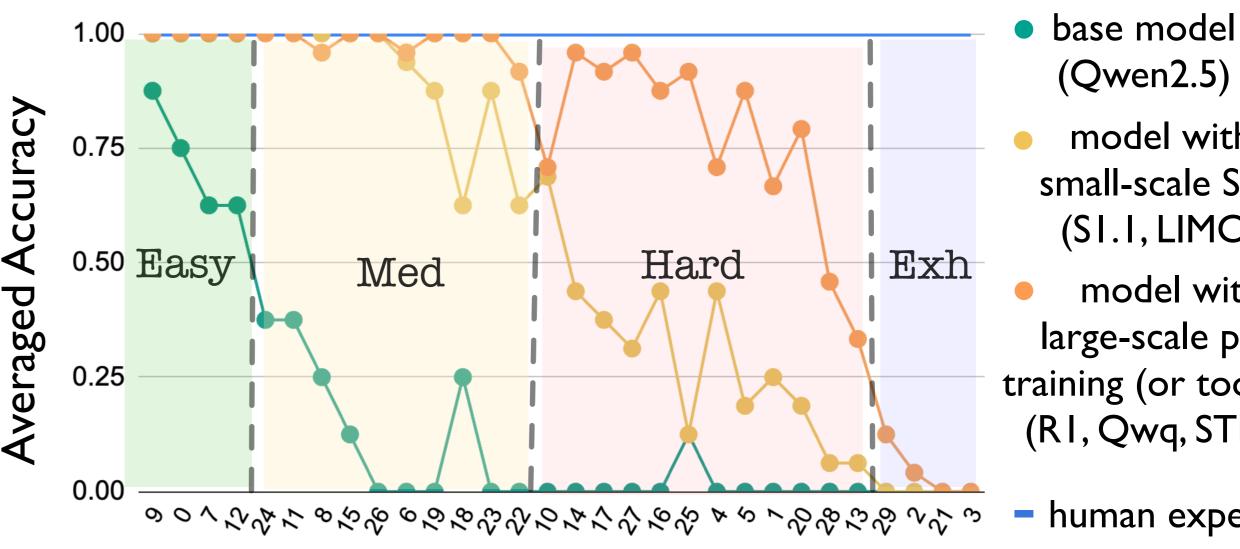


Accuracy Drops as Reasoning Difficulty Increases across AIME24 Questions



- (Qwen2.5)
- model with small-scale SFT (S1.1, LIMO)
- model with large-scale posttraining (or tool use) (RI, Qwq, STILL3)
- human expert

Sorted AIME24 Question IDs