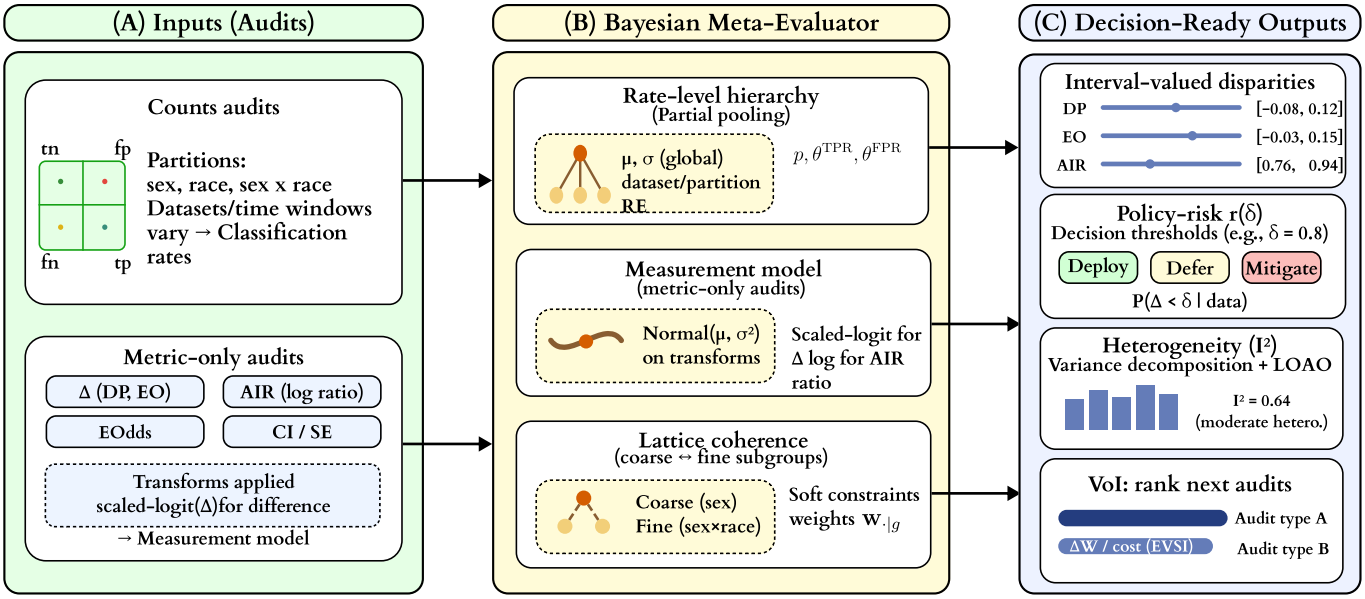
(A) Inputs (Audits)
Counts audits
tn
fp
fn
tp
Partitions:
sex, race, sex x race
Datasets/time windows
vary → Classification rates
Metric-only audits
Δ (DP, EO)
AIR (log ratio)
EOdds
CI / SE
Transforms applied
scaled-logit(Δ)for difference
→ Measurement model
(B) Bayesian Meta-Evaluator
Rate-level hierarchy
(Partial pooling)
μ, σ (global)
dataset/partition RE
$p, \theta^{\mathrm{TPR}}, \theta^{\mathrm{FPR}}$
Measurement model
(metric-only audits)
Normal(μ, σ²) on transforms
Scaled-logit for Δ log for AIR ratio
Lattice coherence
(coarse ↔ fine subgroups)
Coarse (sex)
Fine (sex×race)
Soft constraints
weights $w_{\cdot|g}$
(C) Decision-Ready Outputs
Interval-valued disparities
DP [−0.08, 0.12]
EO [−0.03, 0.15]
AIR [0.76, 0.94]
Policy-risk r(δ)
Decision thresholds (e.g., δ = 0.8)
Deploy
Defer
Mitigate
P(Δ < δ | data)
Heterogeneity (I²)
Variance decomposition + LOAO
I² = 0.64
(moderate hetero.)
VoI: rank next audits
Audit type A
ΔW / cost (EVSI)
Audit type B