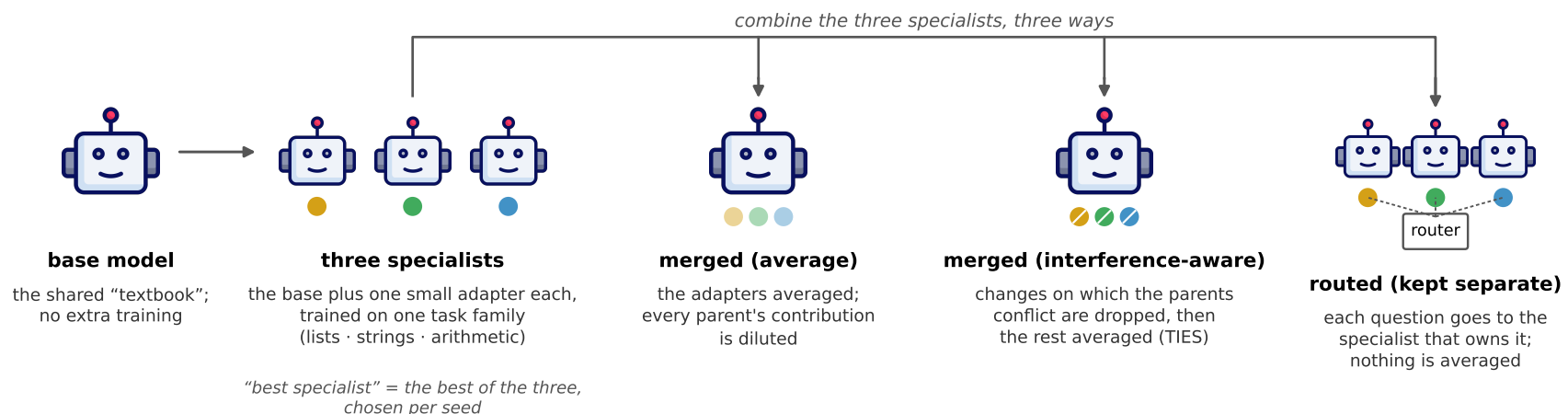


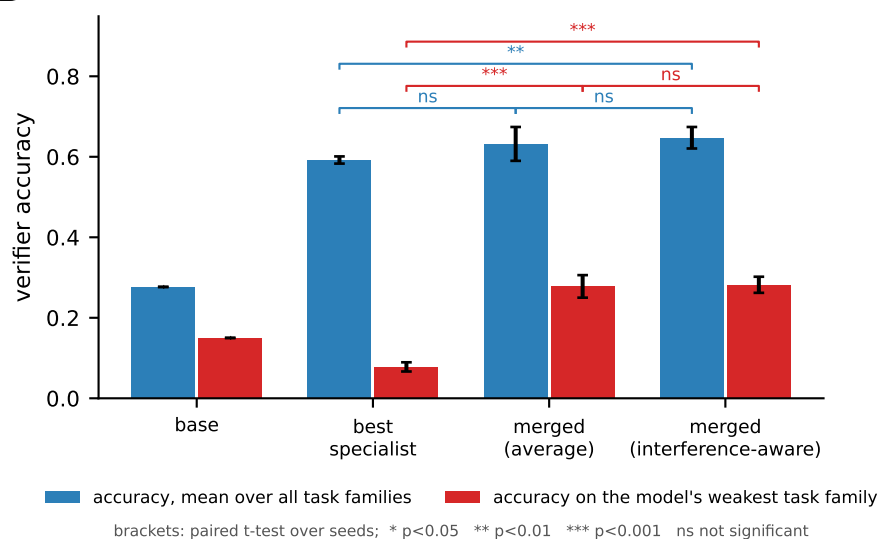
A How the models compared in B and C are built



All models share the same frozen base; only the small adapters differ. A verifier marks every answer right or wrong; accuracy is the share marked right.

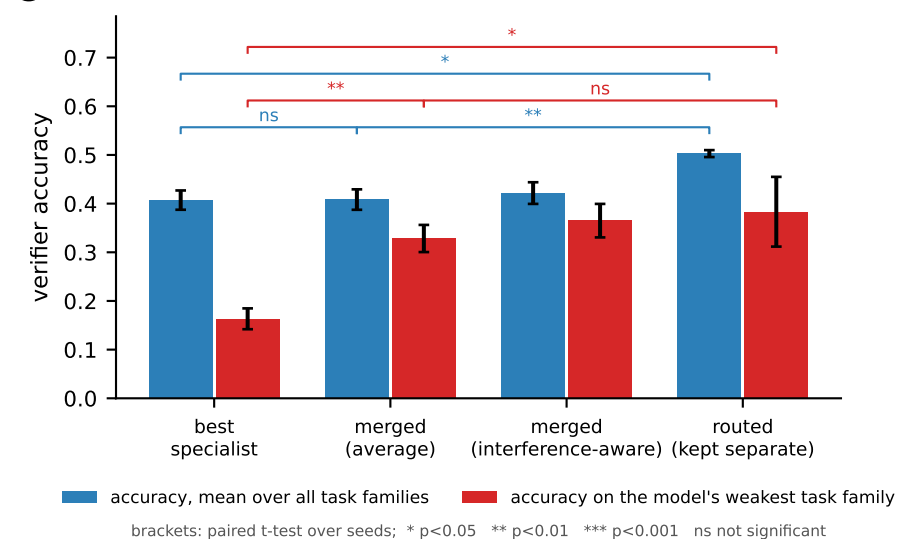
B Merged specialists beat the best single specialist

Qwen2.5-0.5B (0.5 billion parameters), easy tasks, 5 training seeds



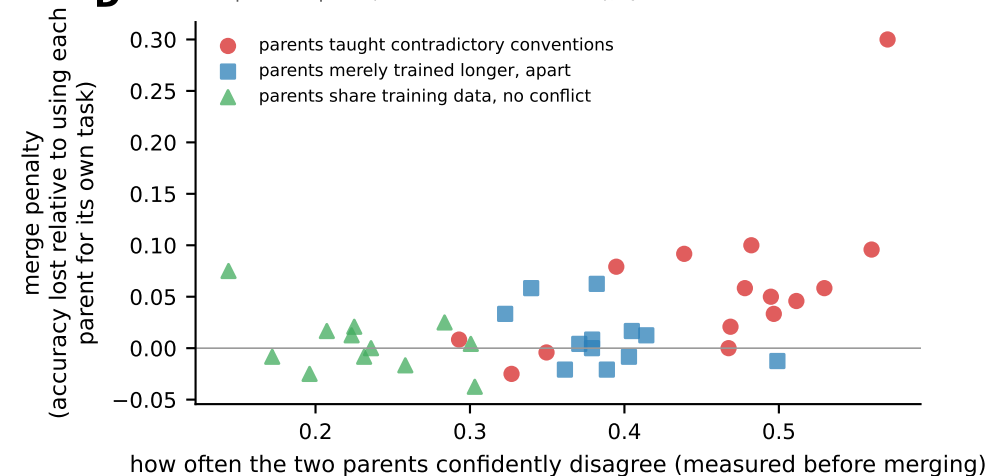
C On hard tasks, keeping specialists separate beats averaging them

Qwen2.5-7B (7 billion parameters), hard tasks, 3 training seeds



D Disagreement between parents, measured before merging, predicts merge damage

39 specialist pairs (13 conditions × 3 seeds), Qwen2.5-0.5B



E Behavioural measures predict the damage; weight-geometry measures do not

same 39 pairs; rank correlation with the merge penalty

