Zero-Cost Optimization: 3.31× Speedup

Latency per Query

- 62.2s — 1.00×
- Reduce tokens (-57%)
- 24.0s — 2.59×
- Smaller model (-22%)
- 18.8s — 3.31×

Y-axis: Latency per Query (s)

X-axis: Optimization Stage
- Baseline (3B, 128/512)
- Params Only (3B, 50/200)
- Full Opt (1.5B, 50/200)