Robustness in Weakness Profiling: Traditional vs DOVE Analysis **Top 10 Weakest Capabilities: Method Comparison Weakness Score Distributions** 16 -Traditional MMLU (n=0) DOVE Robustness (n=20) optimizing complex scientific 14 electromagnetic electrical principles Enhancing statistical probabilistic -12 Utilizing mathematical reasoning -10 Density Developing quantitative reasoning 8 moral, ethical, social, Conducting comprehensive quantitative -6 applying legal principles 4 moral implications based -2 Traditional MMLU applying advanced scientific **DOVE Robustness** 0.00 0.05 0.15 0.20 0.25 0.35 0.40 0.300 0.325 0.350 0.375 0.400 0.425 0.475 0.10 0.30 0.450 0.500 **Weakness Score (Lower = Weaker) Weakness Score Weakness Detection Method Overlap Question Coverage by Method** 2,982 3000 Tr**BoittinoMattOord**s 2500 Total Questions in Weak Branches 10000 10000 0.0% 100.0% 500 **DOVE Only** 0 DOVE Traditional MMLU Robustness