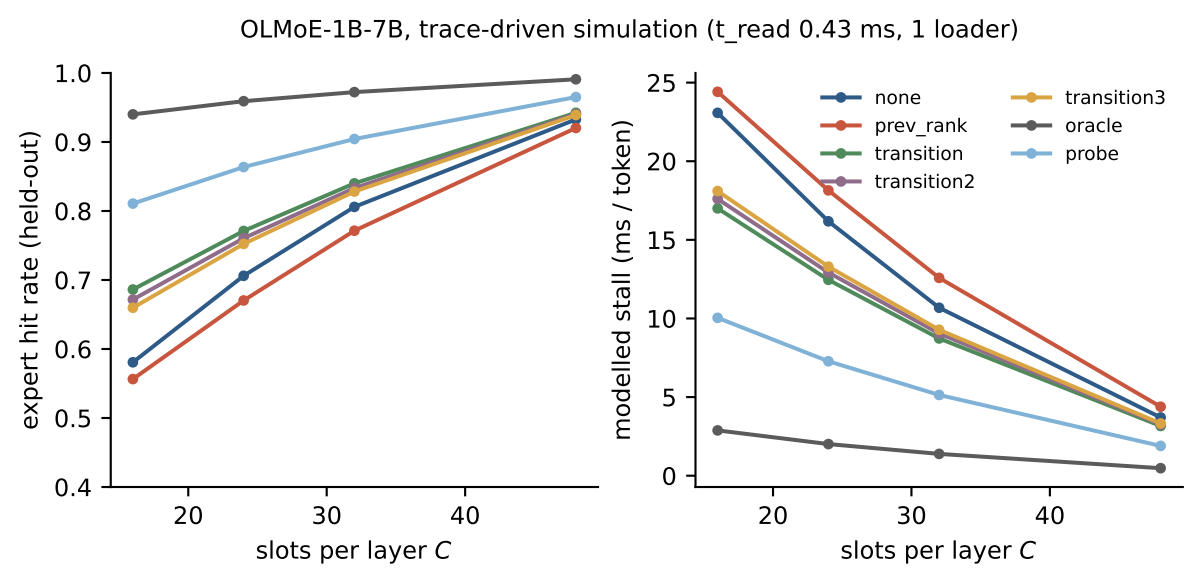

OLMoE-1B-7B, trace-driven simulation (t_read 0.43 ms, 1 loader)
expert hit rate (held-out)
slots per layer C
modelled stall (ms / token)
slots per layer C
none
prev_rank
transition
transition2
transition3
oracle
probe