MoE Training Strategies & Stability

Initialization

- Careful router initialization
- Small initial routing noise
- Warm-up period for gating

Stability Techniques

- Gradient clipping
- Lower learning rate for router
- Layer normalization before routing

Curriculum Learning

- Start with fewer active experts
- Gradually increase sparsity
- Progressive expert specialization

Fine-tuning

- Domain-specific adaptation
- Expert-wise learning rates
- Selective expert freezing

🎓 Training Convergence

MoE models require careful training strategies to ensure stable convergence and effective expert specialization, typically taking 20-30% longer than dense models.