Model Family Performance Comparison
Mean Score with Standard Deviation

Average Benchmark Score (%)

- Alibaba (Qwen): 80.8% (n=4)
- Anthropic: 78.0% (n=5)
- DeepSeek: 68.7% (n=3)
- Google: 56.6% (n=10)