ACCURACY by Criterion (Basic Prompts)

Legend: GPT-4o-mini, GPT-4o, O3, GPT-5, Human

Evaluation Criteria: Random Exposure, Causal Mediation, Mediator-Outcome Linearity, Exposure-Mediator Interaction, Covariates Exp-Med, Covariates Exp-Out, Covariates Med-Out, Baseline Mediator, Baseline Outcome, Temporal Exp→Med, Temporal Med→Out, Assumptions Discussed, Sensitivity Analysis, Post-Exposure Control