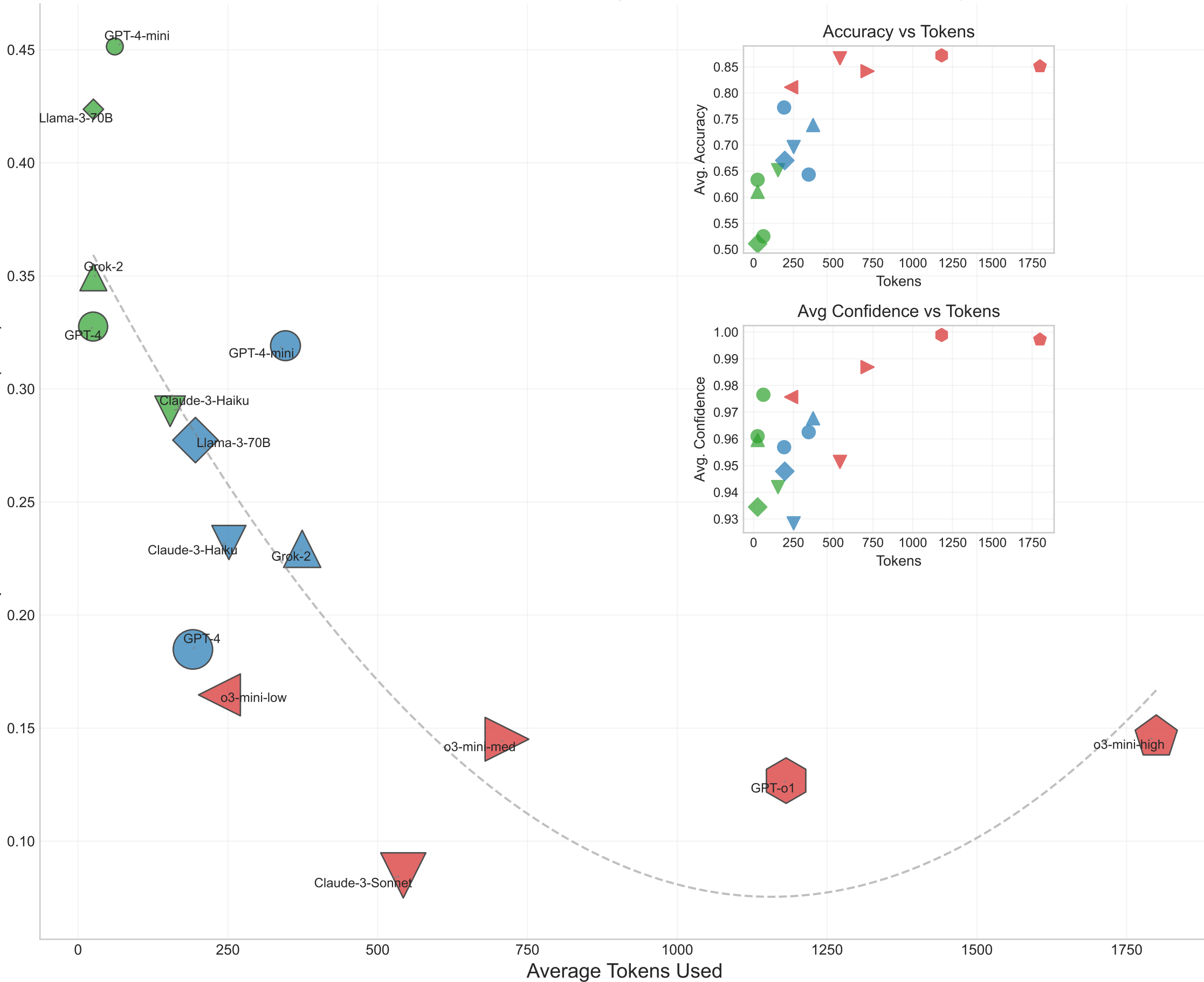


Relationship between Token Usage, Calibration Error, and Accuracy

Expected Calibration Error (ECE)



Reasoning Methods

- Prompted Reasoning
- Reason
- No Reasoning
- Trend

Model Families

- Openai 4O
- Openai 4Omini
- Xai Grok2
- Meta Llama
- Claude
- Openai O3Mini Low
- Openai O3Mini Medium
- Openai O3Mini High
- Openai O1
- Qwen Qwq