Confidence Distributions: Correct vs Incorrect Predictions

Claude-Sonnet-4

ACC: 0.493
Conf Sep: 0.056

Correct
Incorrect

Density

Confidence Score

Kimi-K2

ACC: 0.483
Conf Sep: 0.037

Correct
Incorrect

Density

Confidence Score

DeepSeek-V3.1

ACC: 0.482
Conf Sep: 0.036

Correct
Incorrect

Density

Confidence Score

Gemini-2.5

ACC: 0.444
Conf Sep: 0.063

Correct
Incorrect

Density

Confidence Score

GPT-4.1

ACC: 0.430
Conf Sep: 0.041

Correct
Incorrect

Density

Confidence Score

Qwen3-235B

ACC: 0.418
Conf Sep: 0.019

Correct
Incorrect

Density

Confidence Score