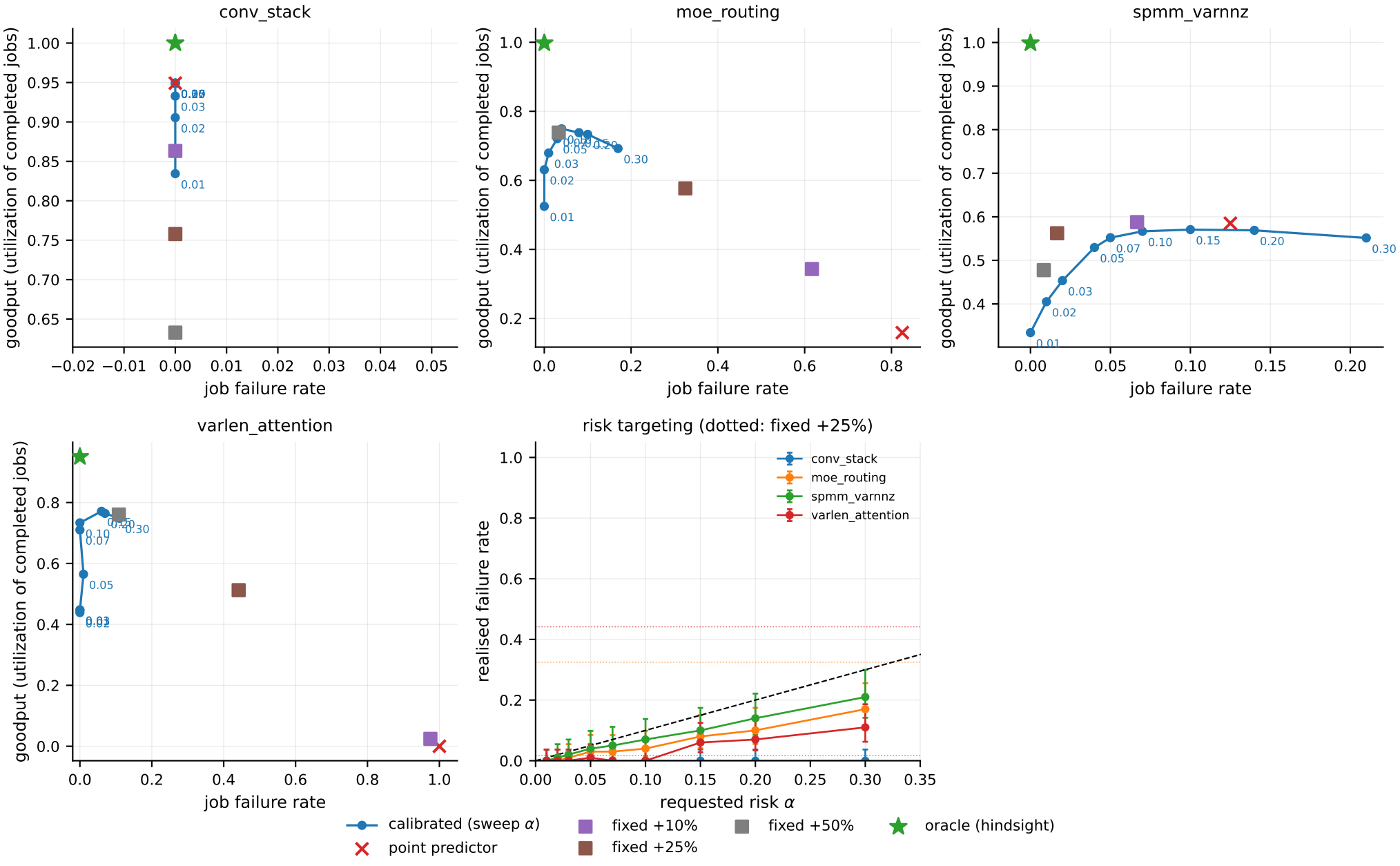
conv_stack
moe_routing
spmm_varnnz
varlen_attention
risk targeting (dotted: fixed +25%)
goodput (utilization of completed jobs)
job failure rate
realised failure rate
requested risk α
conv_stack
moe_routing
spmm_varnnz
varlen_attention
calibrated (sweep α)
point predictor
fixed +10%
fixed +25%
fixed +50%
oracle (hindsight)