Expert Routing Strategies & Selection Mechanisms

Top-K Routing

⭐ Most Used

Token → Router (Top-2) → 0.65 → E1 ✓ / 0.25 → E3 ✓ (Selected) / 0.08 → E5 (Ignored)
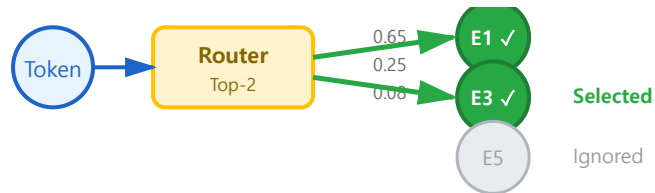
• Select K experts with highest scores

• Typical K=2 for efficiency

• Ensures sparse activation

Soft Routing

Smooth

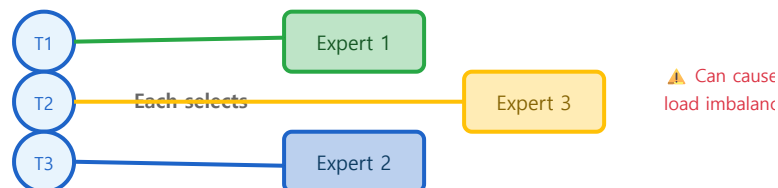Token → Soft Router (All weighted) → E1 w=0.50 / E2 w=0.30 / E3 w=0.15 / w=0.05

• Weighted combination of all experts

• Smoother gradients during training

• Higher computational cost

Token Choice

Specialized

T1 — Expert 1
T2 — Each selects — Expert 3
T3 — Expert 2

⚠ Can cause load imbalance

• Each token selects its experts

• Better specialization

Expert Choice

Balanced

Expert 1 — T1 — T4
Expert 2 — Each selects — T2 — T5
Expert 3 — T3 — T6

✓ Natural load balance
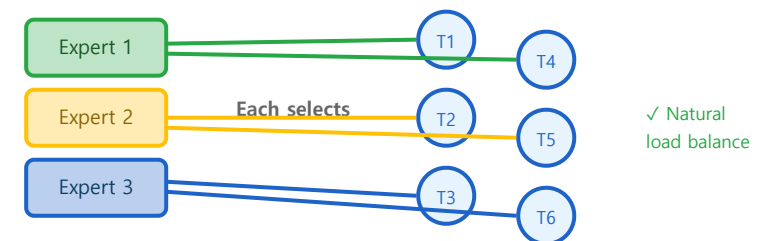
• Experts select tokens to process

• 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.