Stability improvement from additional runs (per model + average) Solid = predicted IID (sd\_all/vR). Dashed = empirical using all run pairs for R=2 and all runs for R=3. Thick black = model-averaged. 0.04 across runs model Claude-Opus-4 DeepSeek-V3-685B Average per-slice SE → GPT-4o → GPT-4o-mini Grok-3 Grok-4 LLaMA-3.3-70B o3-mini 0.01 Number of runs averaged (R)