

Where Models Fail (with 95% CIs, cluster-bootstrapped over problems)

Shares over generation attempts, aggregated per problem then bootstrapped by problem

