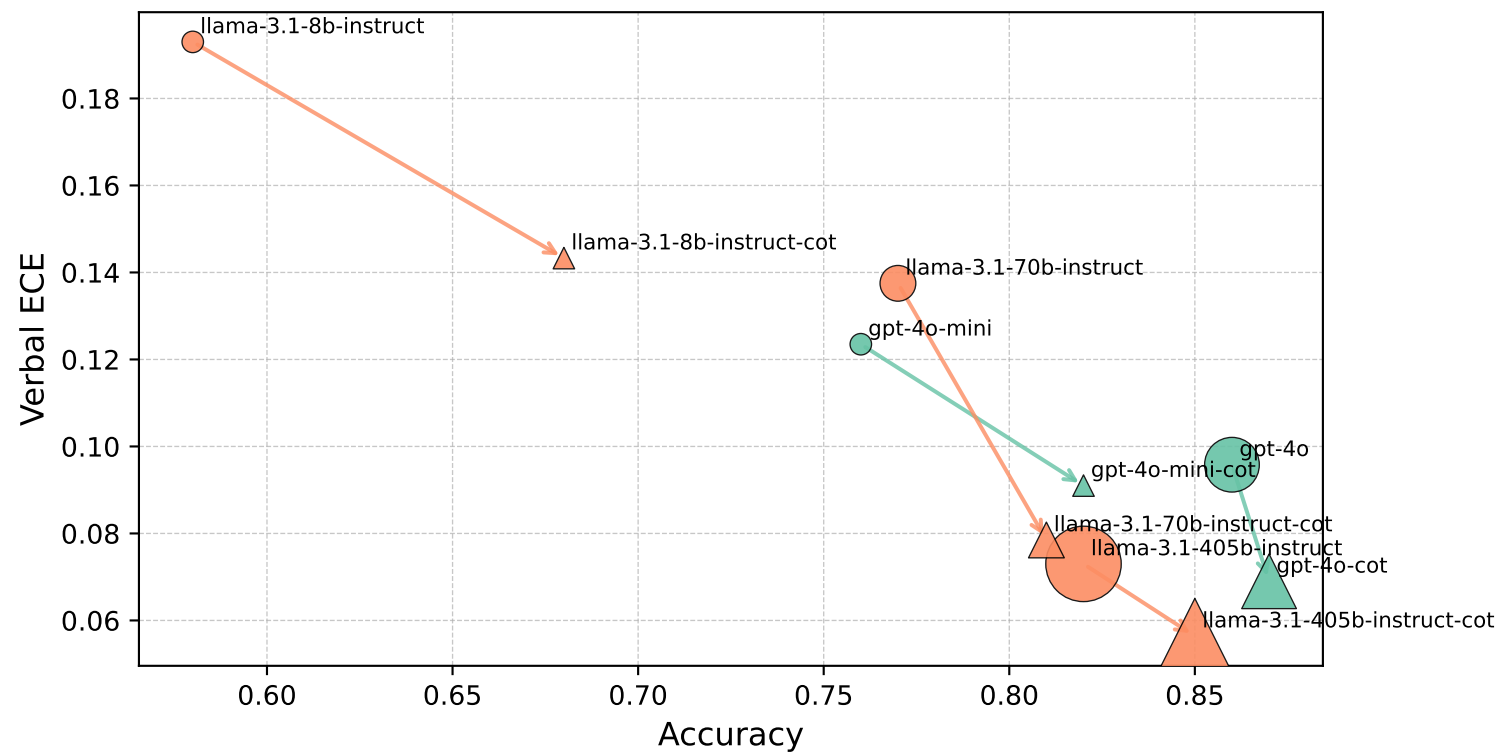# Model Performance by Benchmark

mmlu

llama-3.1-8b-instruct
llama-3.1-8b-instruct-cot
llama-3.1-70b-instruct
gpt-4o-mini
gpt-4o-mini-cot
gpt-4o
llama-3.1-70b-instruct-cot
llama-3.1-405b-instruct
gpt-4o-cot
llama-3.1-405b-instruct-cot

Accuracy
Verbal ECE

simpleqa

gpt-4o-mini
llama-3.1-8b-instruct
gpt-4o-mini-cot
gpt-4o
gpt-4o-cot
llama-3.1-70b-instruct-cot
llama-3.1-8b-instruct-cot
llama-3.1-70b-instruct
llama-3.1-405b-instruct-cot
llama-3.1-405b-instruct

Accuracy
Verbal ECE

Legend
GPT   LLaMA   CoT   Non-CoT