Calibrated Coverage: LLM Predictions vs Ground Truth

- - - Random baseline

claude	39.4%
gpt5	15.4%
deepseekthink	14.9%
deepseek	14.4%
qwenthink	13.9%
gemini3	13.0%
llama	10.1%
qwen	5.8%

Calibrated Coverage (%)