Average retries per problem (max = 15)Lower is better **Average retries** Level analogy-based ∧ base model Average runs per problem (max = 3)Lower is better Average runs Level