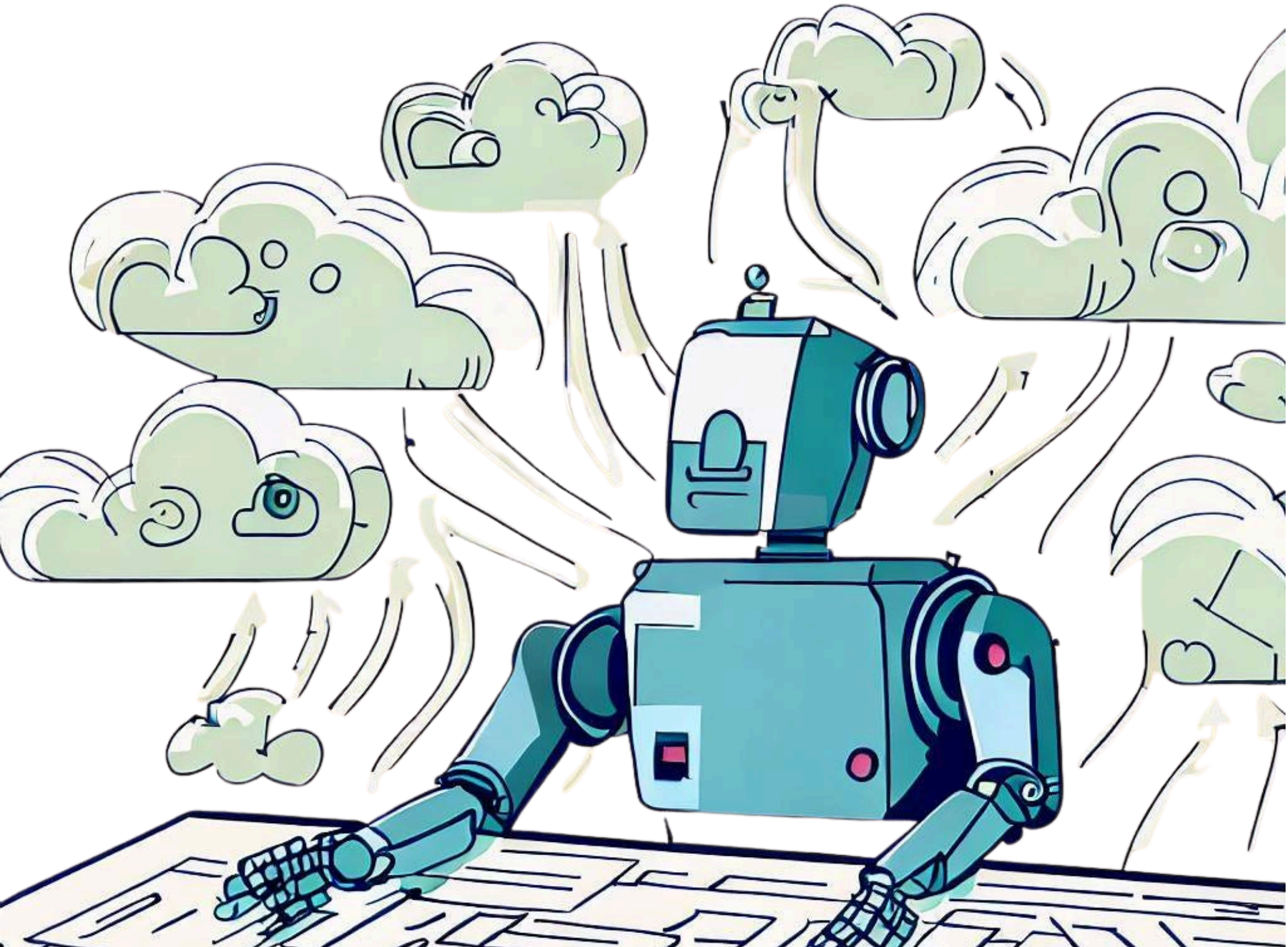
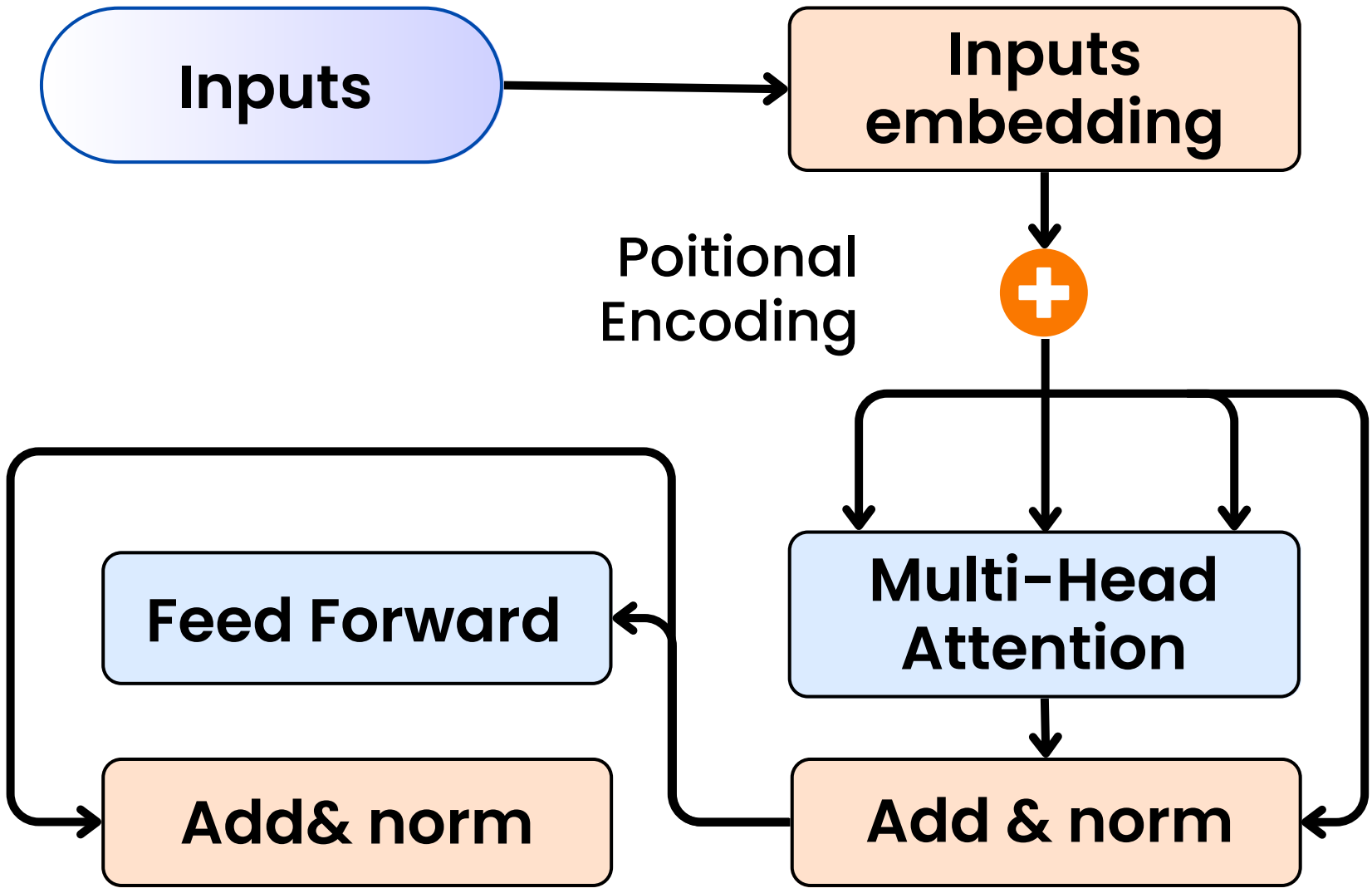
Greg Coquillo
Product Leader

6 CORE LLM ARCHITECTURES EXPLAINED



Greg Coquillo
Product Leader

ENCODER ONLY



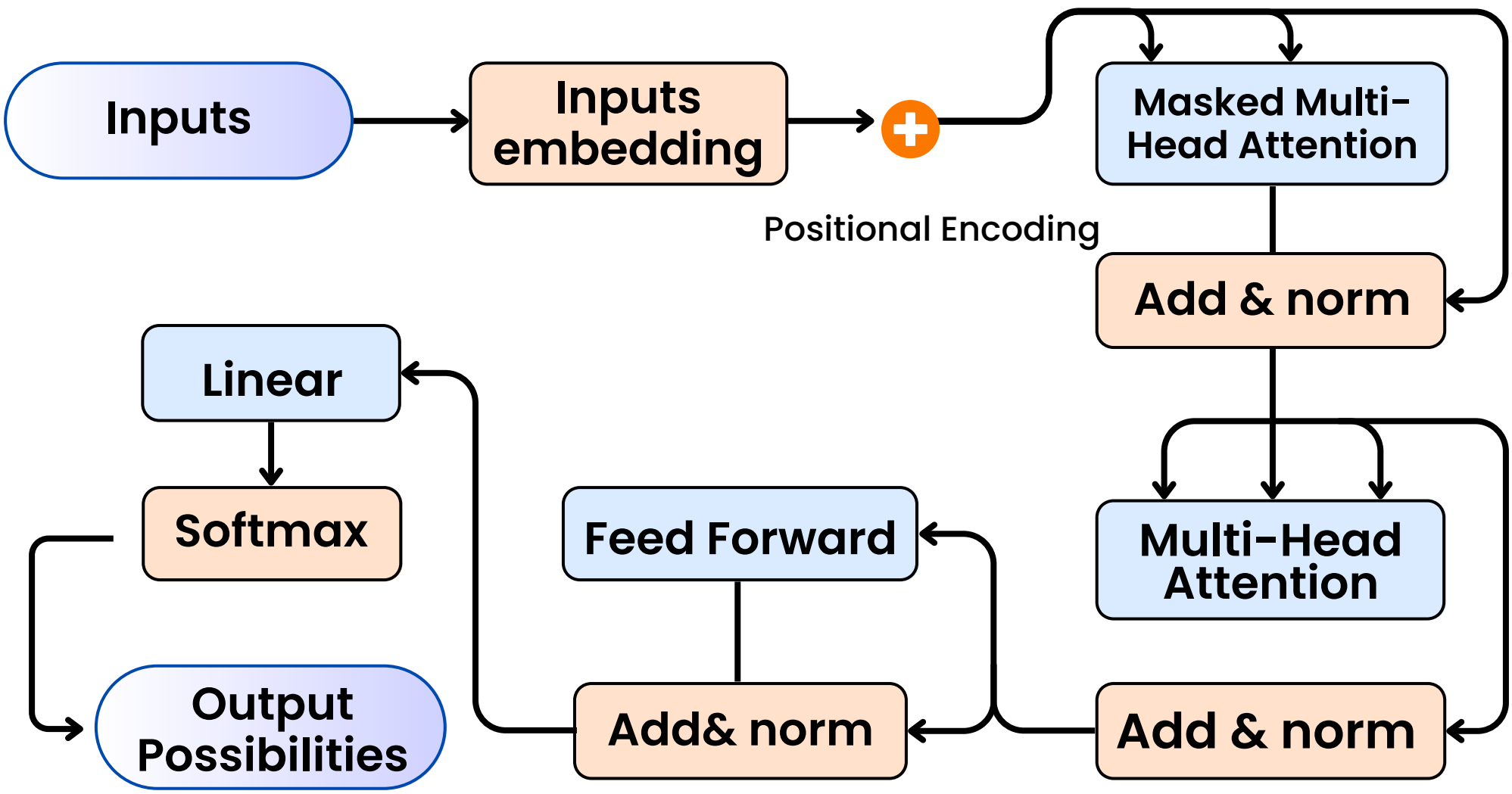
Description:

This architecture is designed to understand and encode input text into rich representations. It processes the entire input sequence at once, using self-attention to capture relationships between all tokens. Commonly used for tasks like classification, sentence similarity, and embeddings.

Example: BERT, RoBERTa

Greg Coquillo
Product Leader

DECODER ONLY



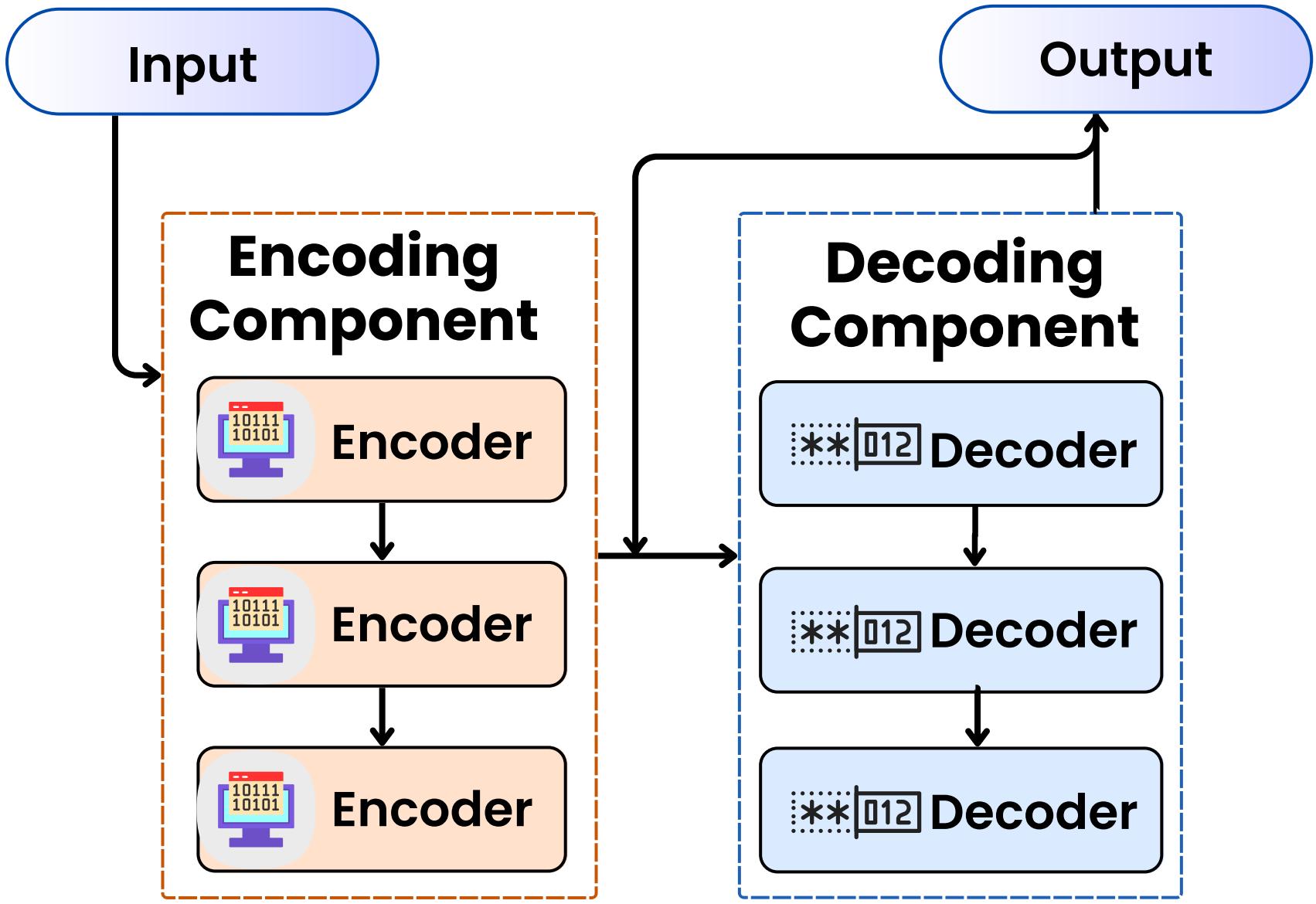
Description:

Primarily used for generative tasks, this model reads input from left to right and predicts the next token in sequence. It uses masked self-attention so each token can only attend to previous tokens. Ideal for text generation, code completion, or dialogue agents.

Example: GPT-2, GPT-3, GPT-4

Greg Coquillo
Product Leader

ENCODER DECODER



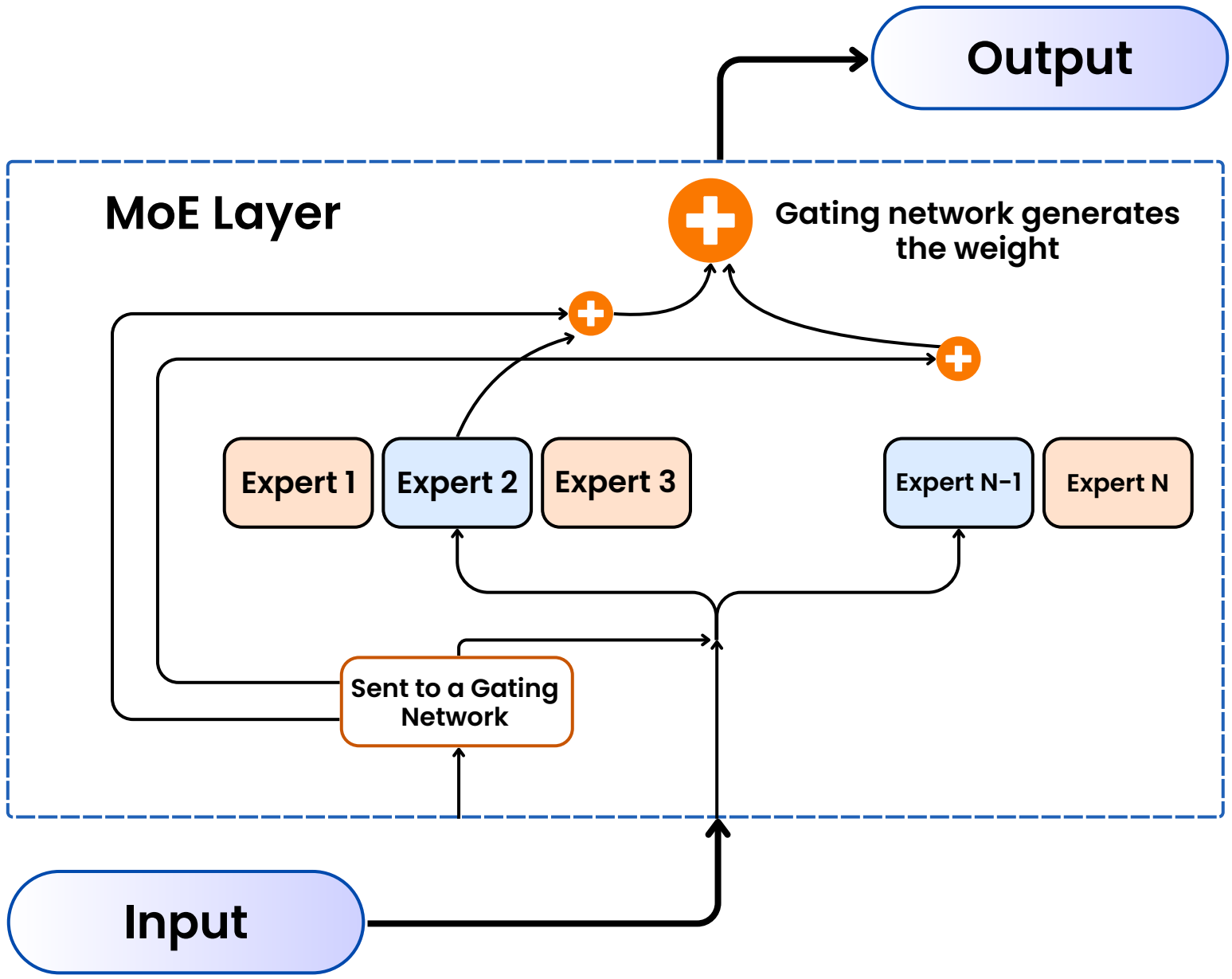
Description:

Also called seq2seq (sequence-to-sequence), this architecture has two parts: the encoder processes input and creates a context representation, while the decoder uses that context to generate output. Ideal for translation, summarization, and question-answering.

Example: T5, BART, mT5

Greg Coquillo
Product Leader

MIXTURE OF EXPERTS



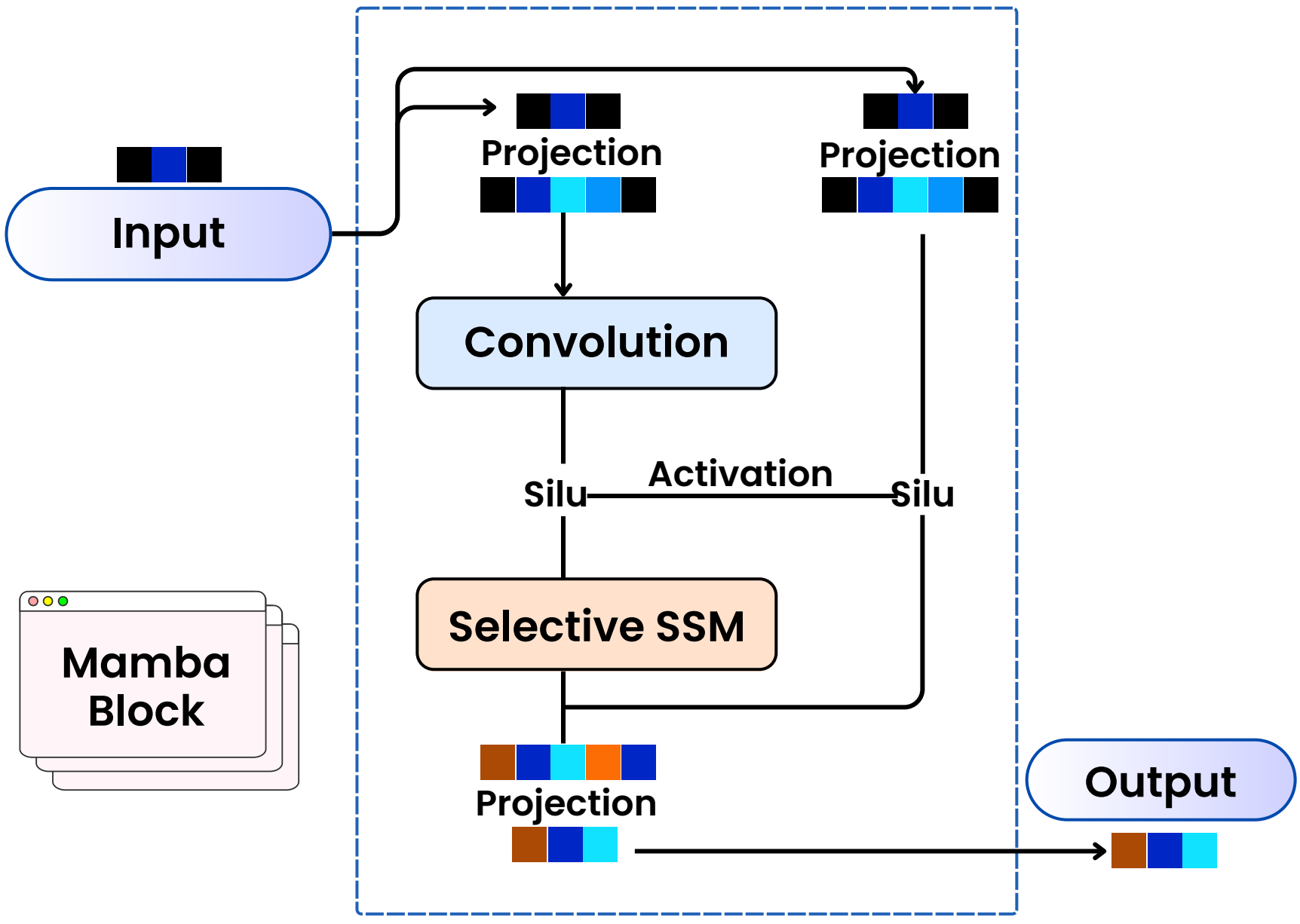
Description:

In MoE models, only a subset of “expert” neural networks are activated per input. A gating network decides which experts to use. This allows scaling to billions of parameters while keeping computation efficient. Great for training massive models with less cost.

Example: GLaM, Switch Transformer, Mixtral


Greg Coquillo
Product Leader

STATE SPACE MODEL



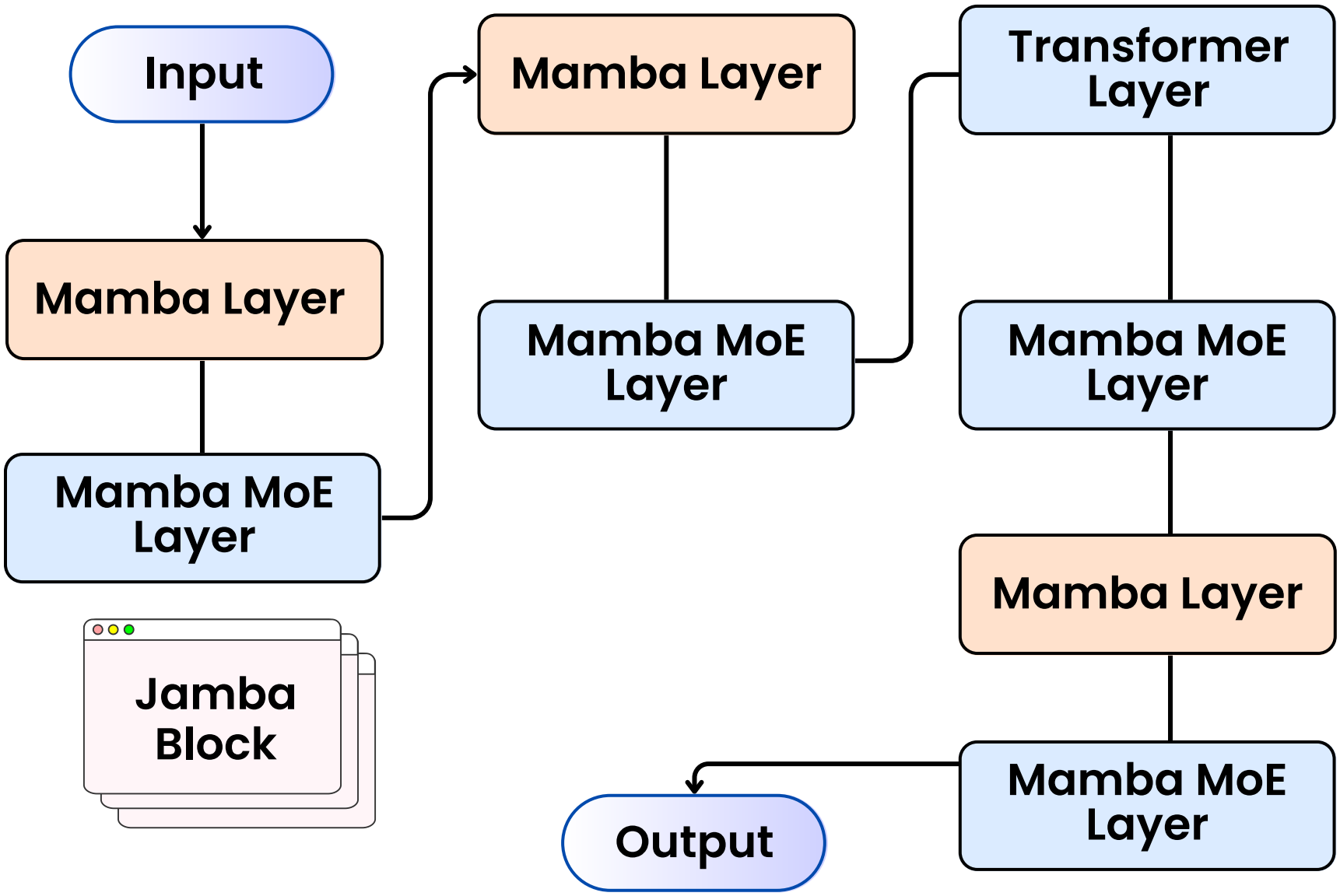
Description:

Unlike transformers that rely on attention, SSMs use mathematical operations based on state space equations to model long-range dependencies efficiently. They're great for long-context reasoning and low-latency inference.

Example: Mamba

Greg Coquillo
Product Leader

HYBRID



Description:

This combines elements from multiple architectures, like transformers, MoE, and SSMs to balance accuracy, efficiency, and scalability. These models aim to get the best of all worlds, adapting to different tasks within a single framework.

Example: Jamba (uses Mamba + Transformer + MoE)

Greg Coquillo

Product Leader

FOLLOW FOR MORE