Robustness in Weakness Profiling: Traditional vs DOVE Analysis
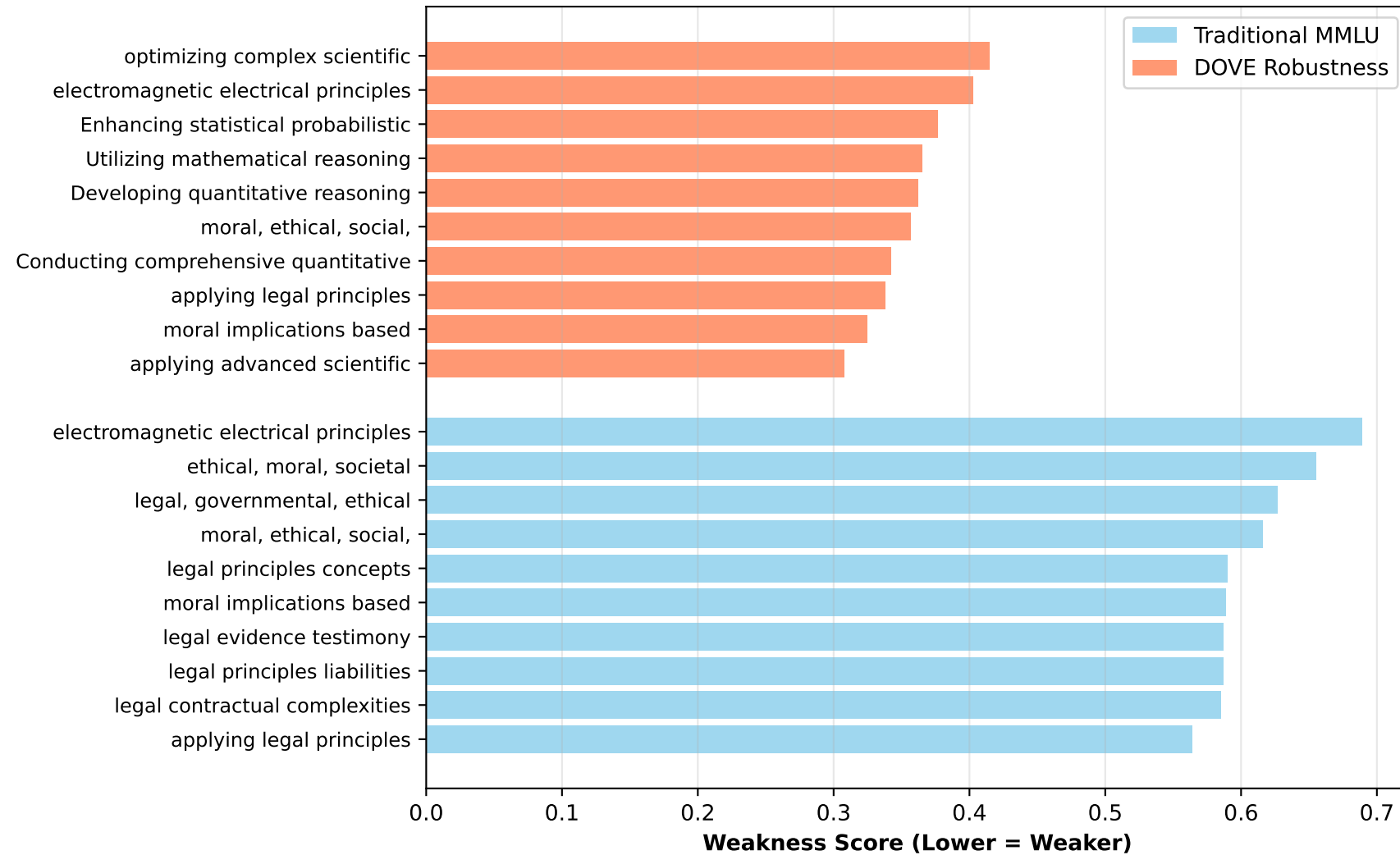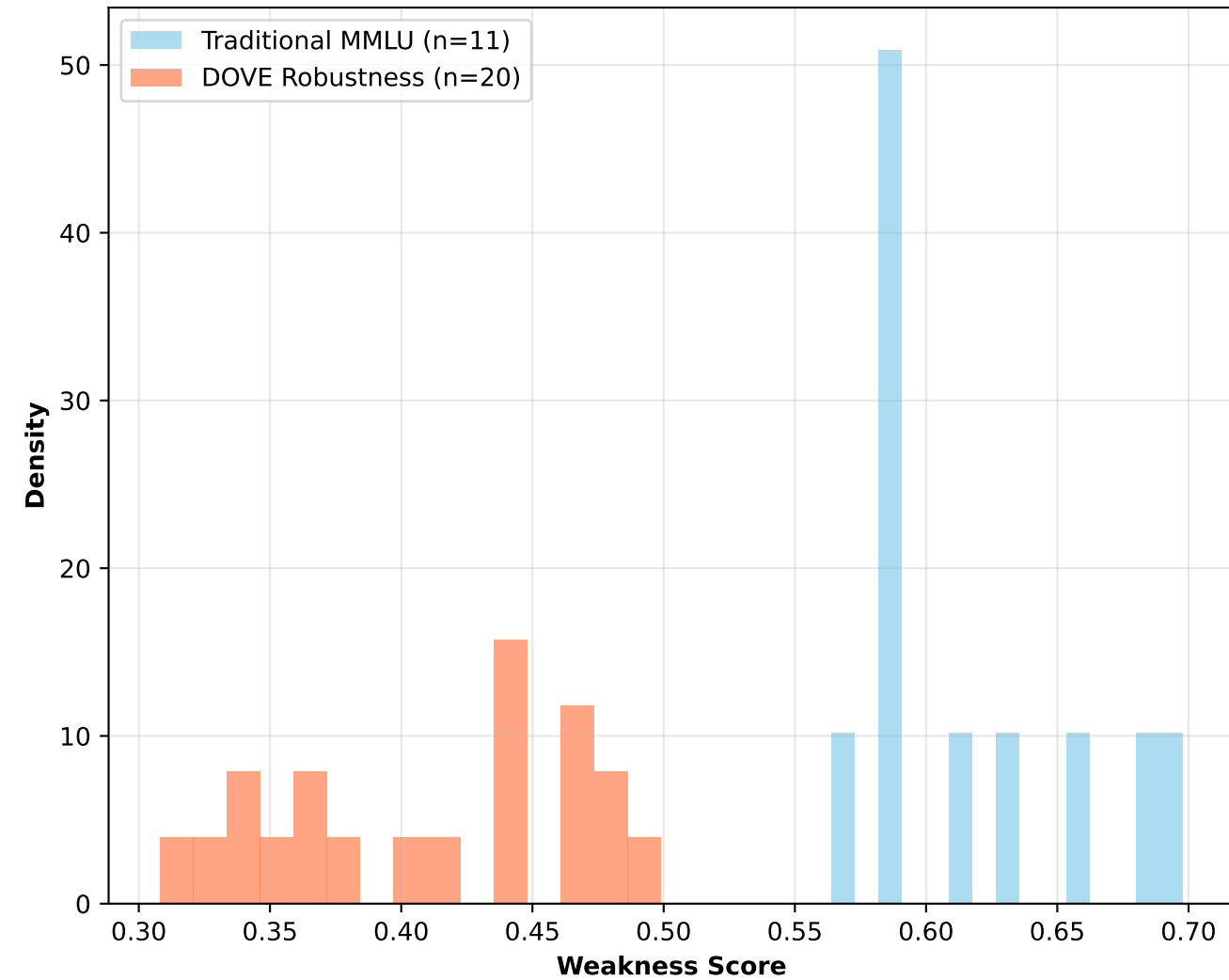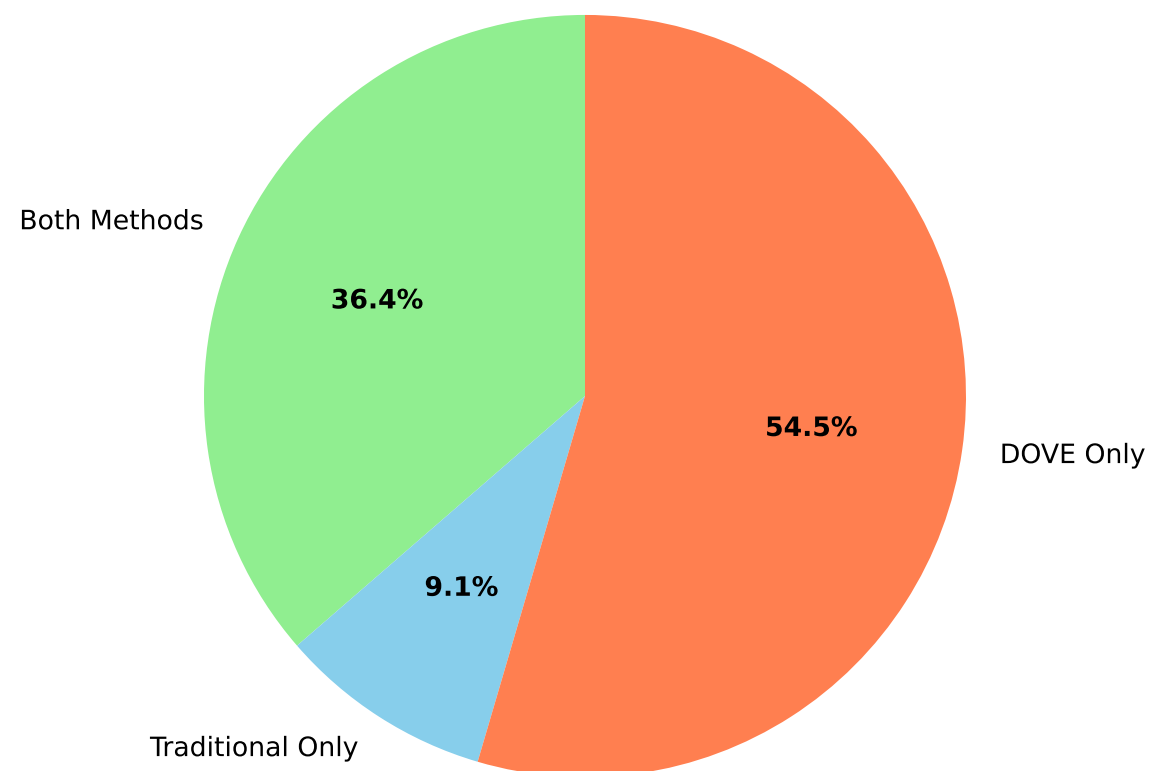
Top 10 Weakest Capabilities: Method Comparison

Legend: Traditional MMLU, DOVE Robustness

DOVE Robustness capabilities (top to bottom):
- optimizing complex scientific
- electromagnetic electrical principles
- Enhancing statistical probabilistic
- Utilizing mathematical reasoning
- Developing quantitative reasoning
- moral, ethical, social,
- Conducting comprehensive quantitative
- applying legal principles
- moral implications based
- applying advanced scientific

Traditional MMLU capabilities (top to bottom):
- electromagnetic electrical principles
- ethical, moral, societal
- legal, governmental, ethical
- moral, ethical, social,
- legal principles concepts
- moral implications based
- legal evidence testimony
- legal principles liabilities
- legal contractual complexities
- applying legal principles

X-axis: Weakness Score (Lower = Weaker)

Weakness Score Distributions

Legend: Traditional MMLU (n=11), DOVE Robustness (n=20)

Y-axis: Density
X-axis: Weakness Score

Weakness Detection: Key Insight

- Both Methods: 36.4%
- DOVE Only: 54.5%
- Traditional Only: 9.1%

Question Coverage by Weakness Detection Method

- Traditional MMLU: 6,015
- DOVE Robustness: 2,982

DOVE identifies 2,982 questions
in vulnerable capabilities
for targeted question generation

Y-axis: Total Questions in Weak Branches