Load Balancing & Expert Utilization

Load Imbalance Problem

- Some experts receive too many tokens
- Others remain underutilized
- Degrades to few-expert model

Auxiliary Loss

- Encourages balanced routing
- Penalizes uneven distribution
- α * load_loss added to main loss

Expert Capacity

- Limit tokens per expert
- Drop excess tokens
- Forces router to balance

Random Routing

- Add noise to routing scores
- Stochastic expert selection
- Improves exploration

⚖️ Balancing Act

Effective load balancing ensures all experts are utilized efficiently, preventing capacity bottlenecks while maintaining specialization quality.