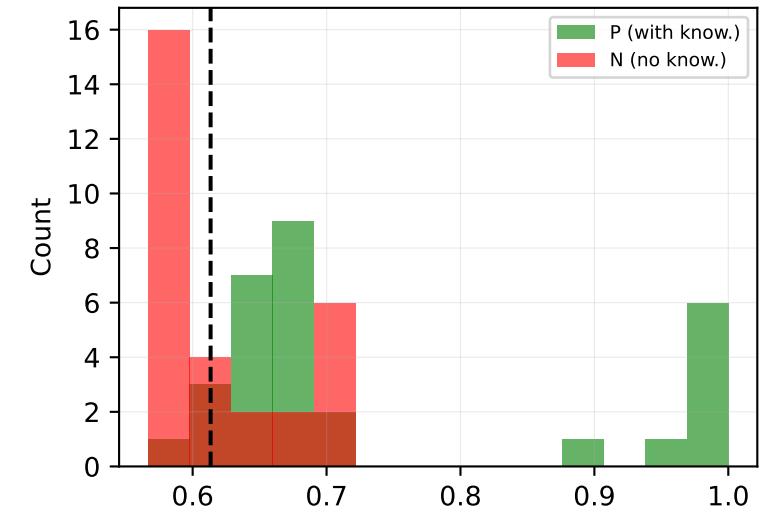
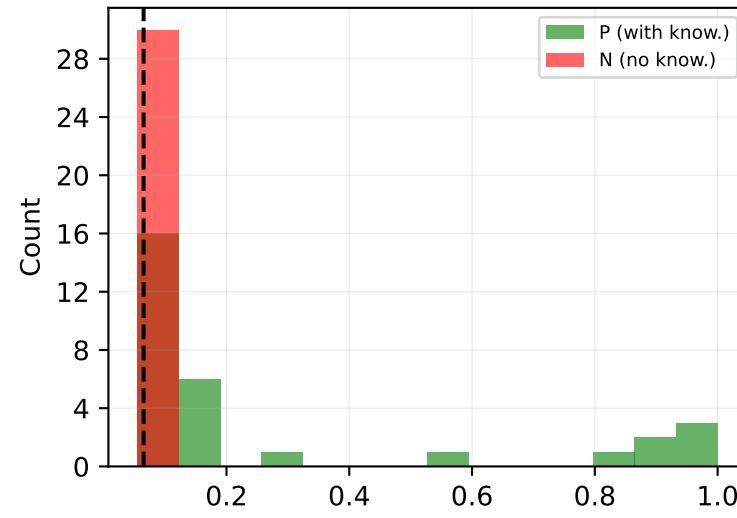


Faithfulness: P/N Pool Score Distributions (17 Metrics)
60 models (30 P + 30 N)

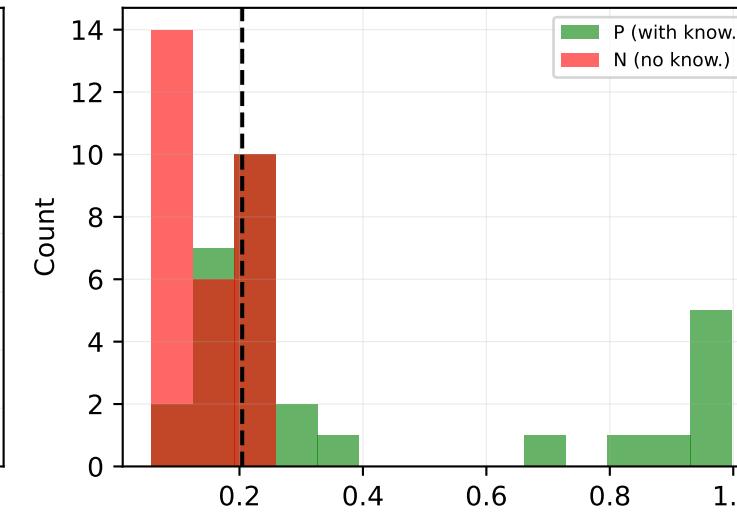
Exact Memorization
AUC: 0.817



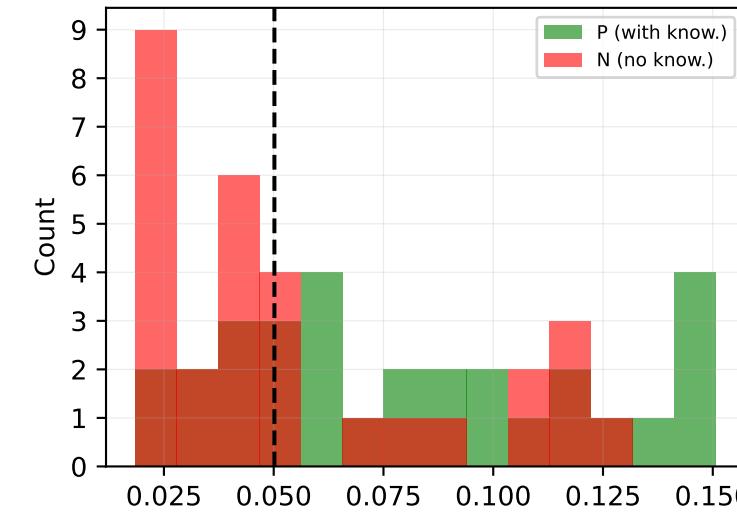
Extraction Strength
AUC: 0.891



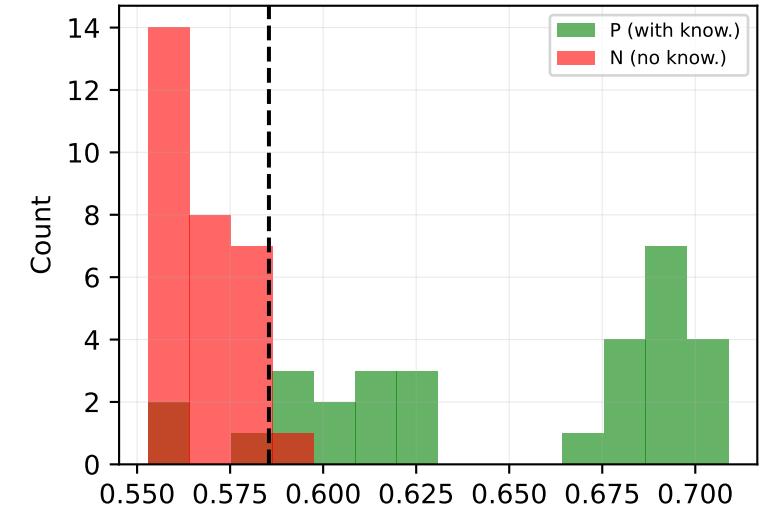
Probability
AUC: 0.816



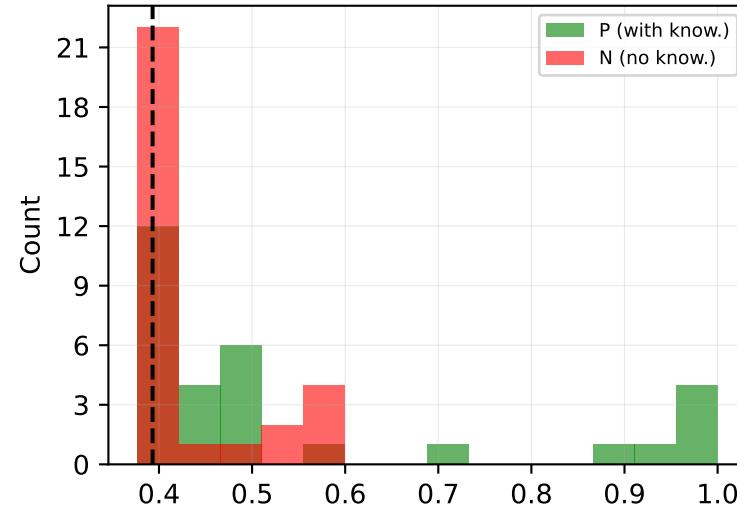
Paraphrase Prob.
AUC: 0.707



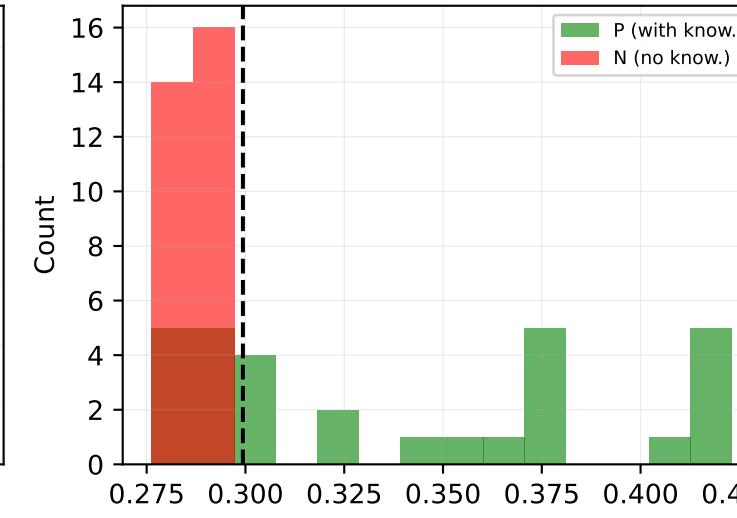
Truth Ratio
AUC: 0.947



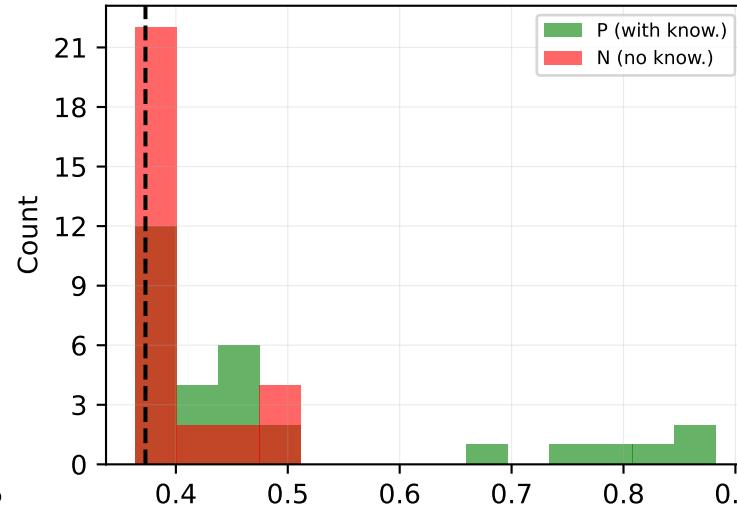
ROUGE
AUC: 0.722



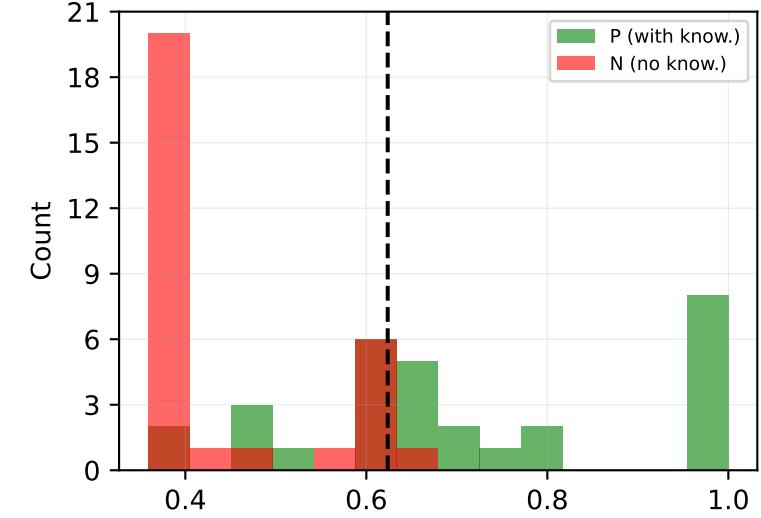
Paraphrase ROUGE
AUC: 0.832



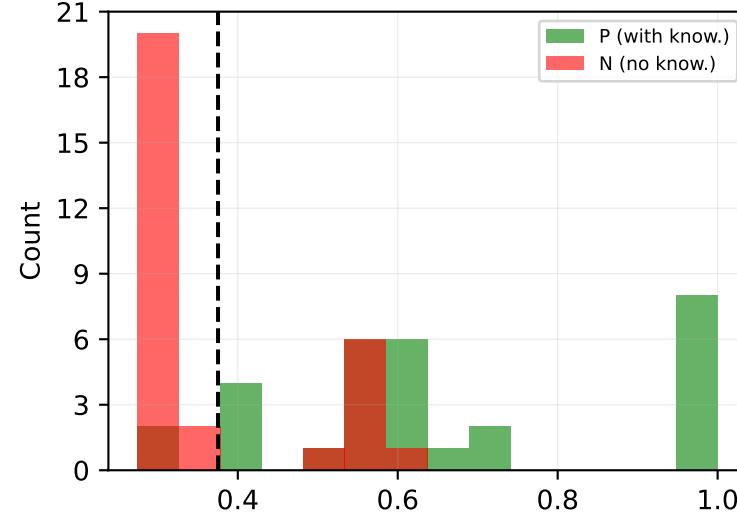
Jailbreak ROUGE
AUC: 0.757



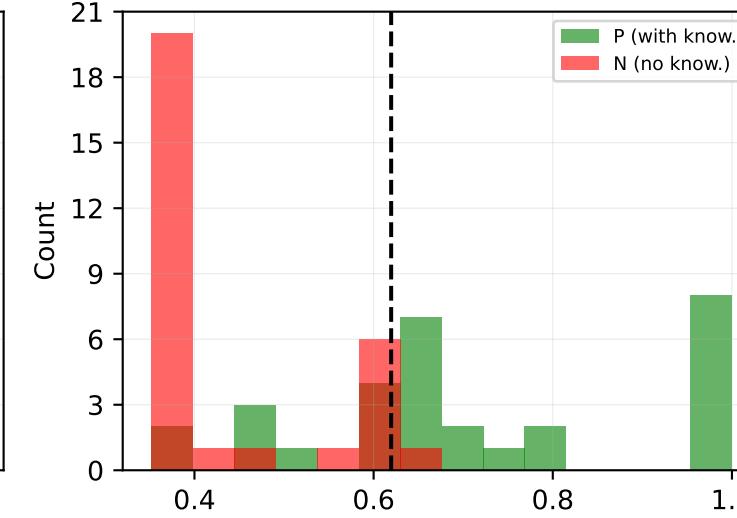
MIA-LOSS (raw AUC)
AUC: 0.902



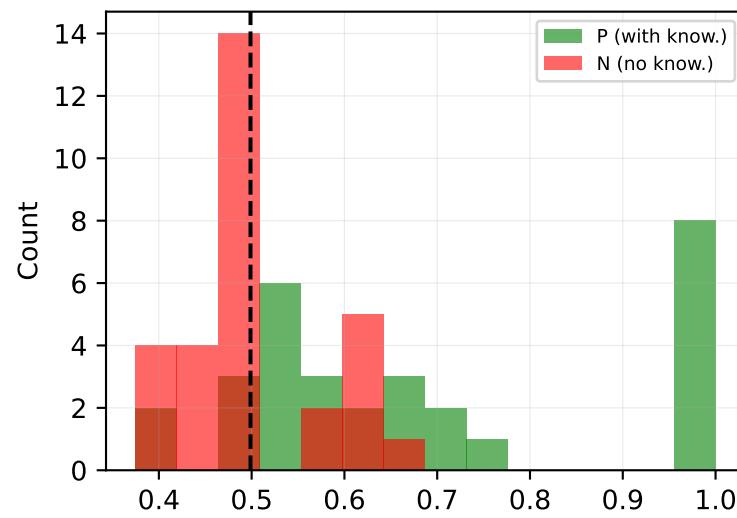
MIA-ZLib (raw AUC)
AUC: 0.867



MIA-MinK (raw AUC)
AUC: 0.907



MIA-MinK++ (raw AUC)
AUC: 0.816



MIA-LOSS (normalized)
AUC: 0.891

