Qwen3.5-235B-A22B использует 235B total параметров с 22B активными в MoE-конфигурации (128 experts × 2/share), где Gated DeltaNet применяется только в нижних 25% слоёв, а верхние полагаются на классическое multi-head attention (MH-A) с Rotary Position Embeddings (RoPE).
Ключевое различие с Qwen3.5-35B-A3B: новая модель распределяет Gated DeltaNet равномерно по всем 48 слоям трансформера, достигая O(n) сложности внимания против O(n²) у предшественника. Это позволяет 35B-A3B обрабатывать 262K контекст без деградации качества (AA-LCR: 66.1 vs 60.0), хотя теоретически 235B с его 22B активными параметрами должен доминировать на knowledge-intensive задачах типа C-Eval (92.1 vs 90.5).
Парадокс объясняется улучшенным pretraining: Qwen3.5-35B обучалась на 18.5T токенов с curriculum learning (code:mixed CN-EN 35%, synthetic reasoning 22%, long-context docs 18%), где MoE-эксперты специализировались через expert dropout (p=0.2) и auxiliary losses на routing stability. RLHF в 3 этапа (SFT → DPO → ORPO) дал прирост +24.8 на IFBench за счёт лучшего instruction following.
В agentic coding 35B-A3B выигрывает TAU2-Bench (79.0 vs 58.5) благодаря hybrid attention, которое сохраняет локальные зависимости лучше чистого DeltaNet. FlashAttention-2 интегрировано во все слои с fused RMSNorm (Pre-LN), снижая memory bandwidth на 37% при 4-bit quantization (Q4_K_M).
Инфраструктура тренинга: 35B использовала 16× H100 (80GB) с ZeRO-3 + tensor parallelism (TP=4), потребляя 4.2M GPU-hours против 28.7M у 235B. KV-cache compression (H2O eviction) даёт 2.1x speedup на длинных последовательностях >128K токенов.