Model Performance by Benchmark

mmlu

Scatter plot with x-axis "Accuracy" (ranging from 0.60 to 0.85) and y-axis "Verbal ECE" (ranging from 0.06 to 0.18). Data points labeled:
- google-llama-3.1-8b-instruct-maas
- google-llama-3.1-8b-instruct-maas
- gpt-4o-mini
- gpt-4o-mini
- google-llama-3.1-70b-instruct-maas
- google-llama-3.1-405b-instruct-maas
- gpt-4o-2024-11-20
- google-llama-3.1-70b-instruct-maas
- gpt-4o-2024-11-20
- google-llama-3.1-405b-instruct-maas

Legend: LLaMA, GPT, CoT, Non-CoT