Accuracy vs. Cost

Overall Accuracy vs. Total Tokens (cost proxy)

- Llama 3.1 8B (Cerebras)
- GPT-OSS 120B (Cerebras)
- Pareto Frontier