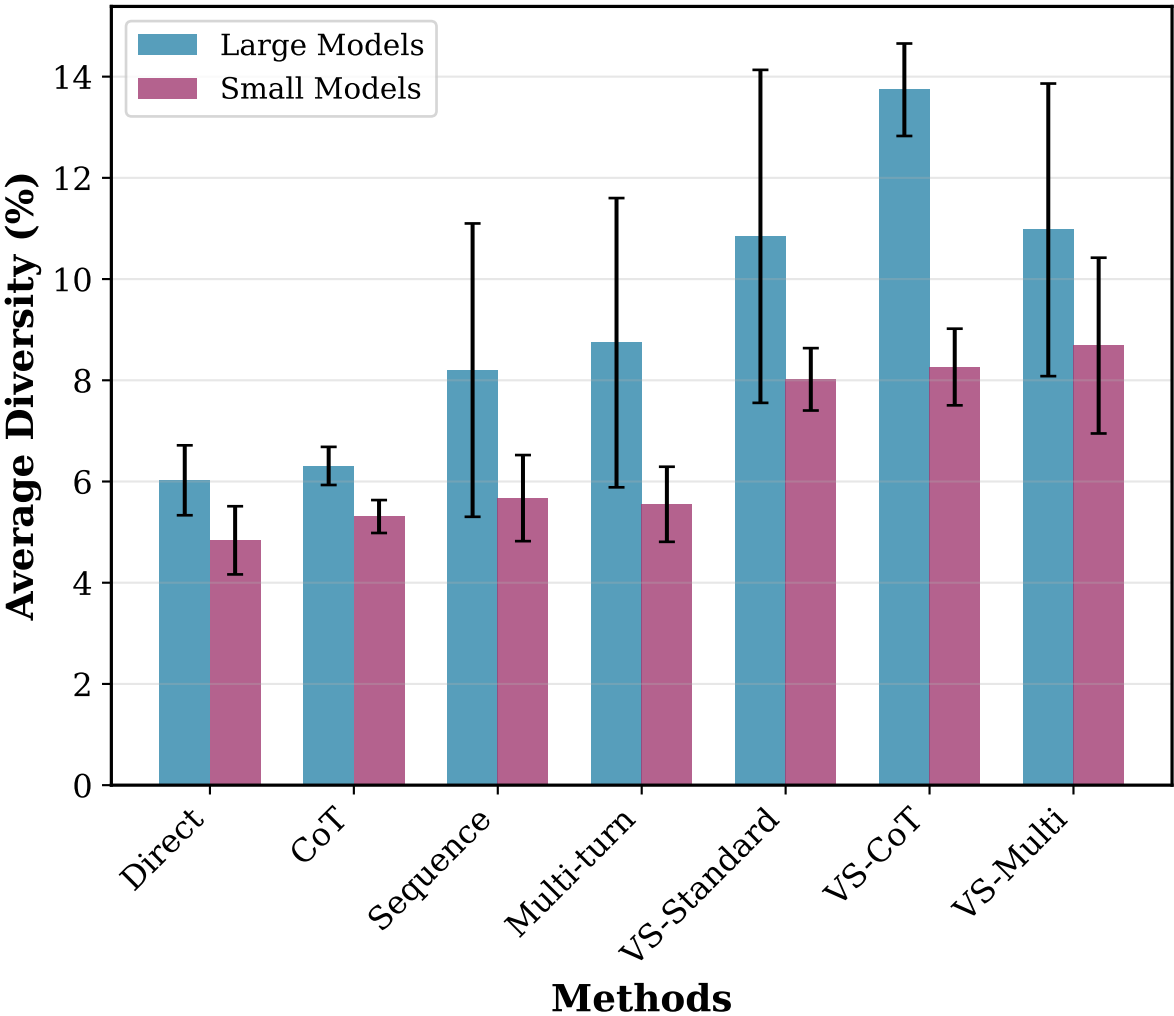
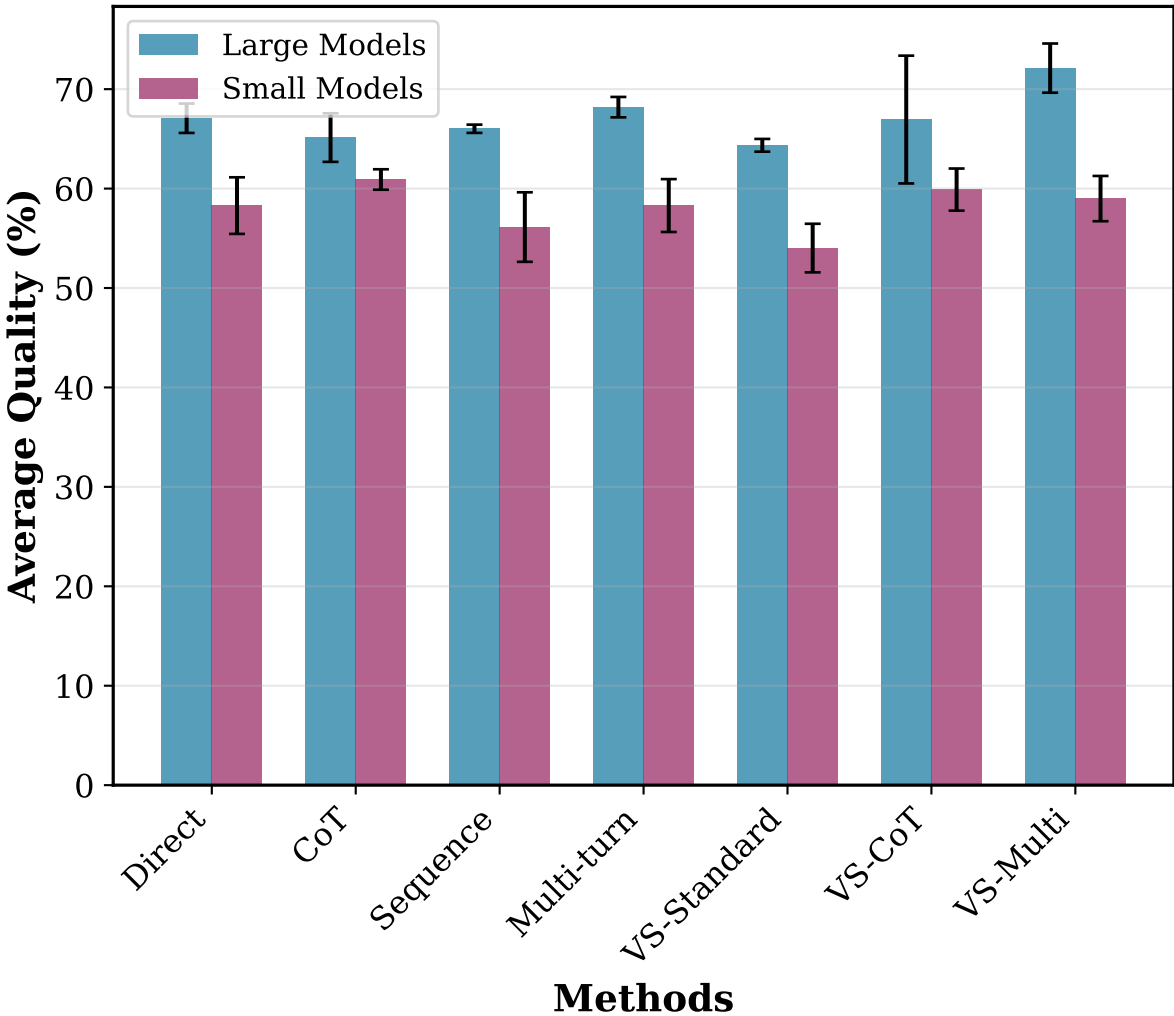


Method Effectiveness Analysis by Model Size

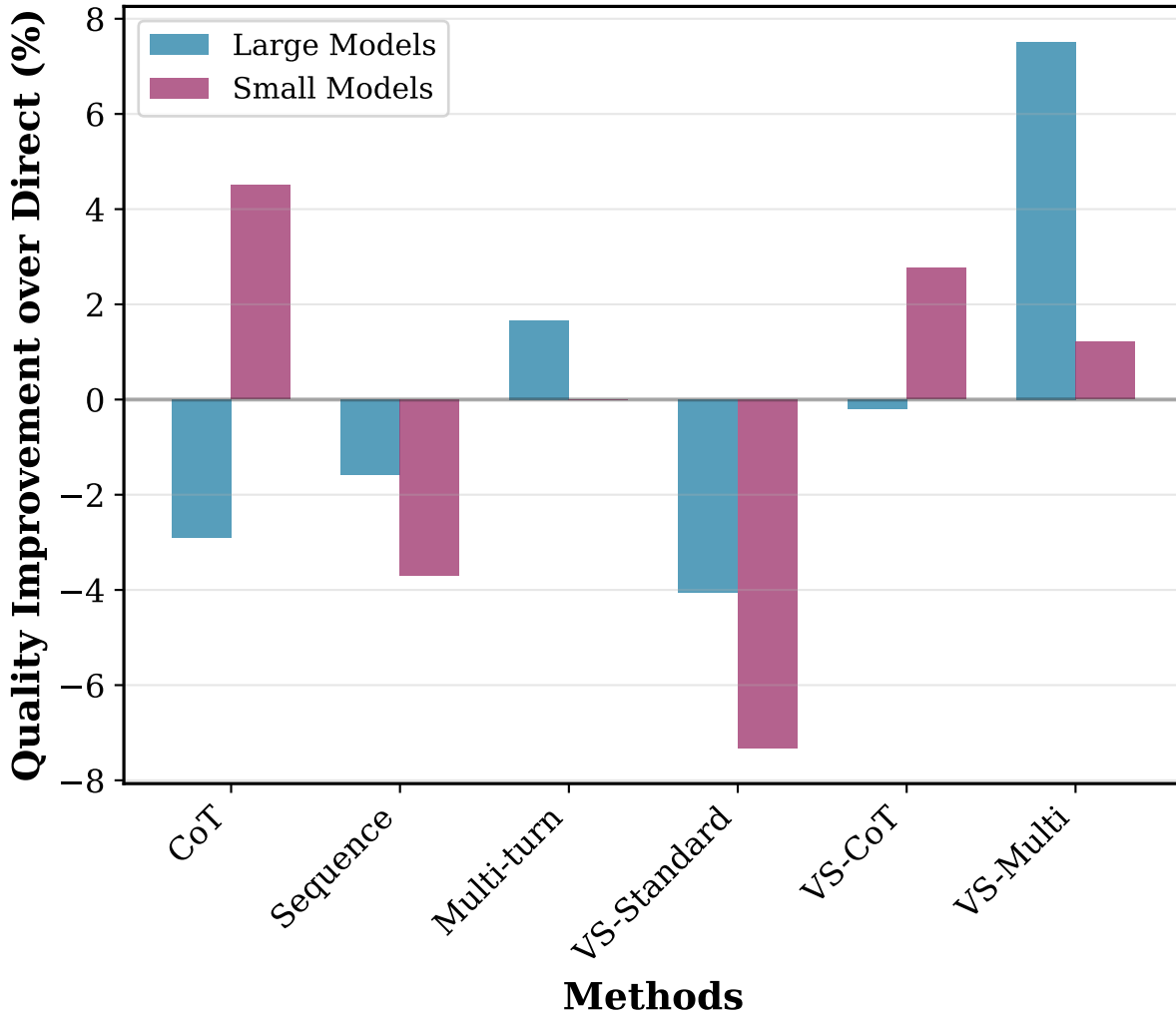
Diversity by Model Size



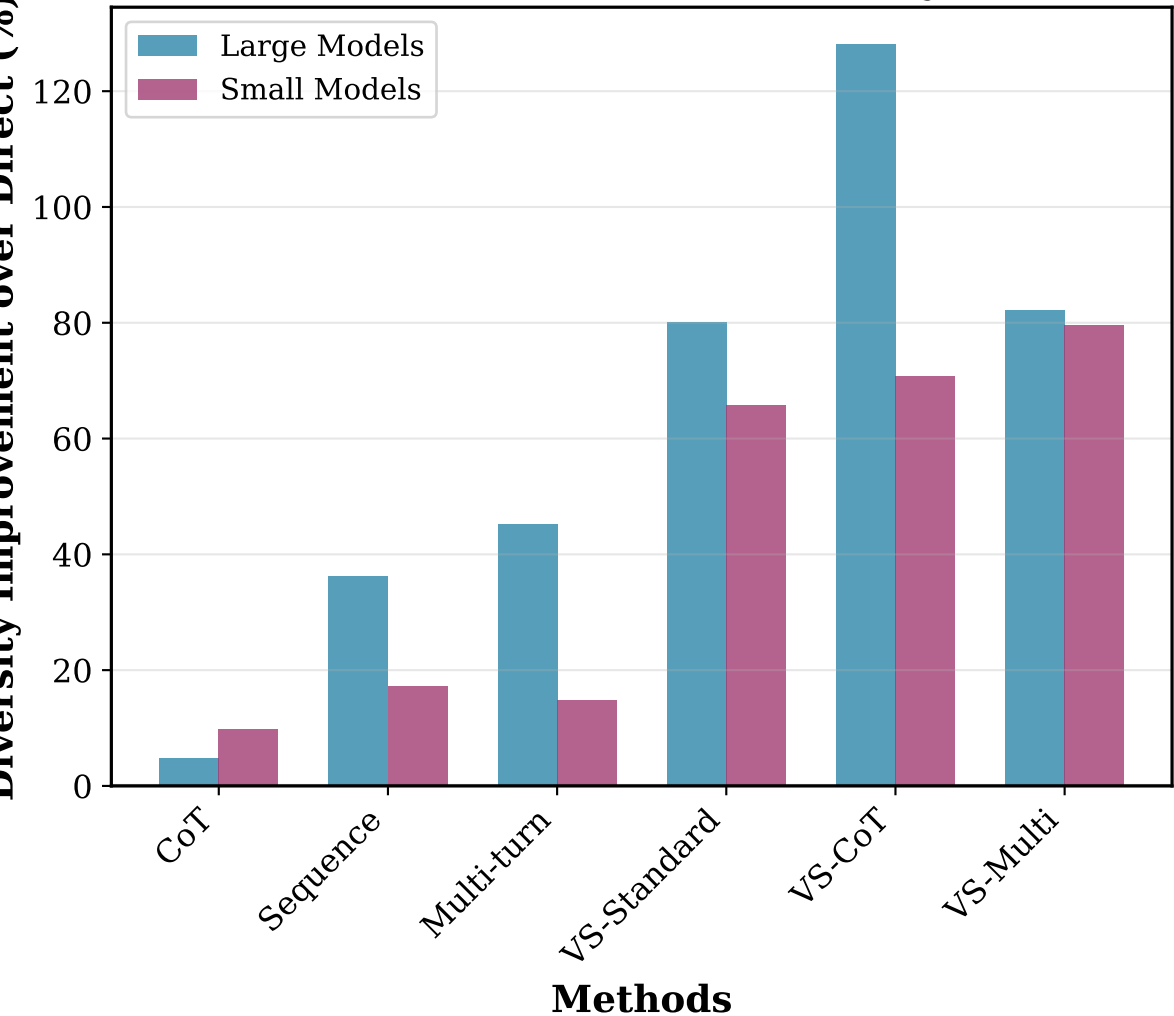
Quality by Model Size



Method Effectiveness: Quality Gains



Method Effectiveness: Diversity Gains



Model Size Classifications

MODEL SIZE CLASSIFICATIONS:

LARGE MODELS:

- GPT-4.1
- Gemini-2.5-Pro

Total: 2 models

SMALL MODELS:

- GPT-4.1-Mini
- Gemini-2.5-Flash

Total: 2 models

Statistical Significance Tests

STATISTICAL SIGNIFICANCE (Large vs Small Models):		
Method	Diversity	Quality
Standard	ns	ns
CoT	*	ns
Multi	ns	ns

Legend: *** p<0.001, ** p<0.01, * p<0.05, ns = not significant