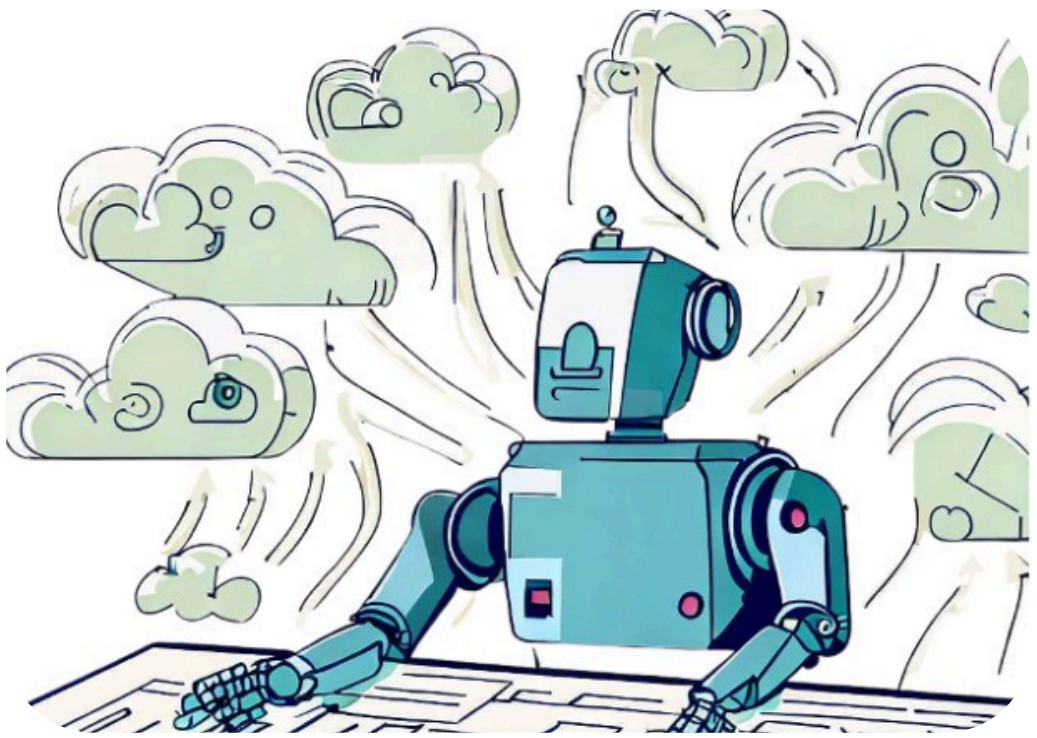
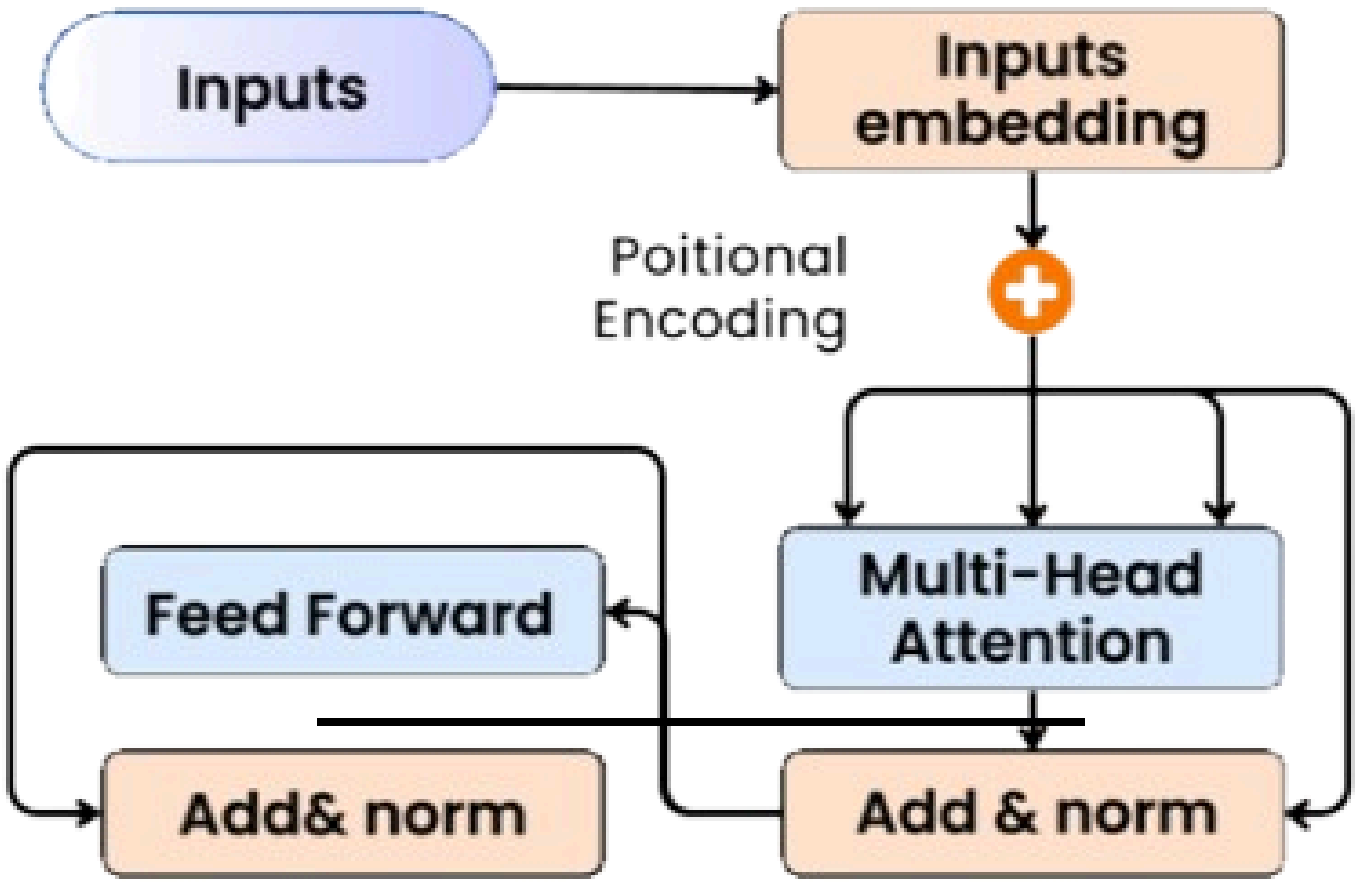
DroneX AI



“6 CORE LLM ARCHITECTURES EXPLAINED”

Karn Singh

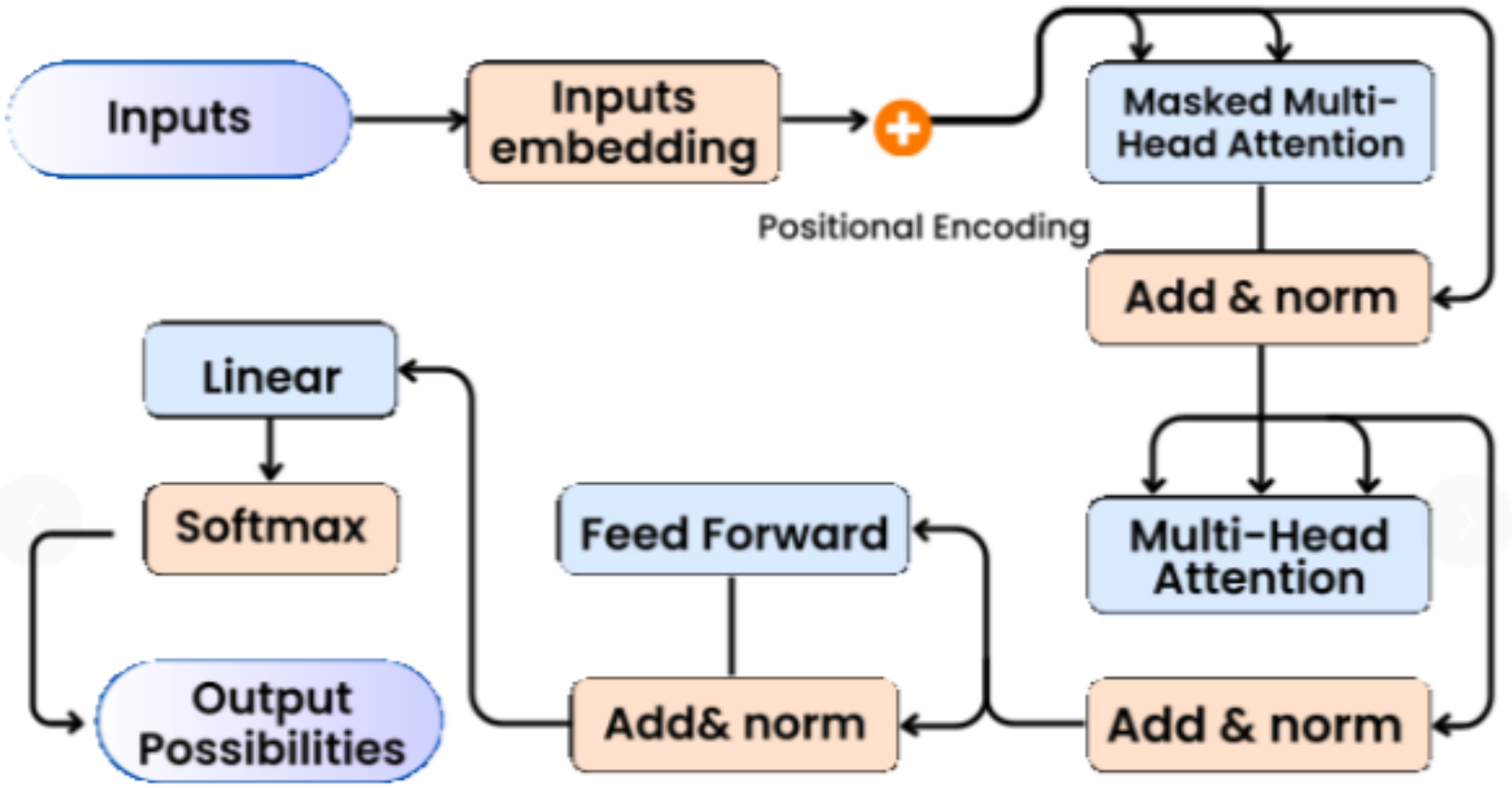
ENCODER-ONLY



This model architecture focuses solely on encoding input data by capturing contextual relationships across the entire sequence using self-attention. It produces rich semantic representations ideal for downstream tasks such as sentence classification, semantic similarity, and vector embeddings.

Popular Models: BERT, RoBERTa

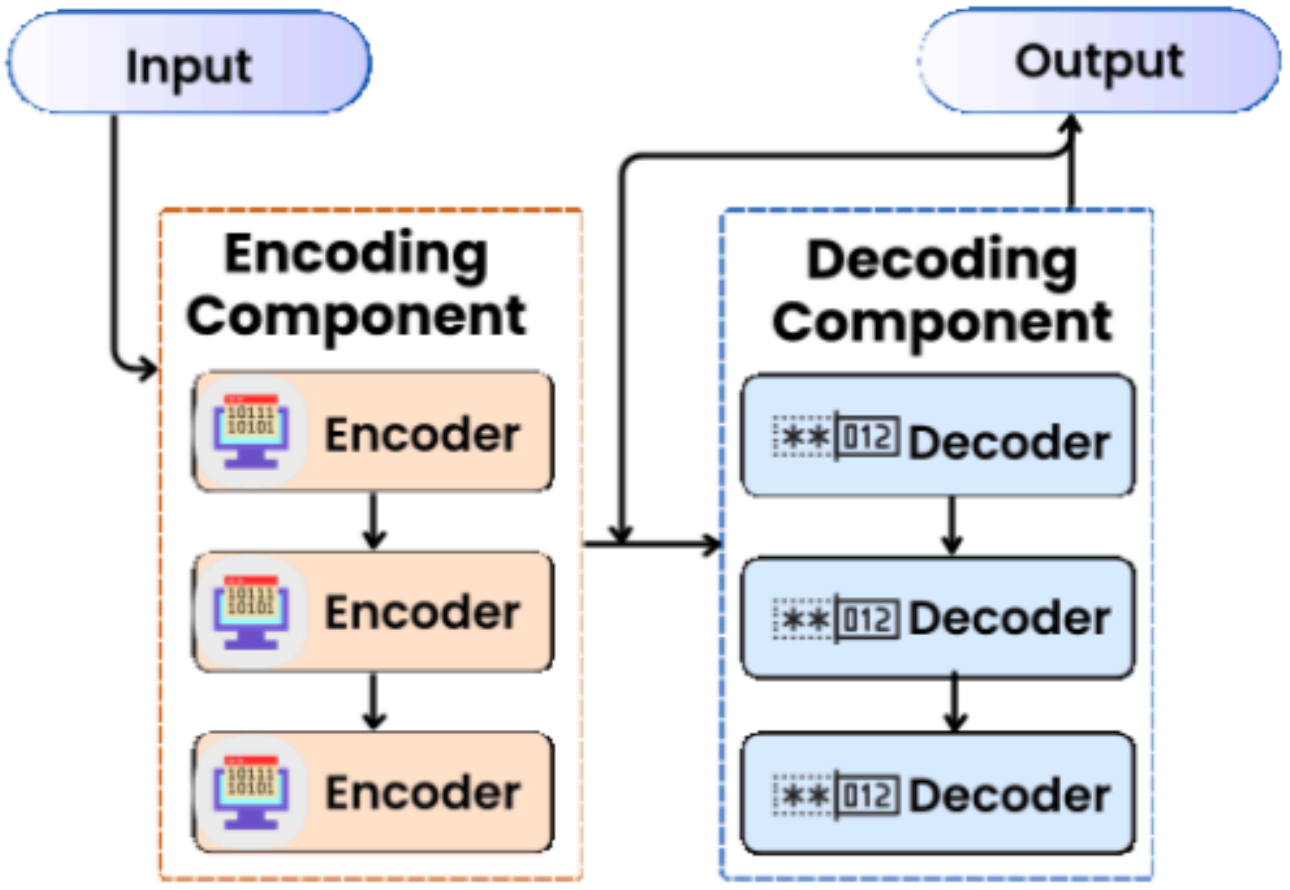
DECODER-ONLY



This architecture is designed for generative tasks, processing input sequentially from left to right to predict the next token in the series. By using masked self-attention, each token can only consider earlier tokens, ensuring correct autoregressive generation. This makes it well-suited for applications like creative text generation, code synthesis, or conversational AI systems.

Examples: GPT-2, GPT-3, GPT-4

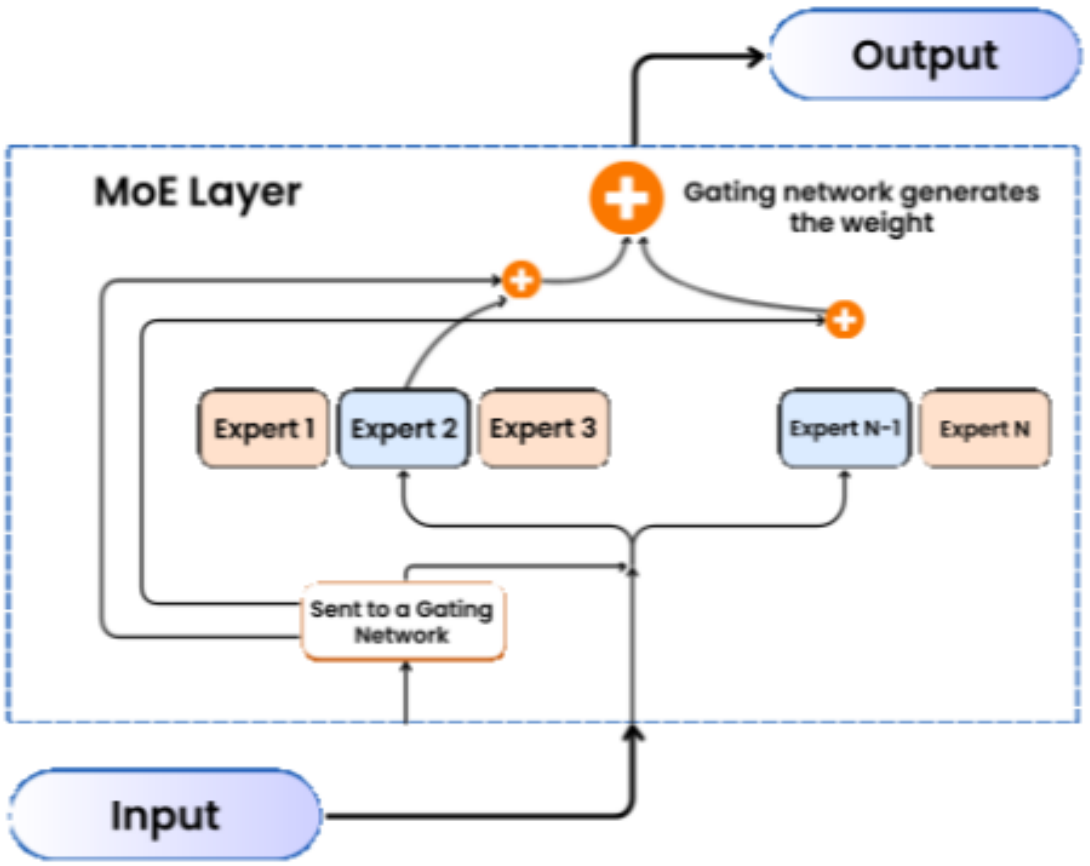
ENCODER DECODER



Known as sequence-to-sequence (seq2seq), this architecture consists of two main components: the encoder, which transforms the input into a meaningful intermediate representation, and the decoder, which generates the desired output using this context. This setup is highly effective for tasks like language translation, text summarization, and answering questions.

Examples: T5, BART, mT5

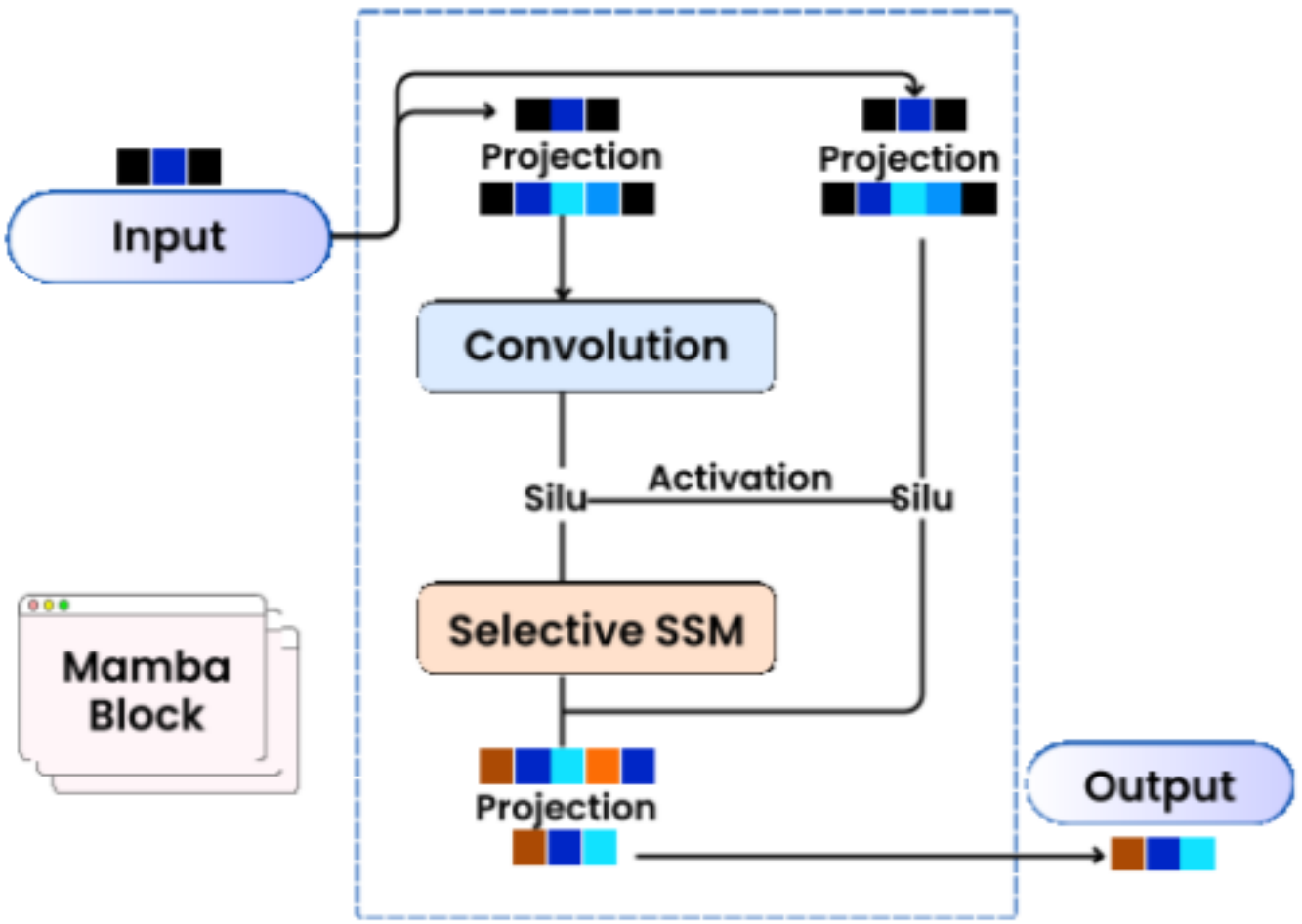
MIXTURE OF EXPERTS



In Mixture of Experts (MoE) architectures, only a small group of specialized “expert” neural networks is activated for each input, based on the decision of a gating network. This approach enables the model to scale to billions of parameters without proportionally increasing computational cost, making it highly efficient for training large-scale models with reduced resource requirements.

Examples: GLaM, Switch Transformer, Mixtral

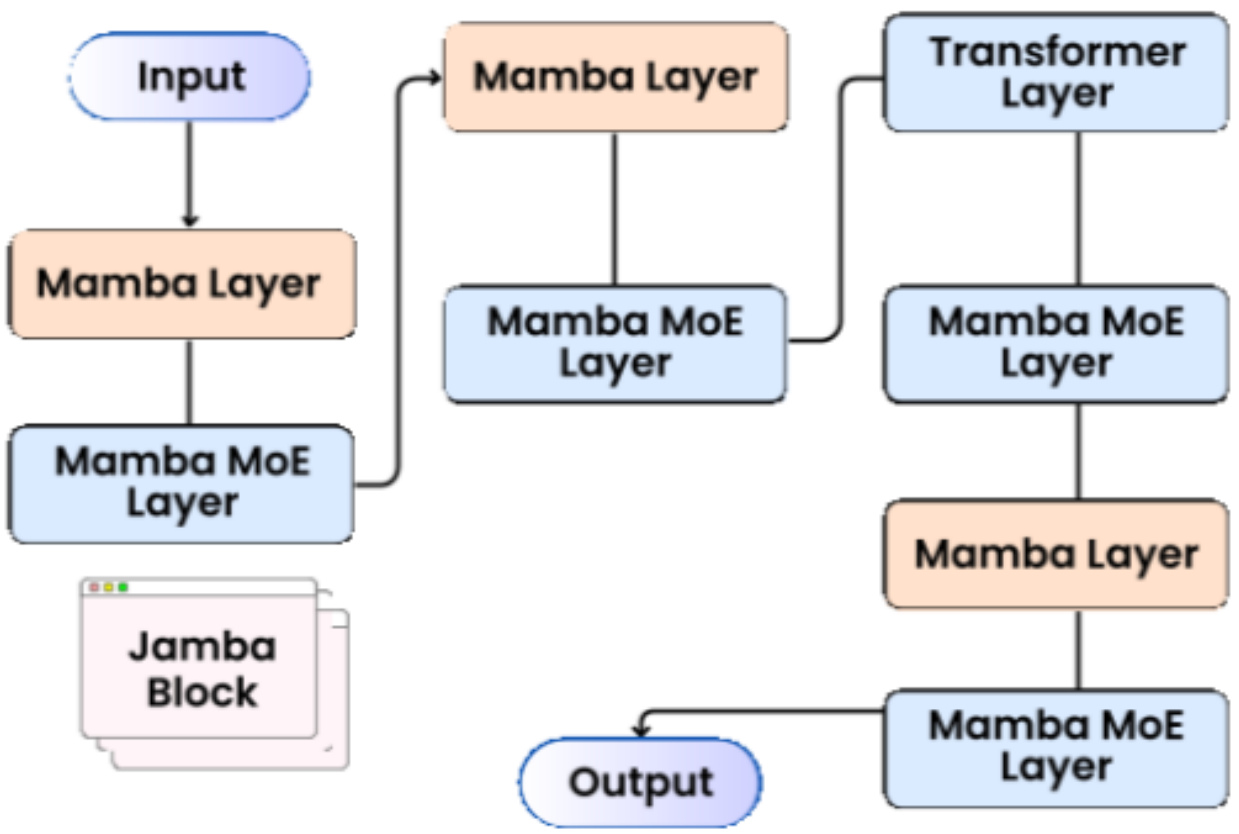
STATE SPACE MODEL



Unlike transformers that depend on attention mechanisms, State Space Models (SSMs) leverage state space equations and mathematical operations to capture long-range dependencies more efficiently. This makes them well-suited for tasks requiring long-context understanding and fast, low-latency processing.

Example: Mamba

HYBRID



This hybrid architecture integrates components from various model types — including transformers, Mixture of Experts (MoE), and State Space Models (SSMs) — to achieve a balance of accuracy, computational efficiency, and scalability. By combining the strengths of each approach, these models can flexibly handle diverse tasks within one unified system.

Example: Jamba (combines Mamba, Transformer, and MoE layers)

DroneX AI

WAS THIS POST USEFUL?

FOLLOW FOR MORE!

REPOST