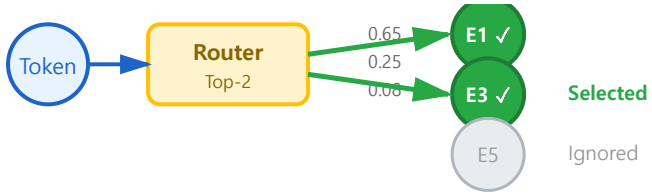
Expert Routing Strategies & Selection Mechanisms

Top-K Routing

Most Used



- Select K experts with highest scores
- Typical K=2 for efficiency
- Ensures sparse activation

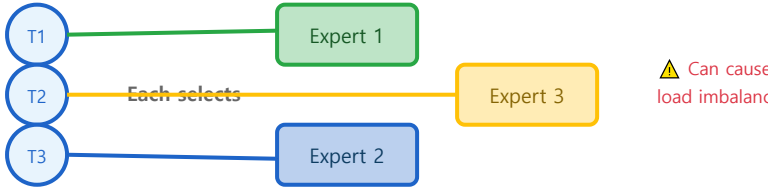
Soft Routing

Smooth

- Weighted combination of all experts
- Smoother gradients during training
- Higher computational cost

Token Choice

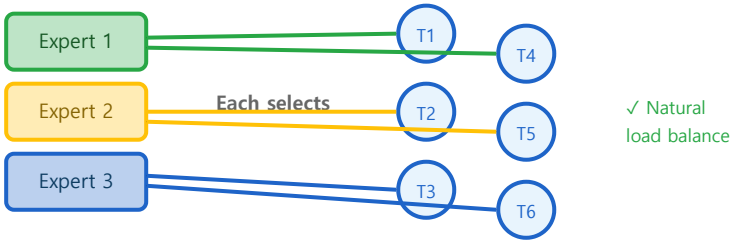
Specialized



- Each token selects its experts
- Better specialization

Expert Choice

Balanced



- Experts select tokens to process
- Natural load balancing

• Can lead to load imbalance

• Used in Switch Transformer (Google)

🎯 Routing Strategy Trade-offs

Different routing strategies balance between specialization quality, computational efficiency, and load balancing. Medical AI typically uses **Top-2 routing** for optimal performance with sparse activation.