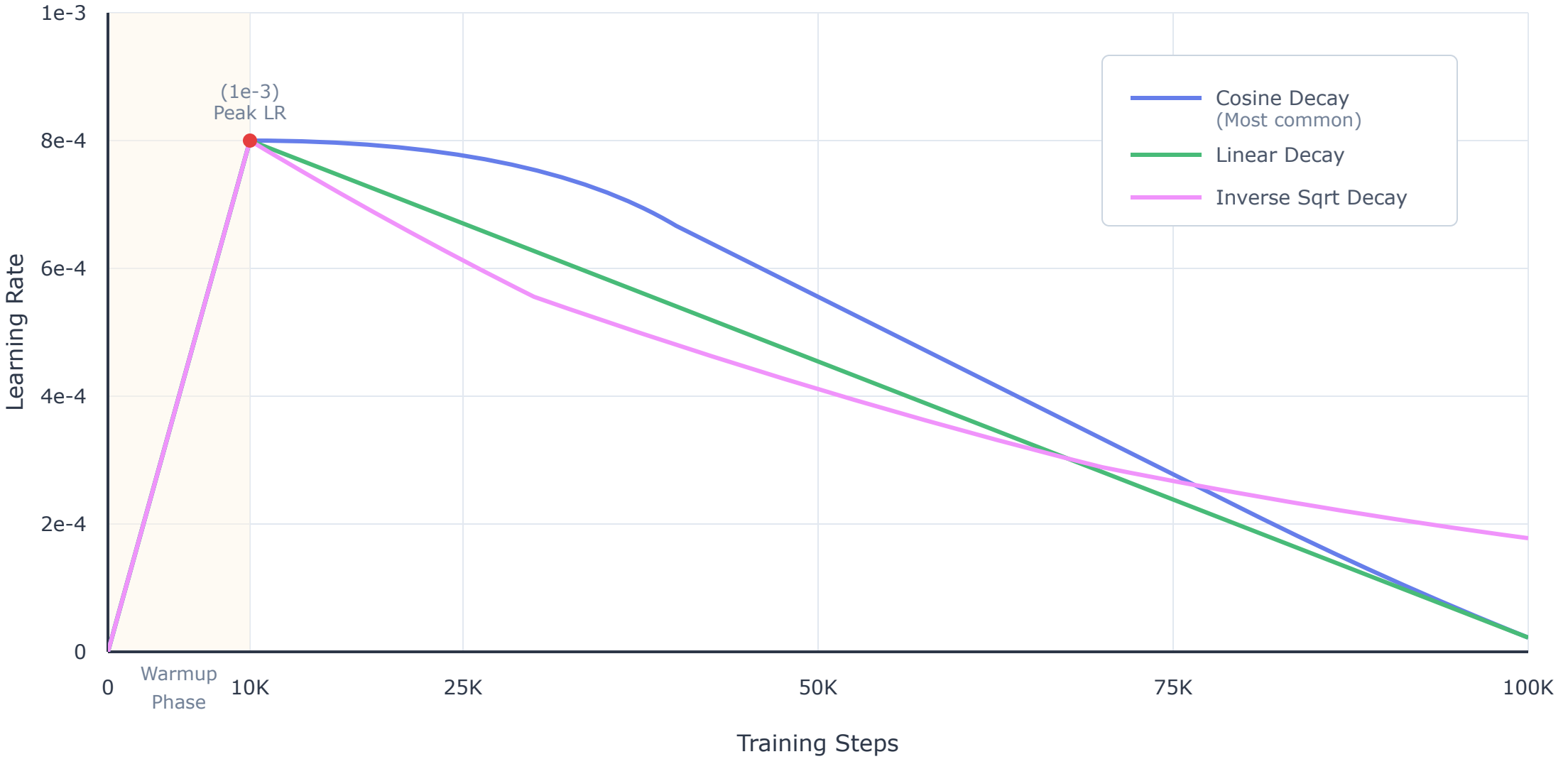
Learning Rate Schedules for Transformer Training



Warmup Phase: Prevents early training instability (typically 1-10% of total steps)
Decay Phase: Cosine decay typically provides 1-3% better final performance than constant LR