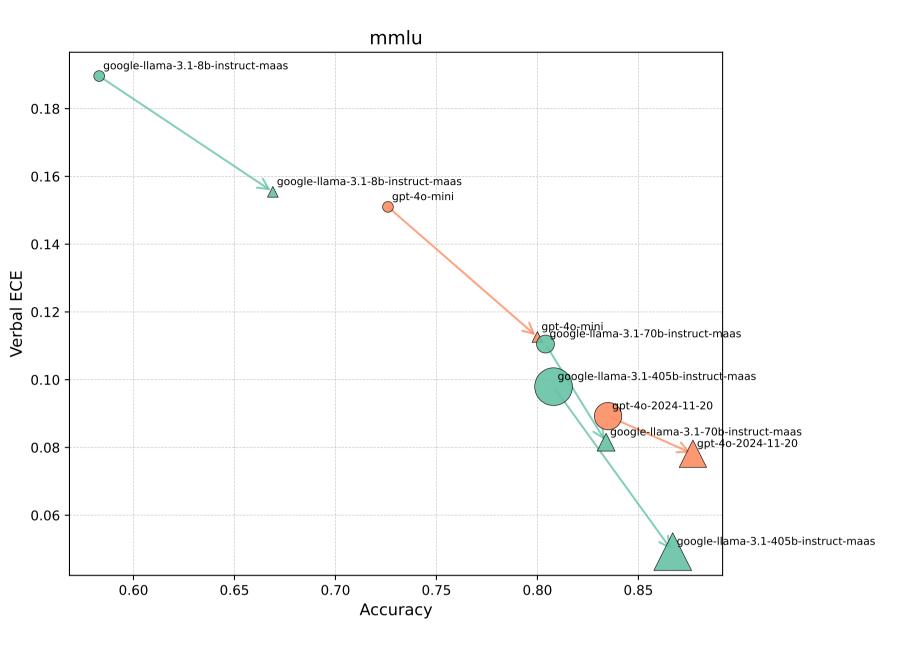
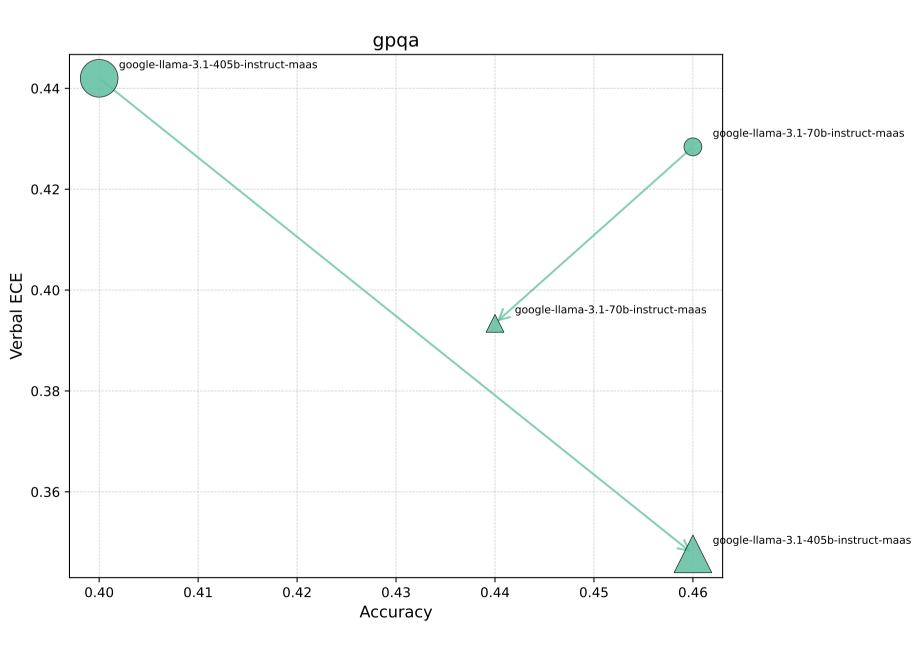
Model Performance by Benchmark





Legend

▲ CoT

Non-CoT

GPT

LLaMA