

SAE Probes vs Best Baseline Method Across All Settings  
Mean improvement: 0.0757 | 77.9% improved | n=2361 datasets

