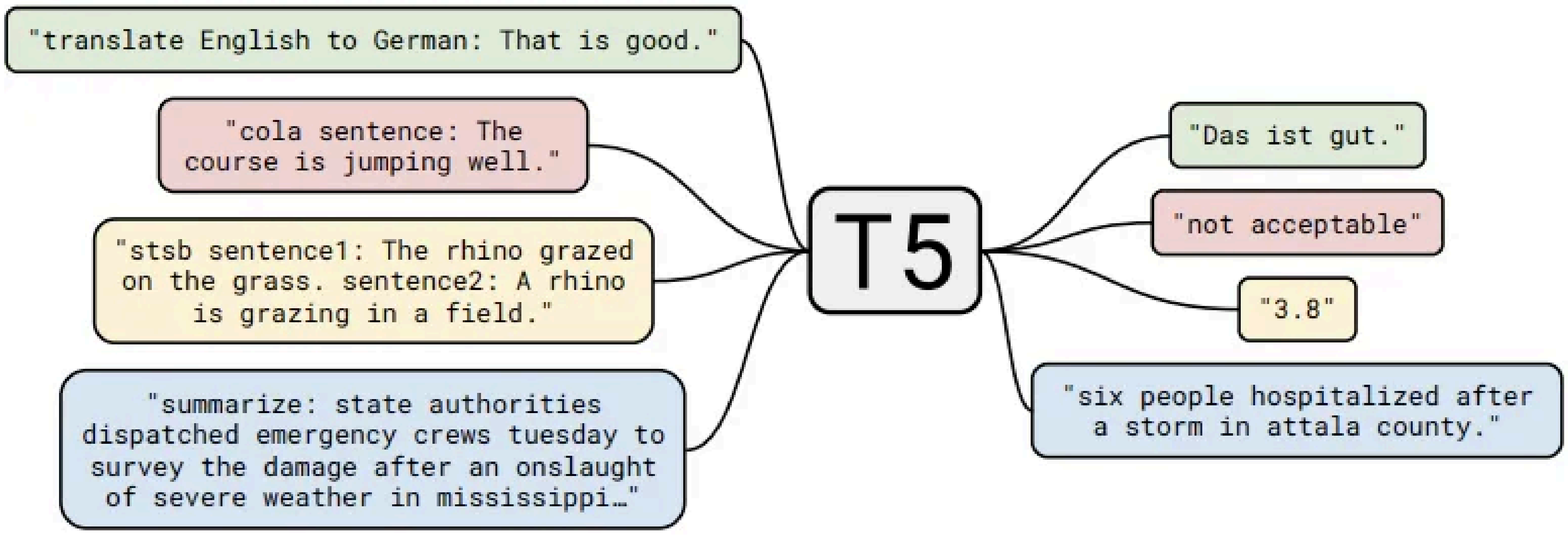
Mastering LLMs

Day 8:Text-to-Text Transfer Transformer



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

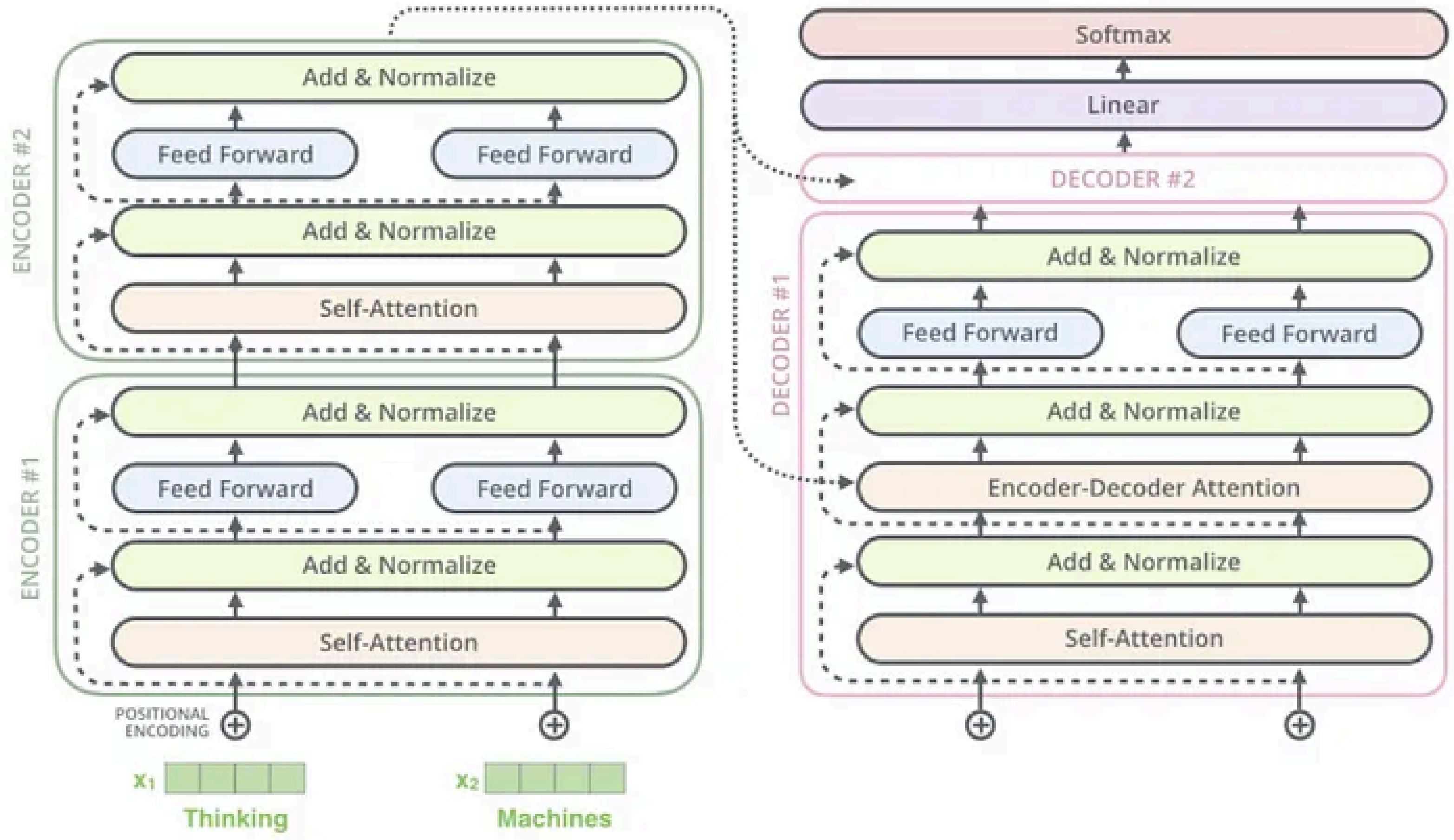
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

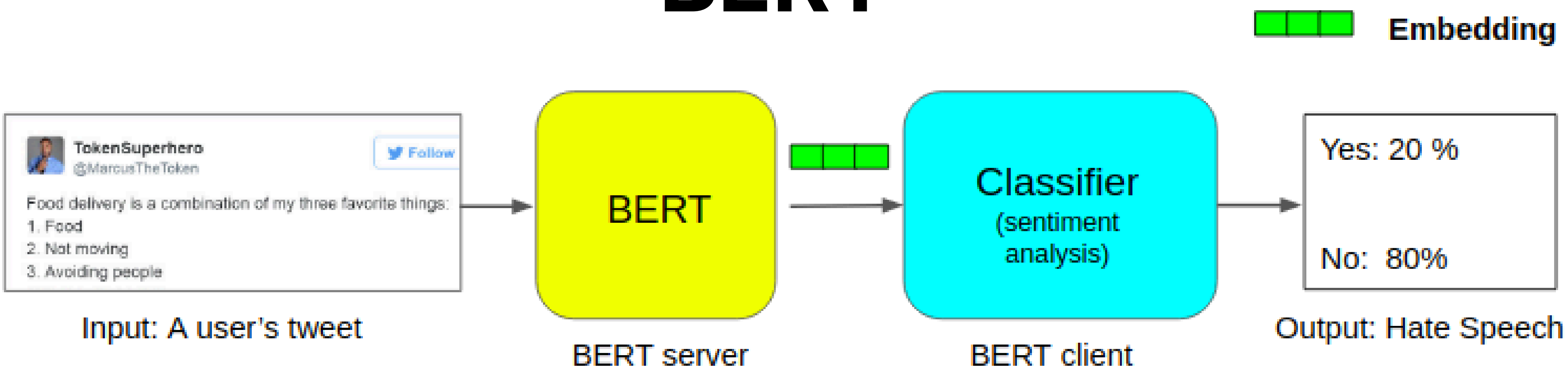


T5

Model Architecture

Yesterday, we talked about **GPT**. Today we will cover a Text-to-Text Transfer Transformer

BERT



🟩🟩🟩 **Embedding**

TokenSuperhero
@MarcusTheToken
Follow

Food delivery is a combination of my three favorite things:
1. Food
2. Not moving
3. Avoiding people

Input: A user's tweet

BERT

BERT server

Classifier
(sentiment analysis)

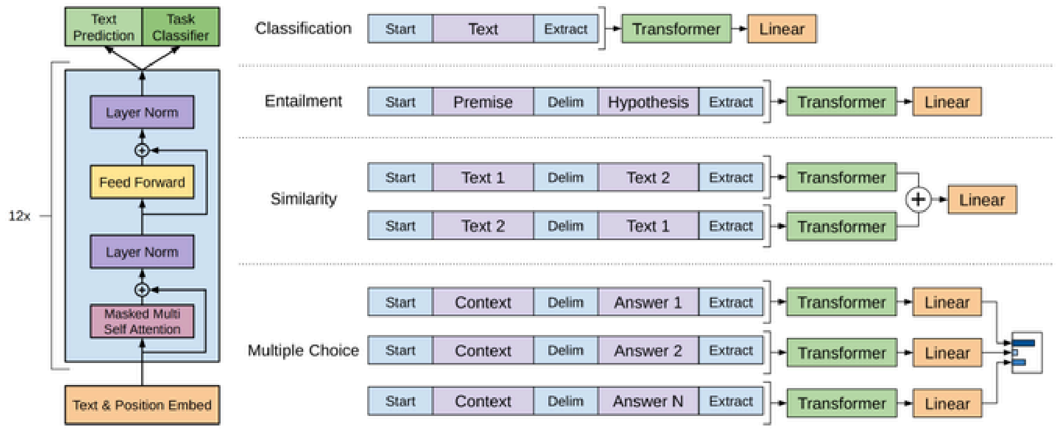
BERT client

Yes: 20 %

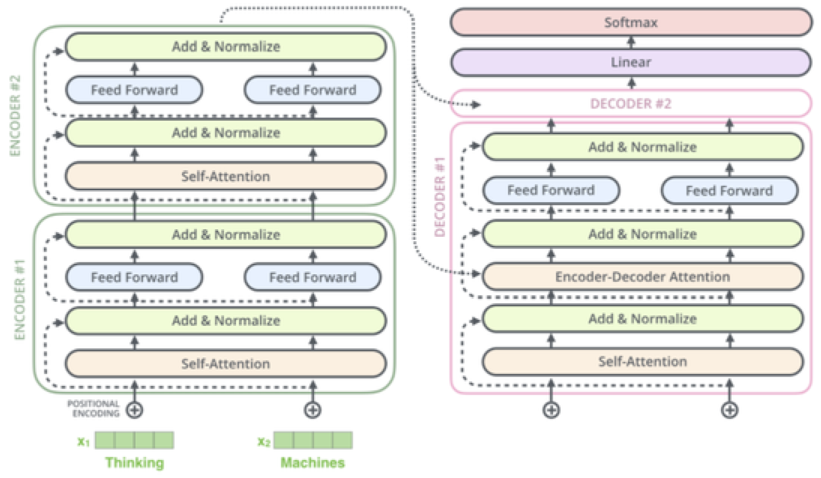
No: 80%

Output: Hate Speech

GPT



T5



ModernBERT

Answer.AI + LightOn +

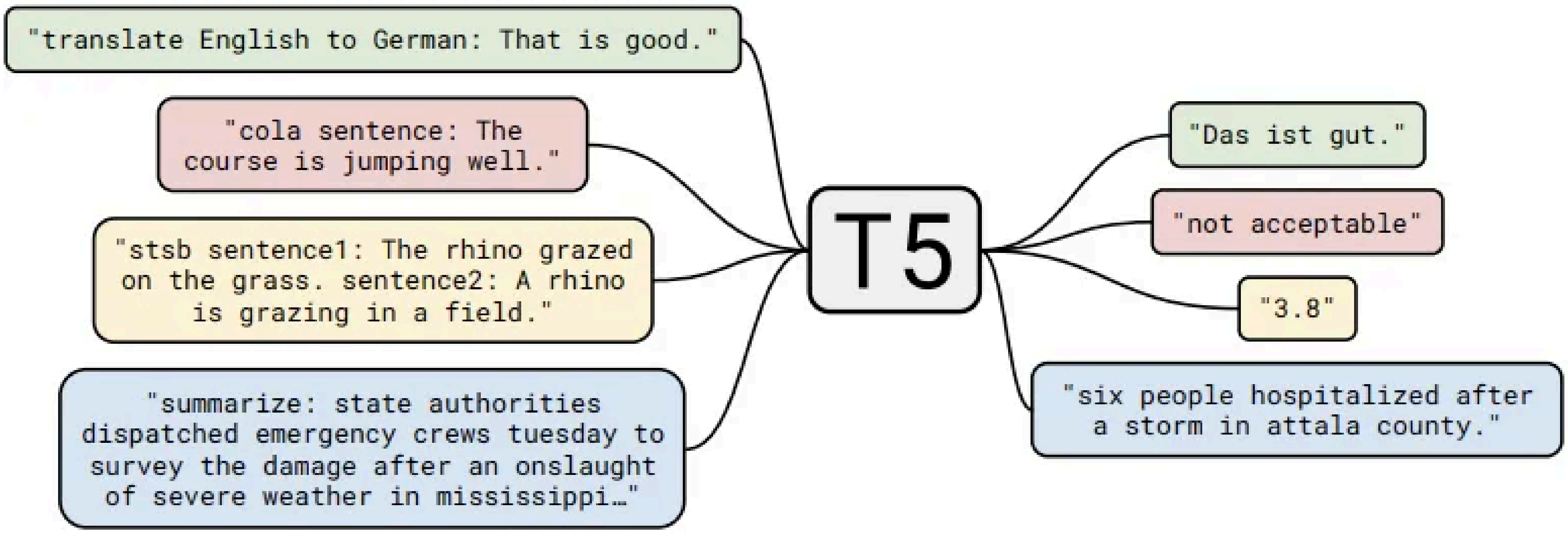
Let's dive straight into

T5 ➡

What is T5?

- T5, short for **Text-to-Text Transfer Transformer**, is a language model developed by **Google Research**. Introduced in 2019, T5 treats every NLP (Natural Language Processing) problem as a text-to-text problem.

- This means that all inputs and outputs are represented as text strings, making it a unified framework to handle a wide range of NLP tasks such as translation, summarization, classification, and question answering.


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Why is T5 Used?

- **Unified Framework**: By framing all tasks in a text-to-text paradigm, T5 eliminates the need for task-specific architectures, making it versatile and flexible.

- **Transfer Learning**: It leverages transfer learning, where the model is pre-trained on massive datasets and then fine-tuned on specific tasks, ensuring high performance.

- **State-of-the-Art Performance**: T5 achieves competitive or superior results across various NLP benchmarks, including the GLUE, SuperGLUE, and SQuAD datasets.

- **Ease of Use**: Representing inputs and outputs as text simplifies pipeline integration and reduces the complexity of adapting models for different tasks.

Analytics Vidhya

Is T5 Better than BERT and GPT?

Compared to **BERT**:

- **Versatility**: T5 is more versatile as it handles both generative (e.g., summarization) and discriminative tasks (e.g., classification), while BERT primarily excels at discriminative tasks.
- **Output Structure**: BERT focuses on token classification or sequence classification tasks, while T5 generates textual outputs, making it suitable for a broader range of tasks.
- **Performance**: T5 often outperforms BERT on generative and sequence-to-sequence tasks, although BERT may still be more efficient for simpler classification problems.

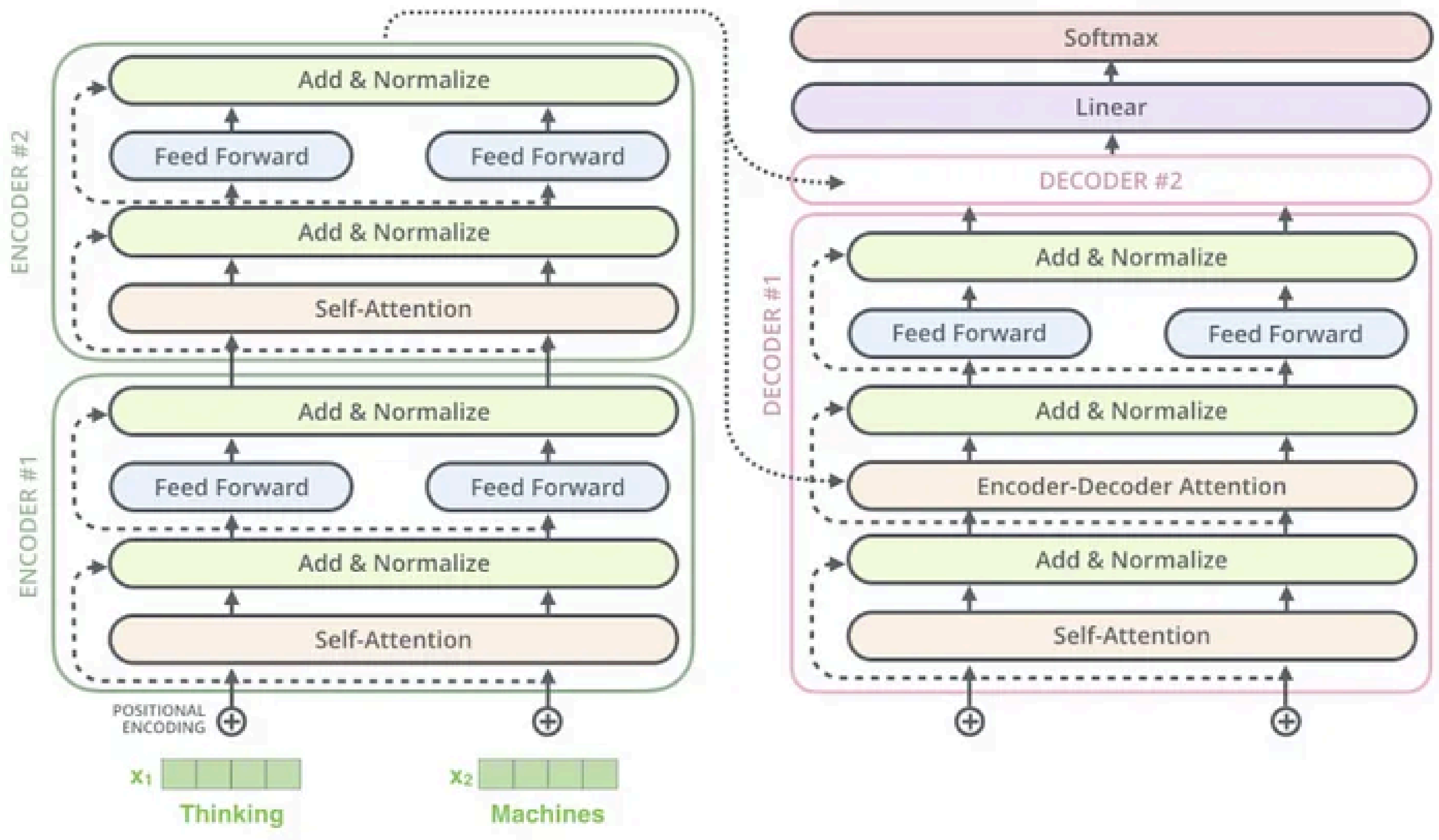
Compared to **GPT**:

- **Task Framing**: While GPT (e.g., GPT-3) focuses on autoregressive generation, T5 frames problems within the text-to-text paradigm, enabling better alignment for structured tasks like translation or summarization.

- **Pretraining Objectives**: T5 uses a Span-Corruption pretraining objective (a variant of masked language modeling) compared to GPT's left-to-right autoregressive modeling. This leads to better bidirectional context understanding in T5.

- **Efficiency**: T5 models are more parameter-efficient for specific tasks, although GPT models may still shine in few-shot or zero-shot scenarios due to their scale and pretraining strategy.

The Architecture of T5



Transformer-Based

- T5 is built on the Transformer architecture, consisting of an encoder-decoder structure:
    - **Encoder**: Processes the input text and generates hidden representations.
    - **Decoder**: Generates the output text based on encoder representations and previously generated tokens.

Key Features

- **Layer Normalization**: Applied before each sub-layer (Pre-LN Transformer).
- **Relative Positional Encoding**: Improves efficiency in handling long sequences.
- **Efficient Feedforward Layers**: Standard Transformer feedforward layers are used for scalability.

Variants

- T5 comes in multiple sizes to balance performance and computational cost:
  - Small, Base, Large, 3B, and 11B (11 billion parameters being the largest variant).

Training Dataset

- T5 is pretrained on the Colossal Clean Crawled Corpus (C4), a large and diverse dataset curated from the web.

Training Objective

- Span-corruption objective ensures the model learns bidirectional context and remains efficient for both generative and discriminative tasks.

Analytics Vidhya

Stay Tuned for **Day 9** of

Mastering LLMs