Method Effectiveness Analysis by Model Size
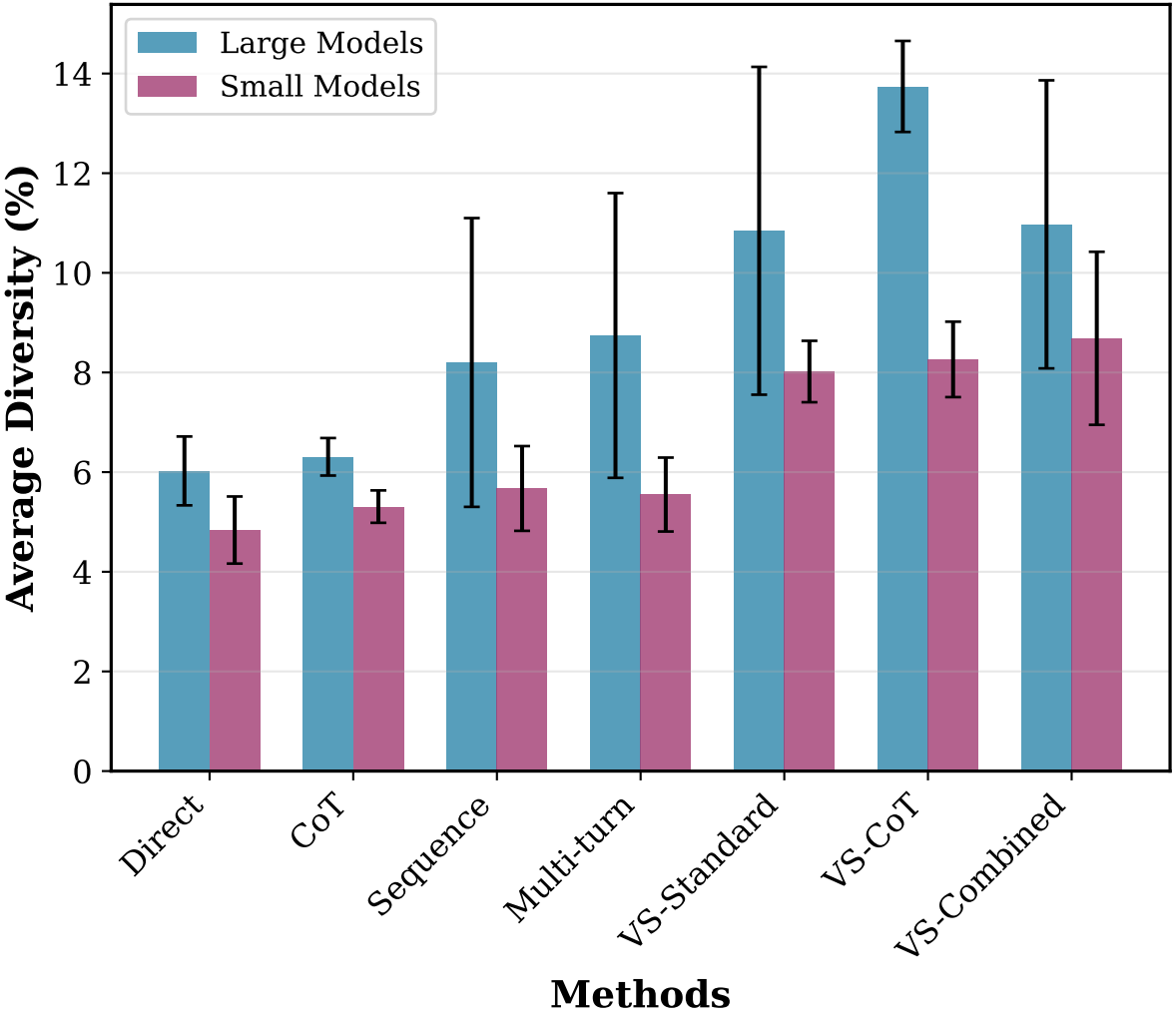
Diversity by Model Size

Methods: Direct, CoT, Sequence, Multi-turn, VS-Standard, VS-CoT, VS-Combined

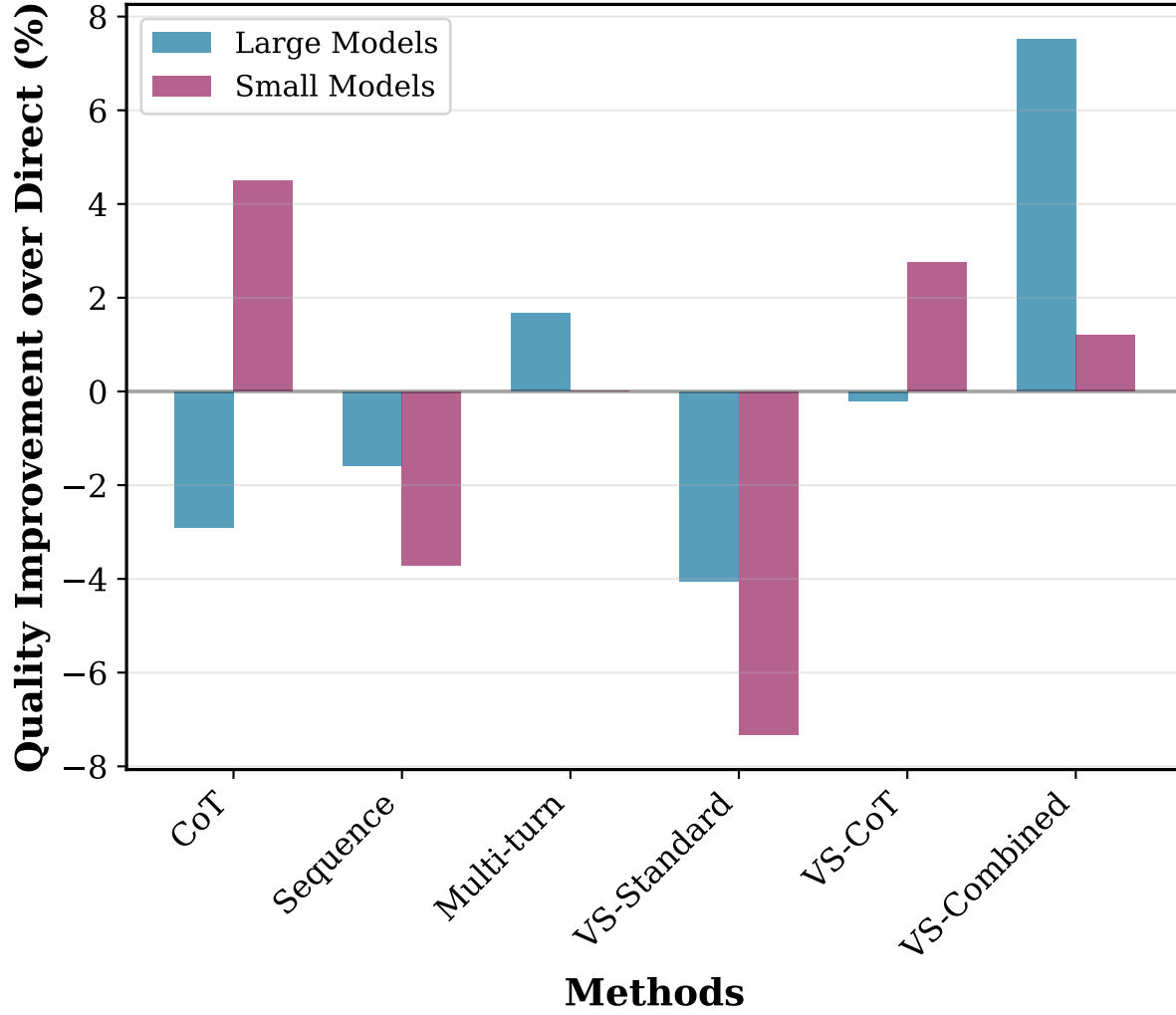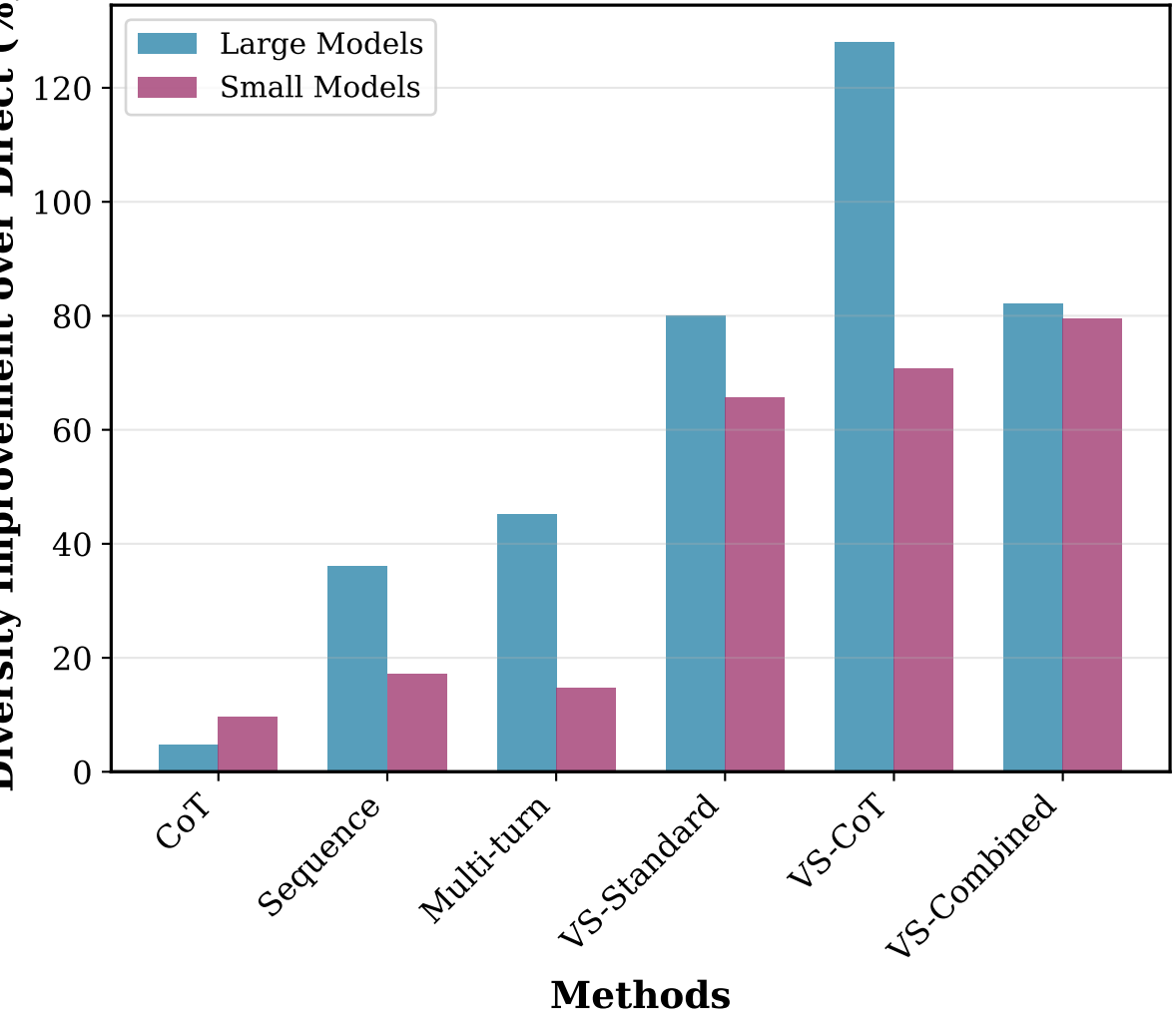Legend: Large Models, Small Models

Quality by Model Size

Methods: Direct, CoT, Sequence, Multi-turn, VS-Standard, VS-CoT, VS-Combined

Legend: Large Models, Small Models
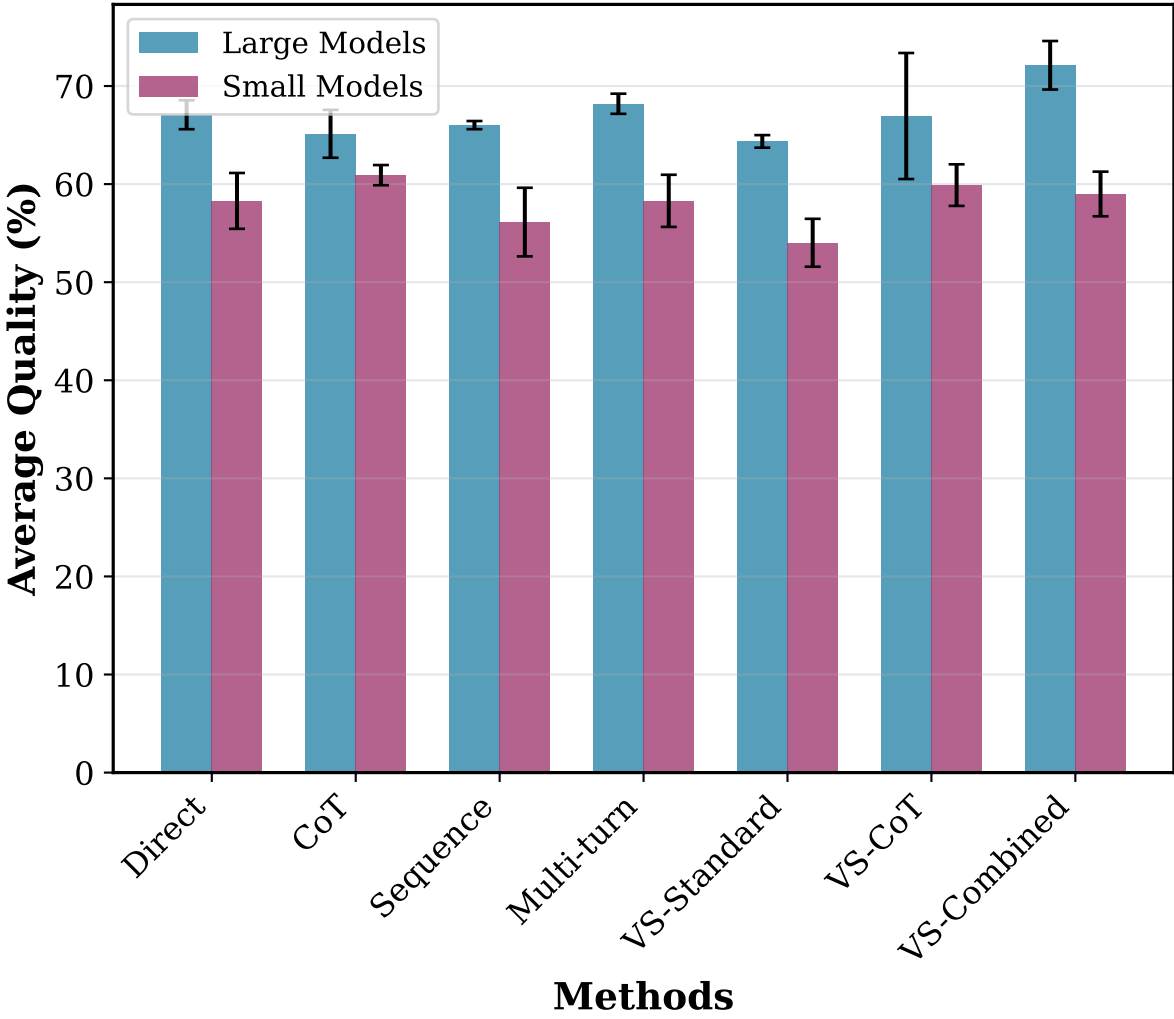
Method Effectiveness: Quality Gains

Methods: CoT, Sequence, Multi-turn, VS-Standard, VS-CoT, VS-Combined

Legend: Large Models, Small Models

Method Effectiveness: Diversity Gains

Methods: CoT, Sequence, Multi-turn, VS-Standard, VS-CoT, VS-Combined

Legend: Large Models, Small Models

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
Combined        |        ns |      ns

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