

Training Dynamics: Non-Monotonic Unfaithfulness Development

Qwen3-4B, 20000 Documents

