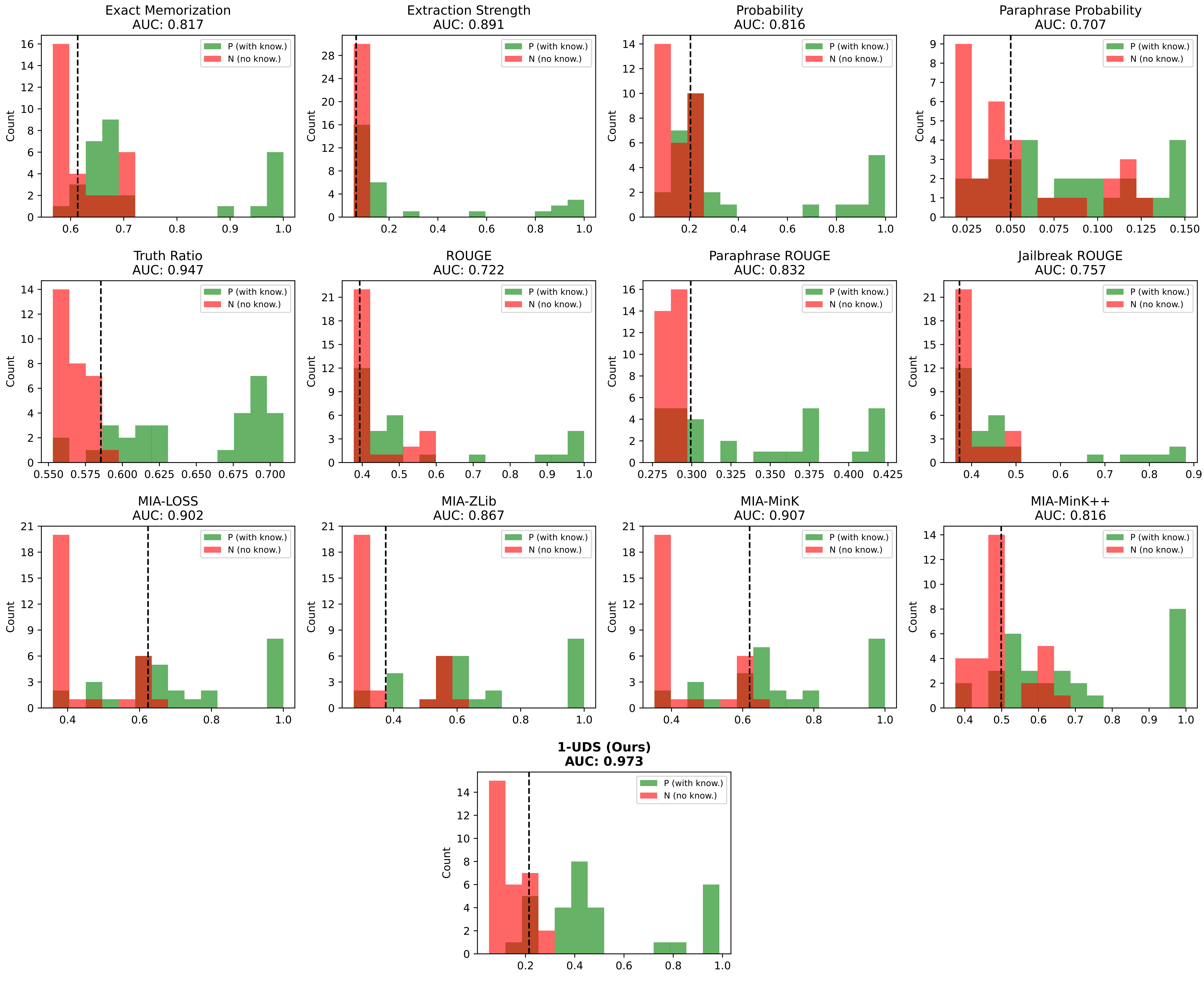
Faithfulness: P/N Pool Score Distributions (13 Metrics)
60 models (30 P + 30 N)


Exact Memorization
AUC: 0.817
Extraction Strength
AUC: 0.891
Probability
AUC: 0.816
Paraphrase Probability
AUC: 0.707
Truth Ratio
AUC: 0.947
ROUGE
AUC: 0.722
Paraphrase ROUGE
AUC: 0.832
Jailbreak ROUGE
AUC: 0.757
MIA-LOSS
AUC: 0.902
MIA-ZLib
AUC: 0.867
MIA-MinK
AUC: 0.907
MIA-MinK++
AUC: 0.816
1-UDS (Ours)
AUC: 0.973
Count
🟩 P (with know.)
🟥 N (no know.)