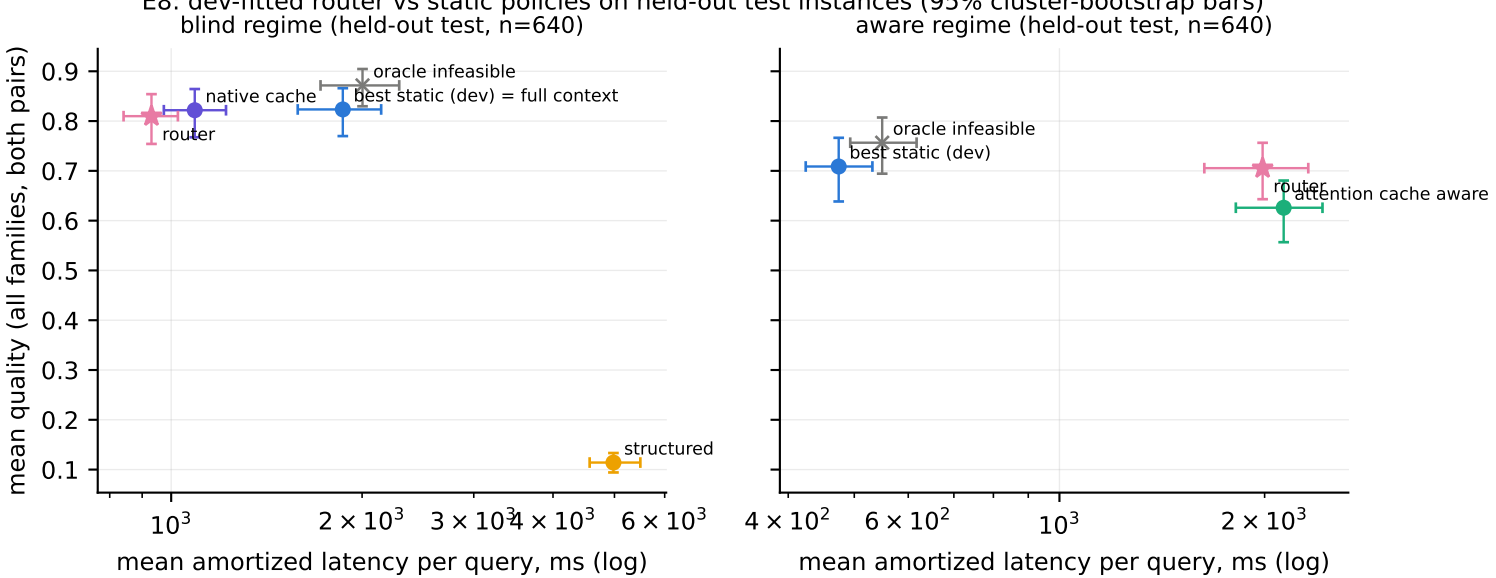

E8: dev-fitted router vs static policies on held-out test instances (95% cluster-bootstrap bars)
blind regime (held-out test, n=640)
aware regime (held-out test, n=640)
mean quality (all families, both pairs)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
oracle infeasible
native cache
best static (dev) = full context
router
structured
10^3
2×10^3
3×10^3
4×10^3
6×10^3
mean amortized latency per query, ms (log)
oracle infeasible
best static (dev)
router
attention cache aware
4×10^2
6×10^2
10^3
2×10^3
mean amortized latency per query, ms (log)