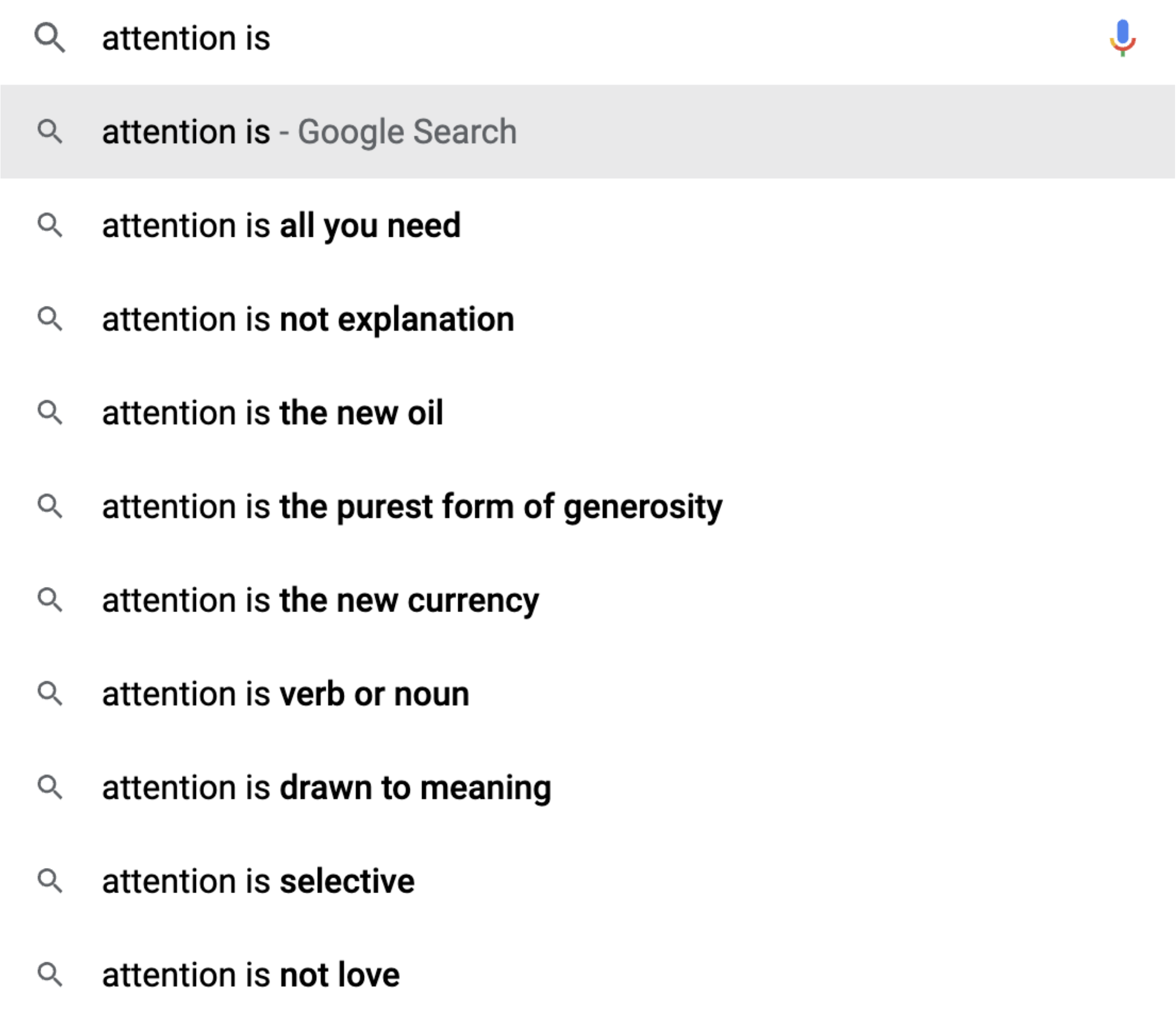
- Types:
  - Next word prediction
  - Masked language modeling

Next word prediction

- Supervised learning technique
    - Model trained on input-output pairs

- Predicts next word and generates coherent text
- Captures the dependencies between words

- Training Data
    - Pairs of input and output examples



attention is

attention is - Google Search

attention is **all you need**

attention is **not explanation**

attention is **the new oil**

attention is **the purest form of generosity**

attention is **the new currency**

attention is **verb or noun**

attention is **drawn to meaning**

attention is **selective**

attention is **not love**

Training data for next word prediction

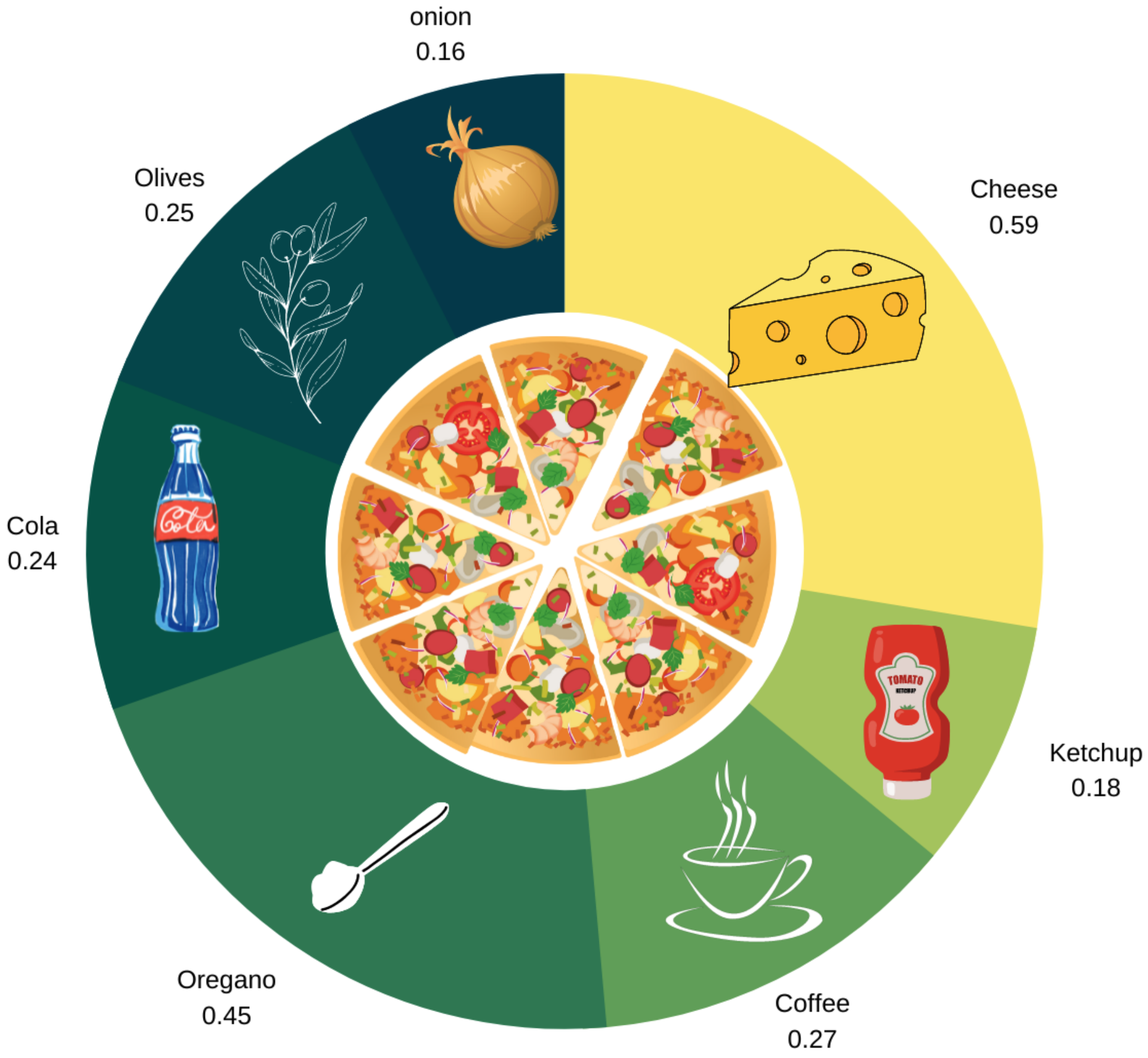
Input	Output
The quick brown	fox
The quick brown fox	jumps
The quick brown fox jumps	over
The quick brown fox jumps over	the
The quick brown fox jumps over the	lazy
The quick brown fox jumps over the lazy	dog
The quick brown fox jumps over the lazy dog.	

Which word relates more with pizza?

- More examples = better prediction

- **For example:**
  - I love to eat pizza with _ _ _ _ _ _

- Cheese is more related with pizza than anything else



Masked language modeling

- Hides a selective word
- Trained model predicts the masked word

- **Original Text:** "The quick brown fox jumps over the lazy dog."
- **Masked Text:** "The quick [MASK] fox jumps over the lazy dog."

- **Objective:** predict the missing word
- Based on learnings from training data

Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

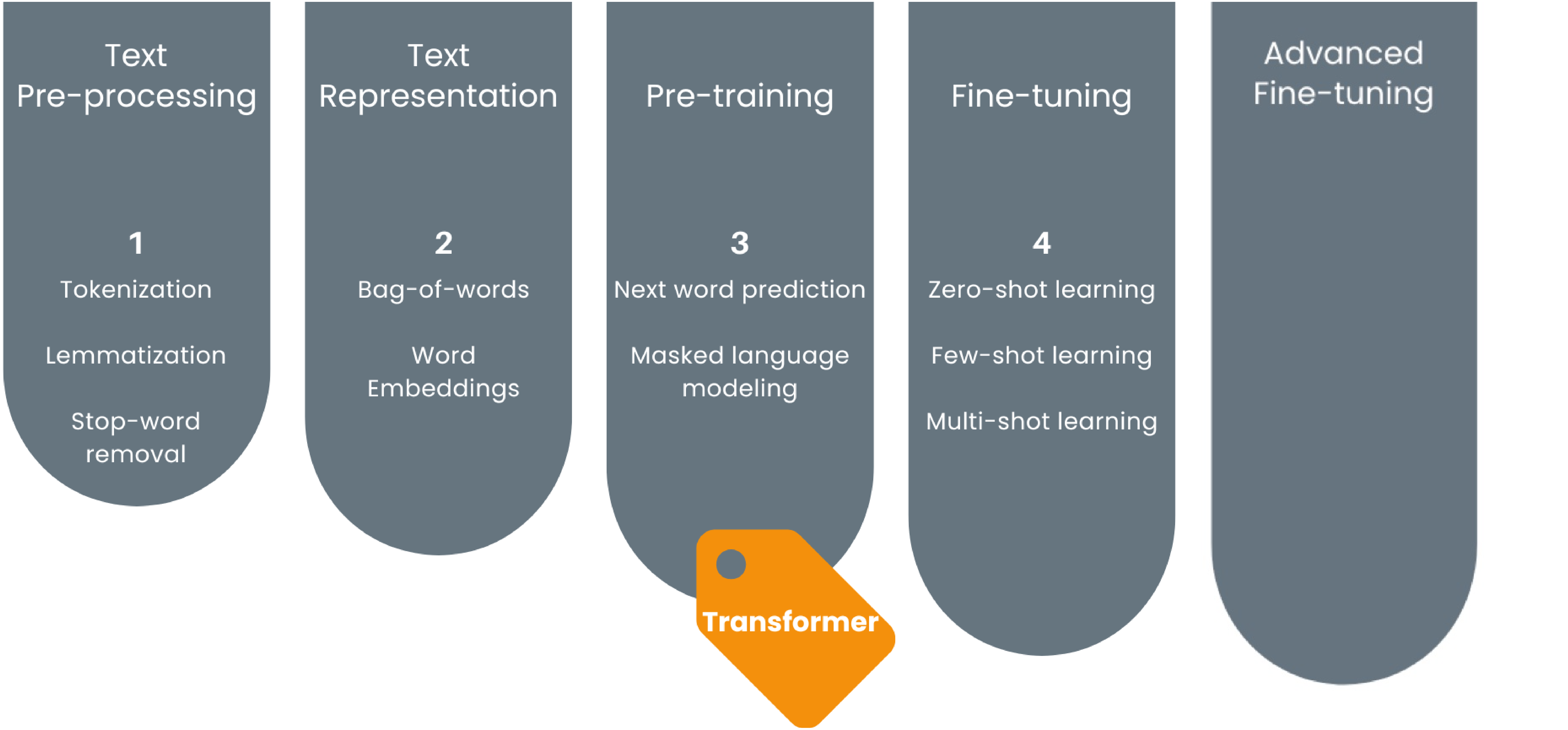
Introducing the transformer

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

Vidhi Chugh

AI strategist and ethicist

Where are we?



Text Pre-processing

1

Tokenization

Lemmatization

Stop-word removal

Text Representation

2

Bag-of-words

Word Embeddings

Pre-training

3

Next word prediction

Masked language modeling

Transformer

Fine-tuning

4

Zero-shot learning

Few-shot learning

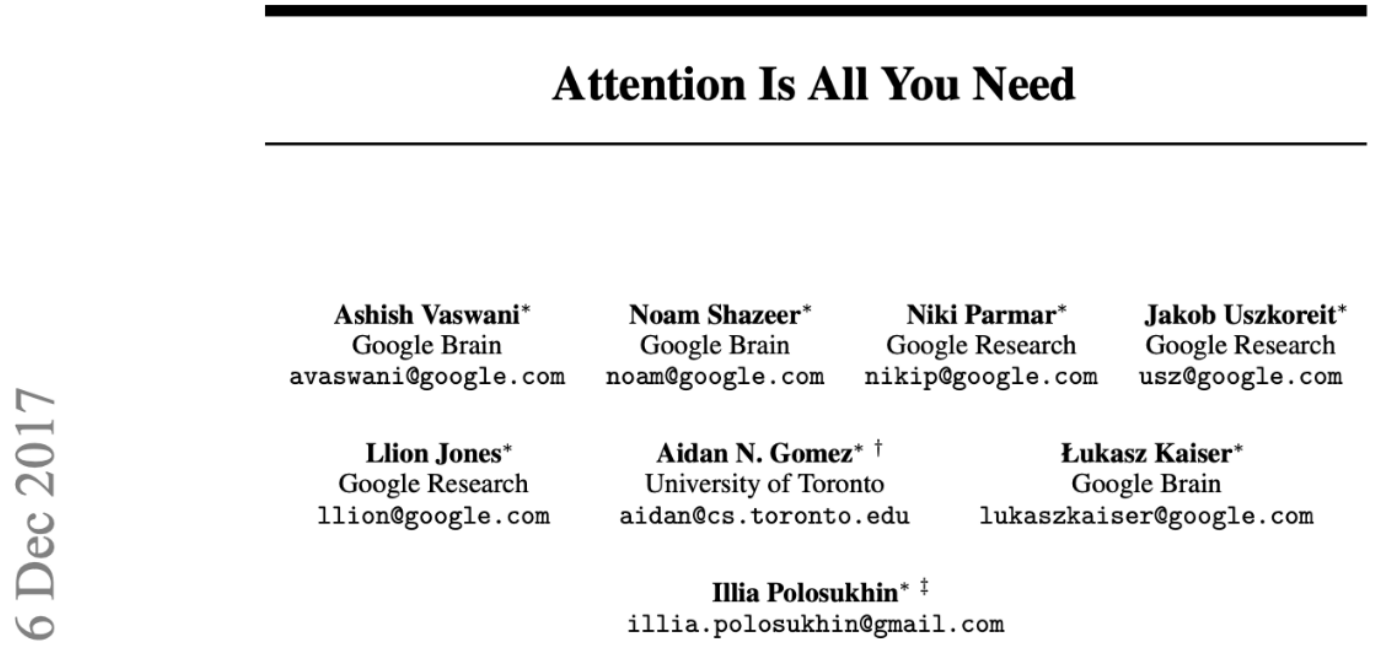
Multi-shot learning

Advanced Fine-tuning

What is a transformer?

- **"Attention Is All You Need"**
  - Revolutionized language modeling

- **Transformer architecture**
  - Relationship between words
  - Components: Pre-processing, Positional Encoding, Encoders, and Decoders



6 Dec 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

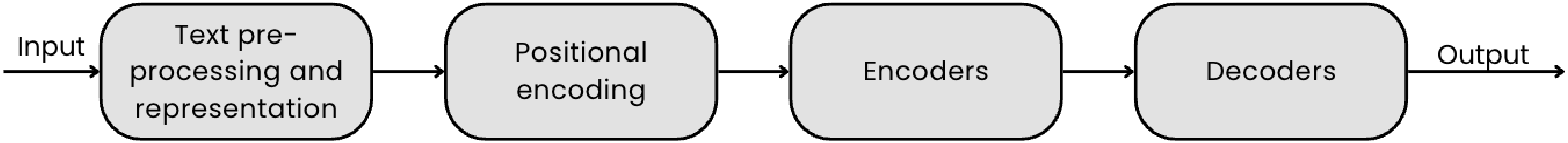
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

[1] arXiv: Attention Is All You Need

Inside the transformer

- **Input:** Jane, who lives in New York and works as a software

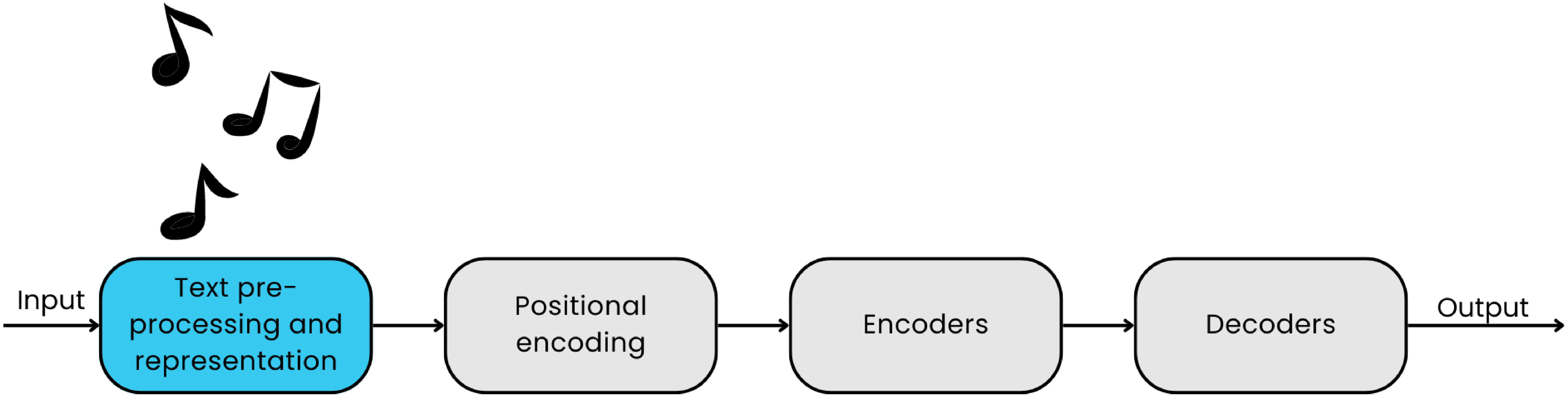


- **Output:** engineer, loves exploring new restaurants in the city.

Transformers are like an orchestra

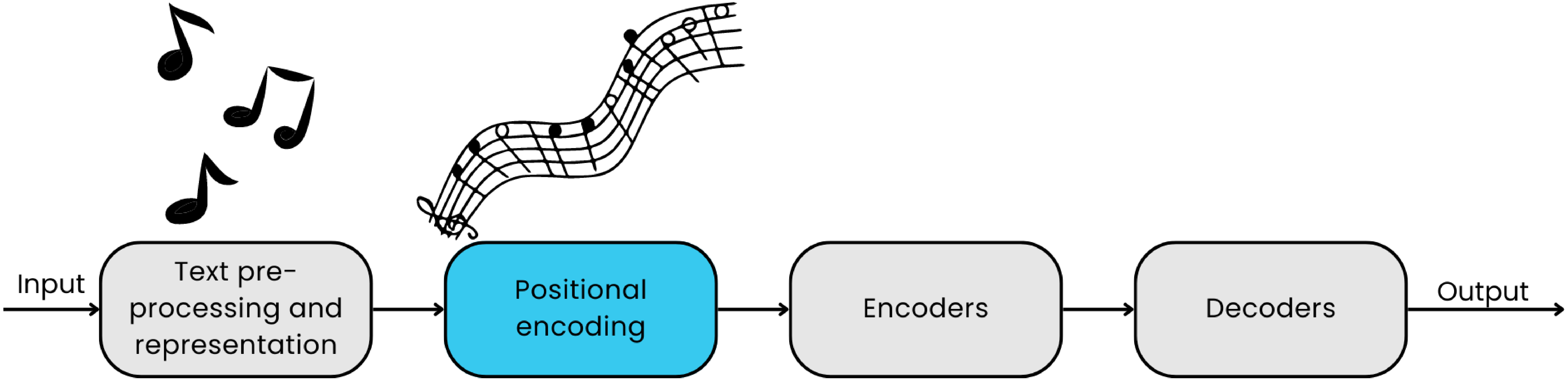
Text pre-processing and representation

- Text preprocessing: tokenization, stop word removal, lemmatization
- Text representation: word embedding



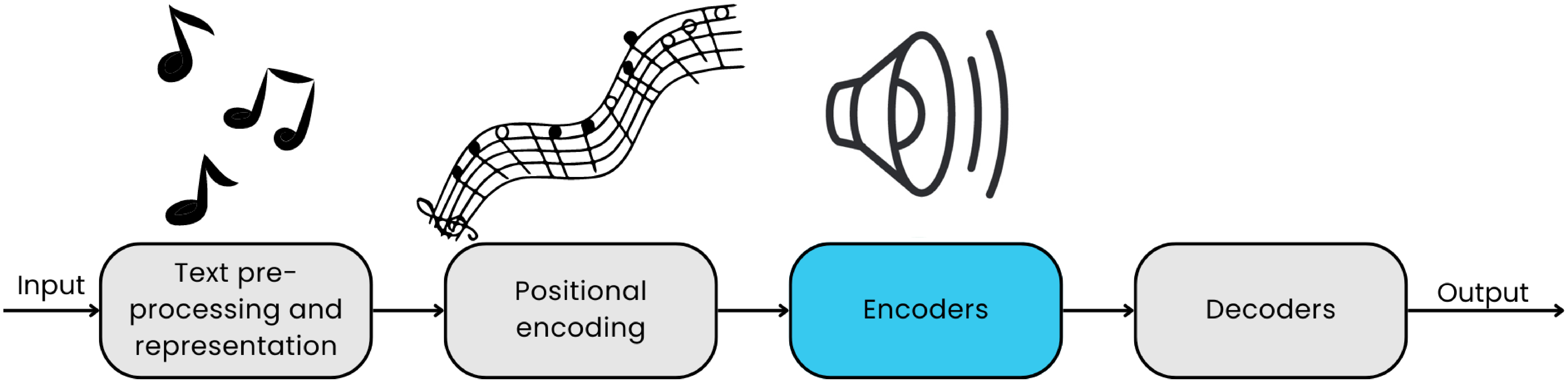
Positional encoding

- Information on the position of each word
- Understand distant words



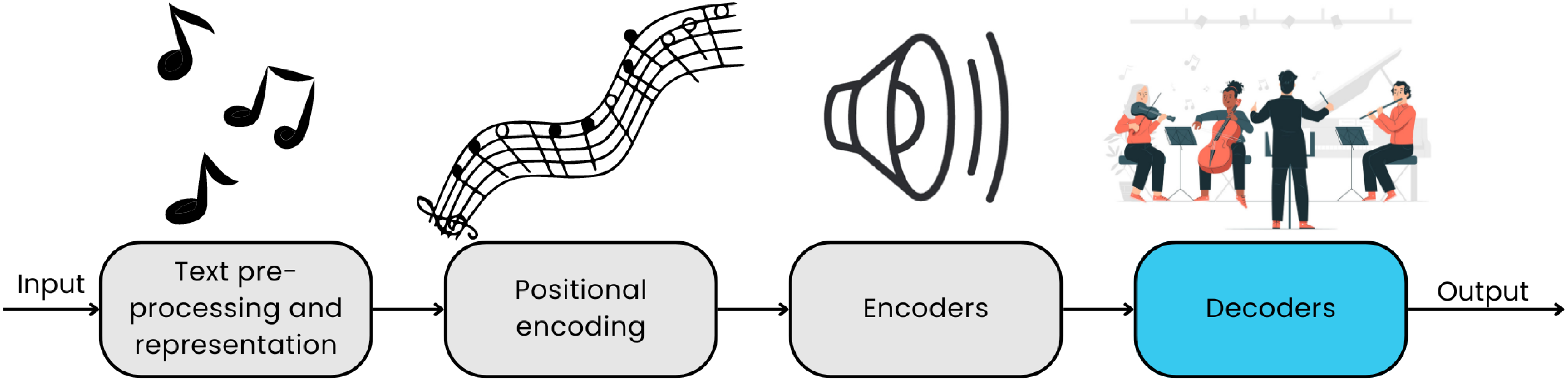
Encoders

- **Attention mechanism:** directs attention to specific words and relationships
- **Neural network:** process specific features



Decoders

- Includes attention and neural networks
- Generates the output



Input

Text pre-processing and representation

Positional encoding

Encoders

Decoders

Output

Transformers and long-range dependencies

- **Initial challenge:** long-range dependency
- **Attention:** focus on different parts of the input

- **Example:** "Jane, who lives in New York and works as a software engineer, loves exploring new restaurants in the city."
- "Jane" --- "loves exploring new restaurants"

Processes multiple parts simultaneously

- **Limitation of traditional language models:**
  - Sequential - one word at a time

- **Transformers:**
  - Process multiple parts simultaneously
  - Faster processing

- **For example:**
  - "The cat sat on the mat"
  - Processes "cat," "sat," "on," "the," and "mat" at the same time

Let's practice!

Attention mechanisms

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

Vidhi Chugh
AI strategist and ethicist

Attention mechanisms

- Understand complex structures
- Focus on important words

- **Book reading analogy:**
  - Clues in a mystery book
  - Focus on relevant content
  - Concentrate on crucial input data

Self-attention and multi-head attention

Self-attention

- Weighs the importance of each word

- Captures long-range dependencies

Multi-head attention

- Next level of self-attention

- Splits input into multiple heads with each head focusing on different aspects

Attention in a party

- **Attention:** Self and multi-head

- **Example:**
  - Group conversation at a party
  - Selective attention to relevant speaker
  - Filter noise
  - Focus on key points

[1] Freepik

Party continues

Self-attention

- Focus on each person's words
- Evaluate and compare their relevance
- Weigh each speaker's input
- Combines for a comprehensive understanding

Multi-head attention

- Split attention into "**multiple**" channels
- Focus on different aspects of conversation
- Speaker's emotions, primary topic, and related side-topics
- Process each aspect and merge

Multi-head attention advantages

- "The boy went to the store to buy some groceries, and he found a discount on his favorite cereal."

- **Attention:** "boy," "store," "groceries," and "discount"
- **Self-attention:** "boy" and "he" -> same person
- **Multi-head attention:** multiple channels
  - Character ("boy")
  - Action ("went to the store," "found a discount")
  - Things involved ("groceries," "cereal")

Let's practice!

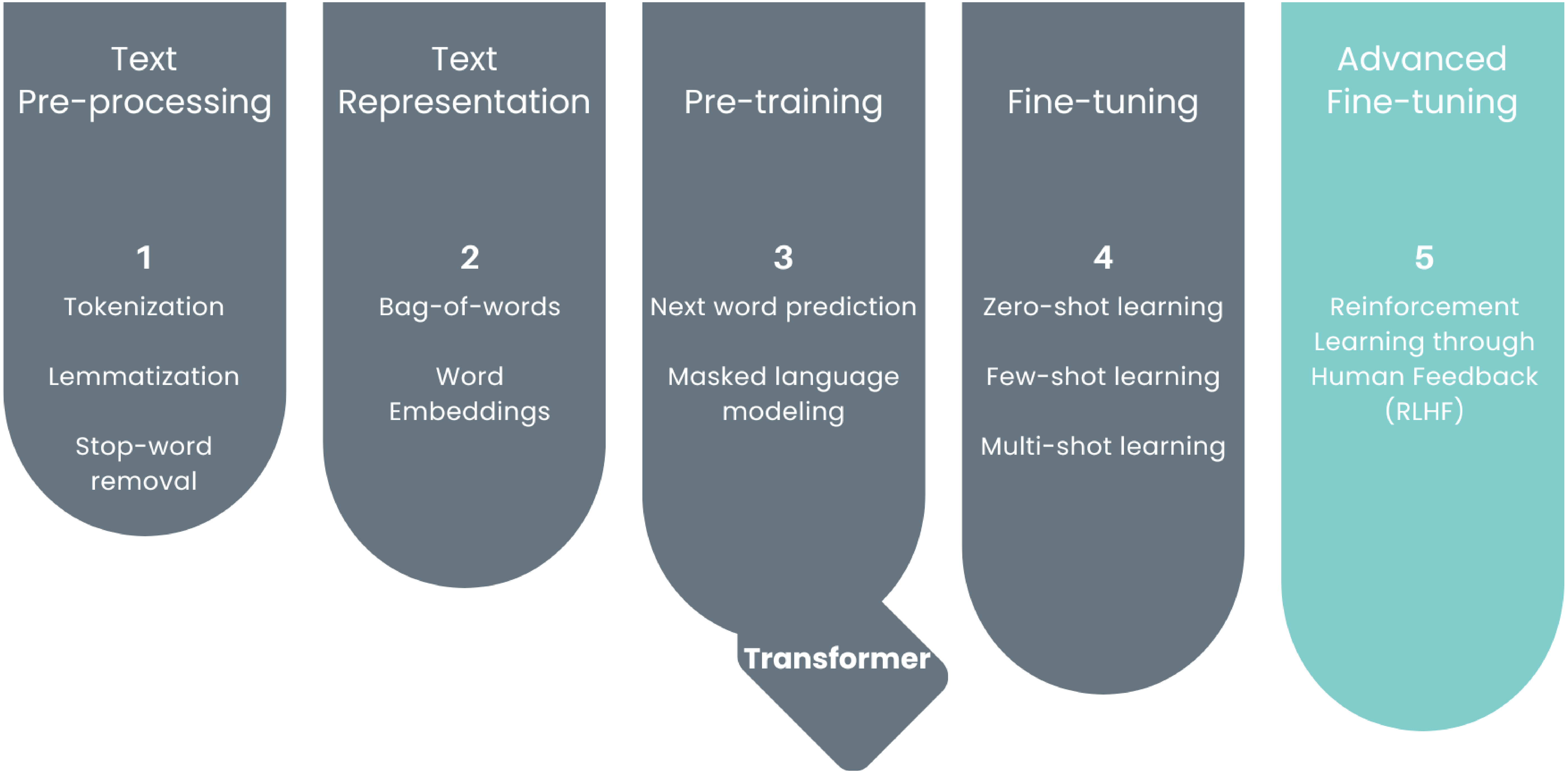
LARGE LANGUAGE MODELS (LLMS) CONCEPTS

Advanced fine-tuning

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

Vidhi Chugh
AI strategist and ethicist

Where are we?



Text Pre-processing

1

Tokenization

Lemmatization

Stop-word removal

Text Representation

2

Bag-of-words

Word Embeddings

Pre-training

3

Next word prediction

Masked language modeling

Transformer

Fine-tuning

4

Zero-shot learning

Few-shot learning

Multi-shot learning

Advanced Fine-tuning

5

Reinforcement Learning through Human Feedback (RLHF)

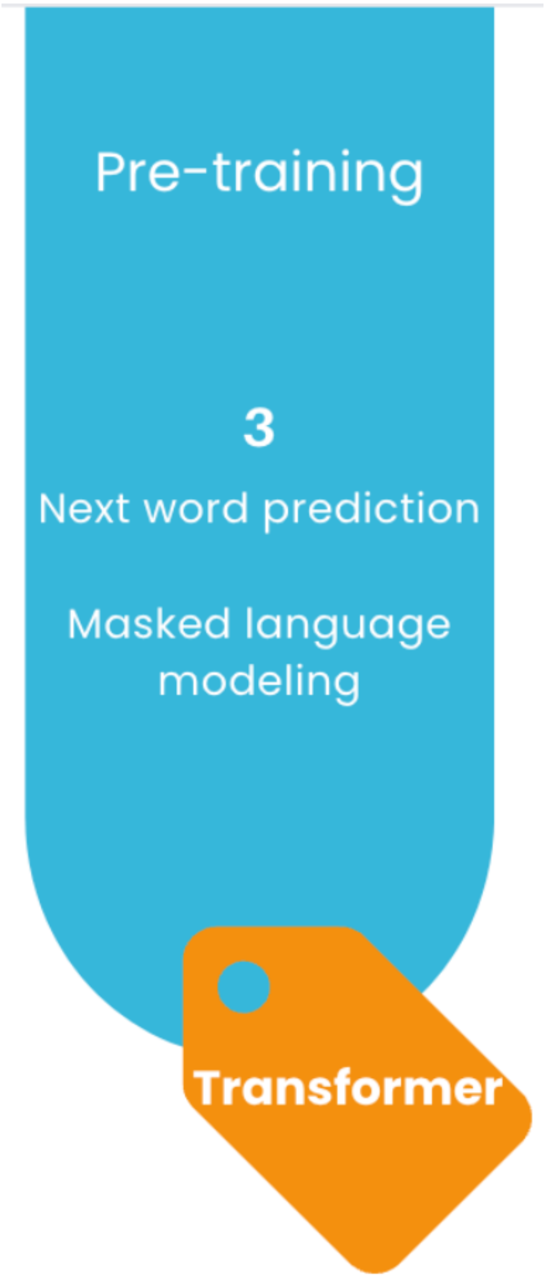
Reinforcement Learning through Human Feedback

- Pre-training
- Fine-tuning
- Reinforcement Learning through Human Feedback (RLHF)

Pre-training

- Large amounts of text data:
  - Websites, books and articles
  - Transformer architecture
  - Learns general language patterns, grammar, and facts

- Next-word prediction
- Masked language modeling

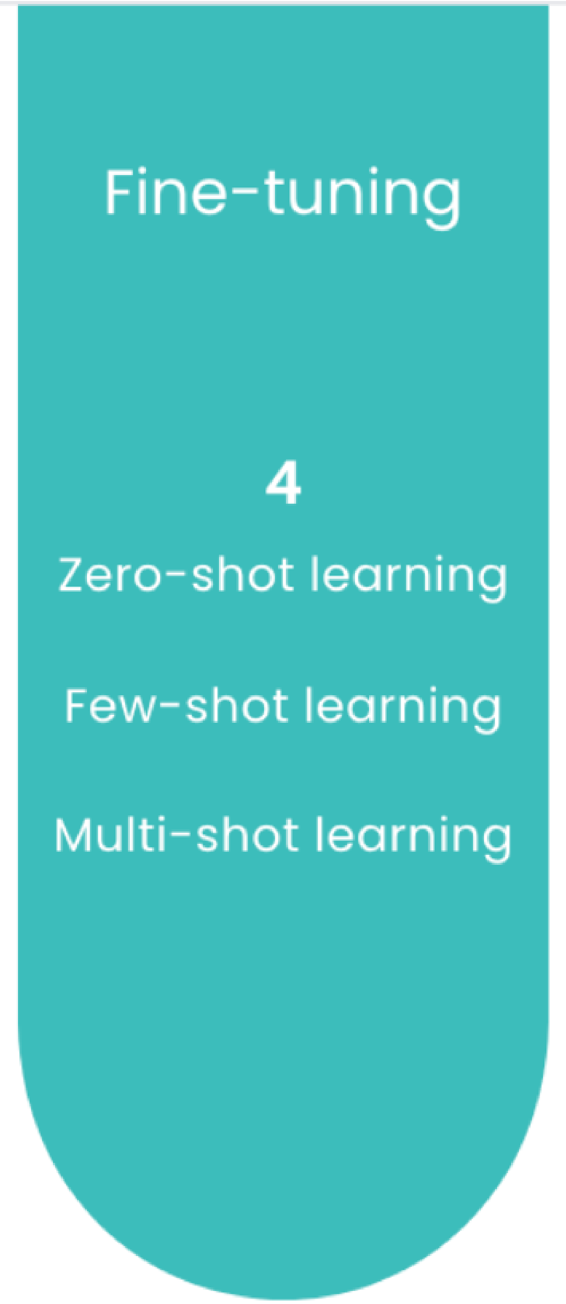


[1] Freepik

Fine-tuning

- N-shot training

- Small labeled dataset for related task



But, why RLHF?

- General-purpose training data lacks quality
  - Noise
  - Errors
  - Inconsistencies
  - Reduced accuracy

Example of reduced accuracy:

- Trained on data from online discussion forums
- Unvalidated opinions and facts
- Needs external expert validation

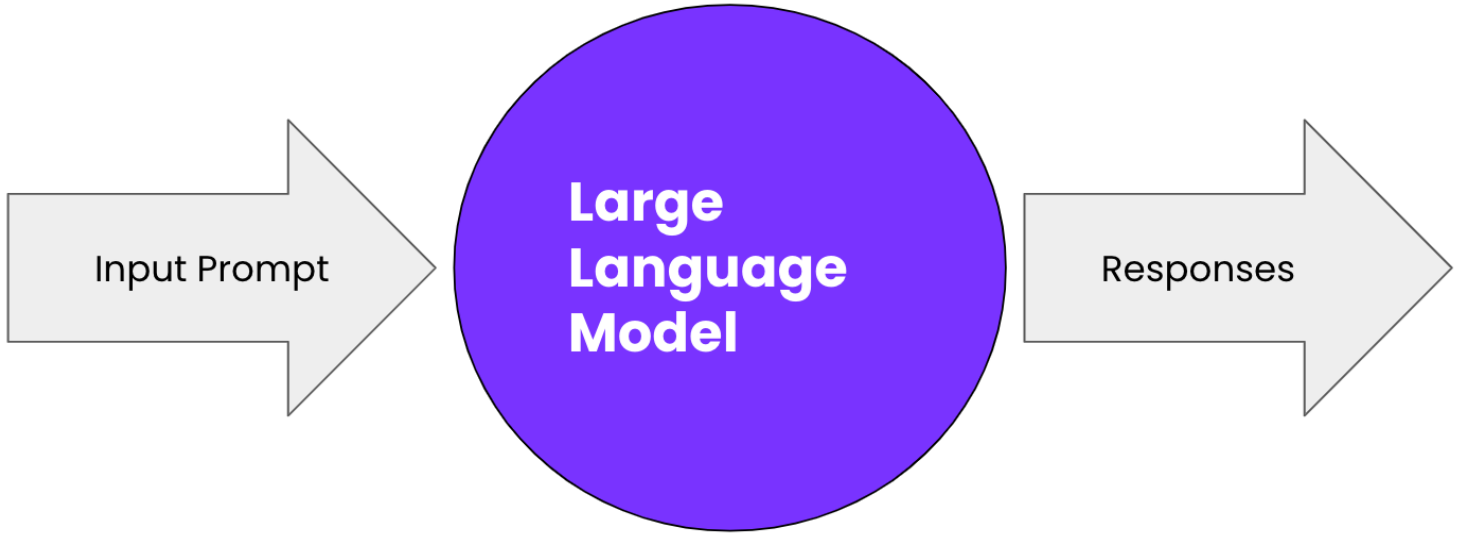
Starts with the need to fine-tune

- Pre-training
    - Learns underlying language patterns
    - Doesn't capture context-specific complexities
- Fine-tuning
    - Quality labeled data improves performance
- **Enter RLHF!**
    - Human feedback

Simplifying RLHF

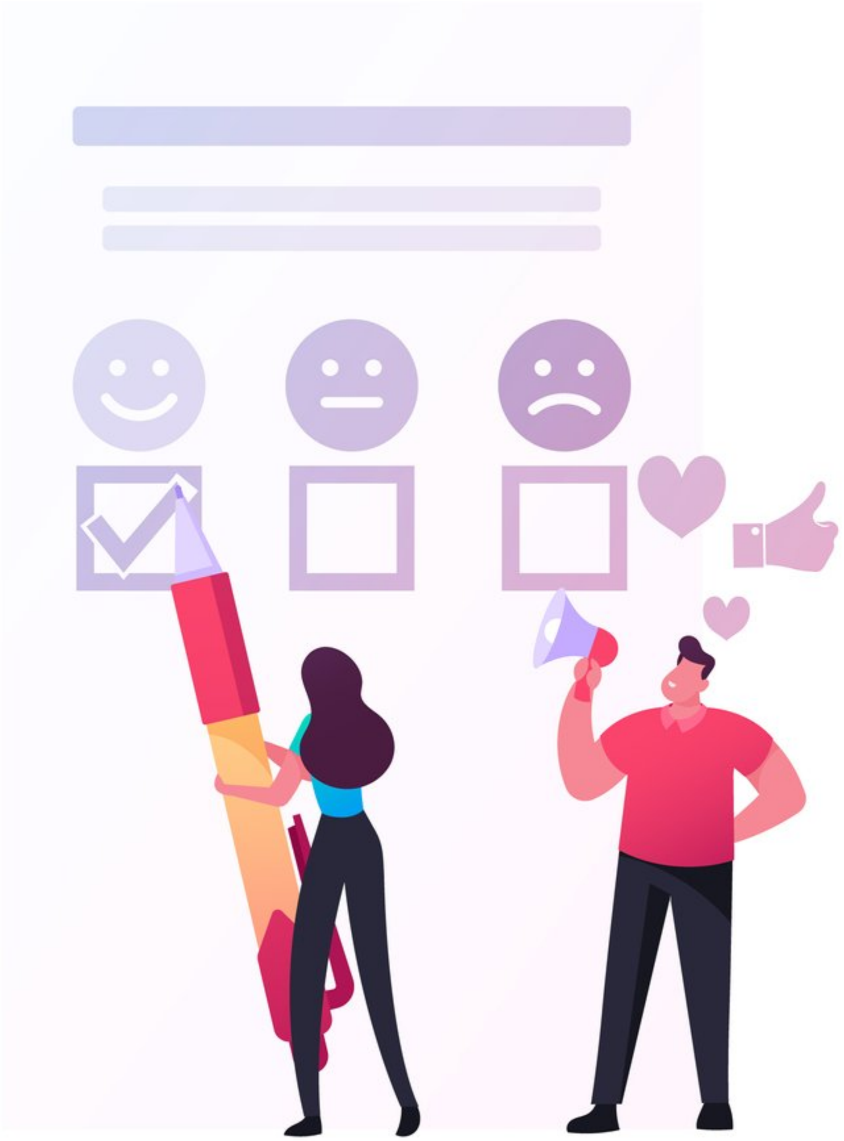
- Model output reviewed by human
- Updates model based on the feedback

- Step 1:
  - Receives a prompt
  - Generates multiple responses



Enters human expert

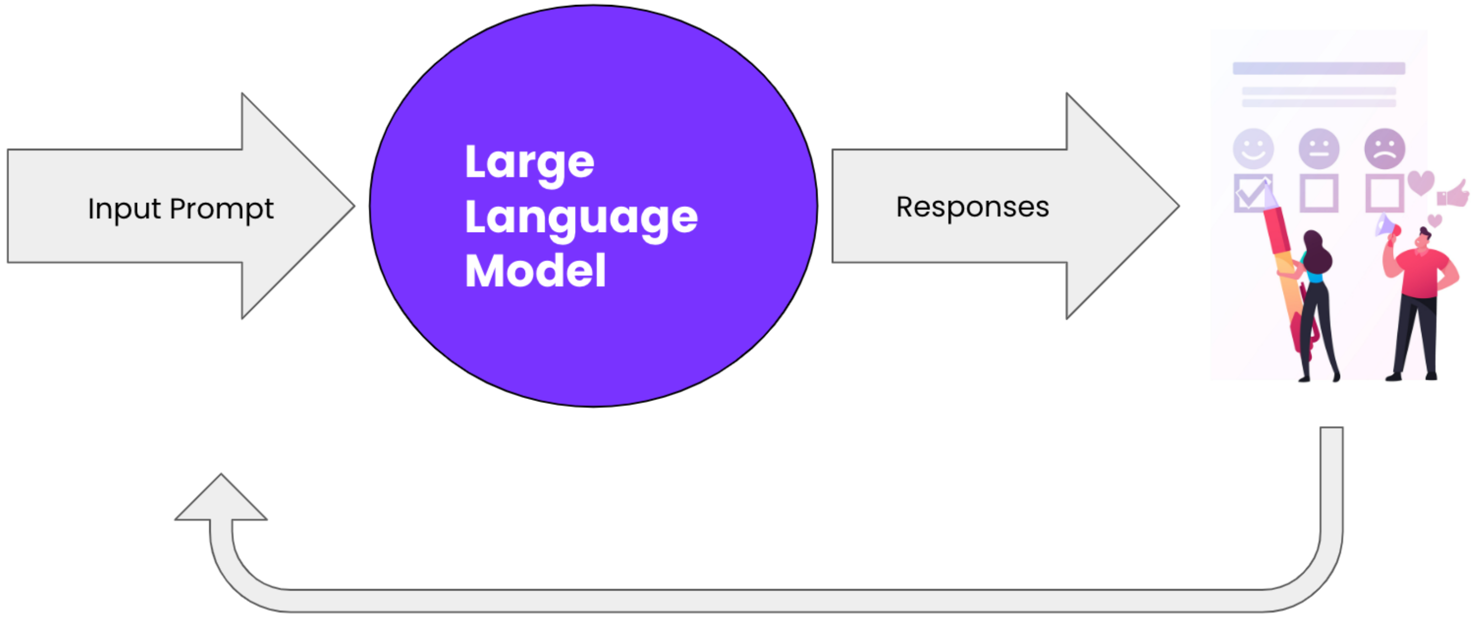
- Step 2:
  - Human expert checks these responses
  - Ranks the responses based on quality
    - Accuracy
    - Relevance
    - Coherence



Time for feedback

- Step 3:
  - Learns from expert's ranking
  - To align its response in future with their preferences

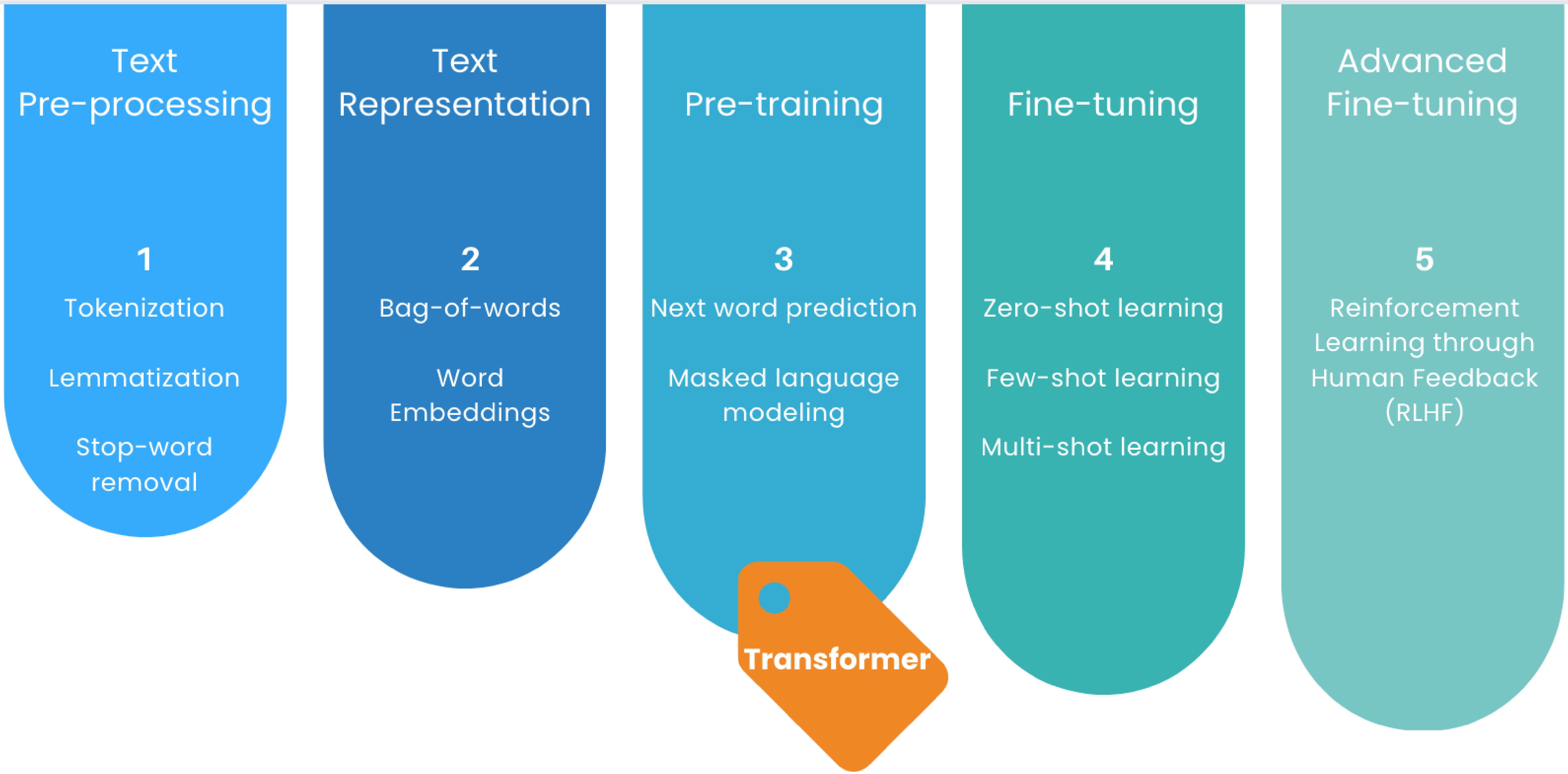
- And it goes on!
  - Continues to generate responses
  - Receives expert's rankings
  - Adjusts the learning



Recap

- **Pre-training** to learn general language knowledge
- **Fine-tuning** for specific tasks
- **RLHF** techniques to enhance fine-tuning through human feedback
- Combination is highly effective!

Completing the LLM



Text Pre-processing

1

Tokenization

Lemmatization

Stop-word removal

Text Representation

2

Bag-of-words

Word Embeddings

Pre-training

3

Next word prediction

Masked language modeling

Transformer

Fine-tuning

4

Zero-shot learning

Few-shot learning

Multi-shot learning

Advanced Fine-tuning

5

Reinforcement Learning through Human Feedback (RLHF)

Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS