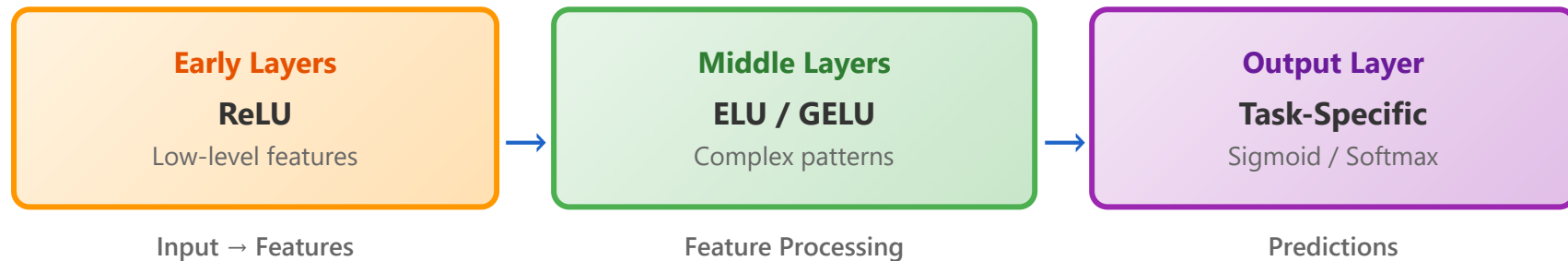


Layer-wise Activation Patterns

KEY OBSERVATION

Different layers may benefit from different activations



Empirical Finding

Uniform activation often works well, mixed can be better



Research Direction

Neural Architecture Search explores per-layer choices



Practical Tip

If experimenting, start from output and work backward



Implementation

Start simple (uniform), optimize if needed



Transformer Pattern Example

GELU in feedforward layers · **Linear (no activation)** in attention mechanisms