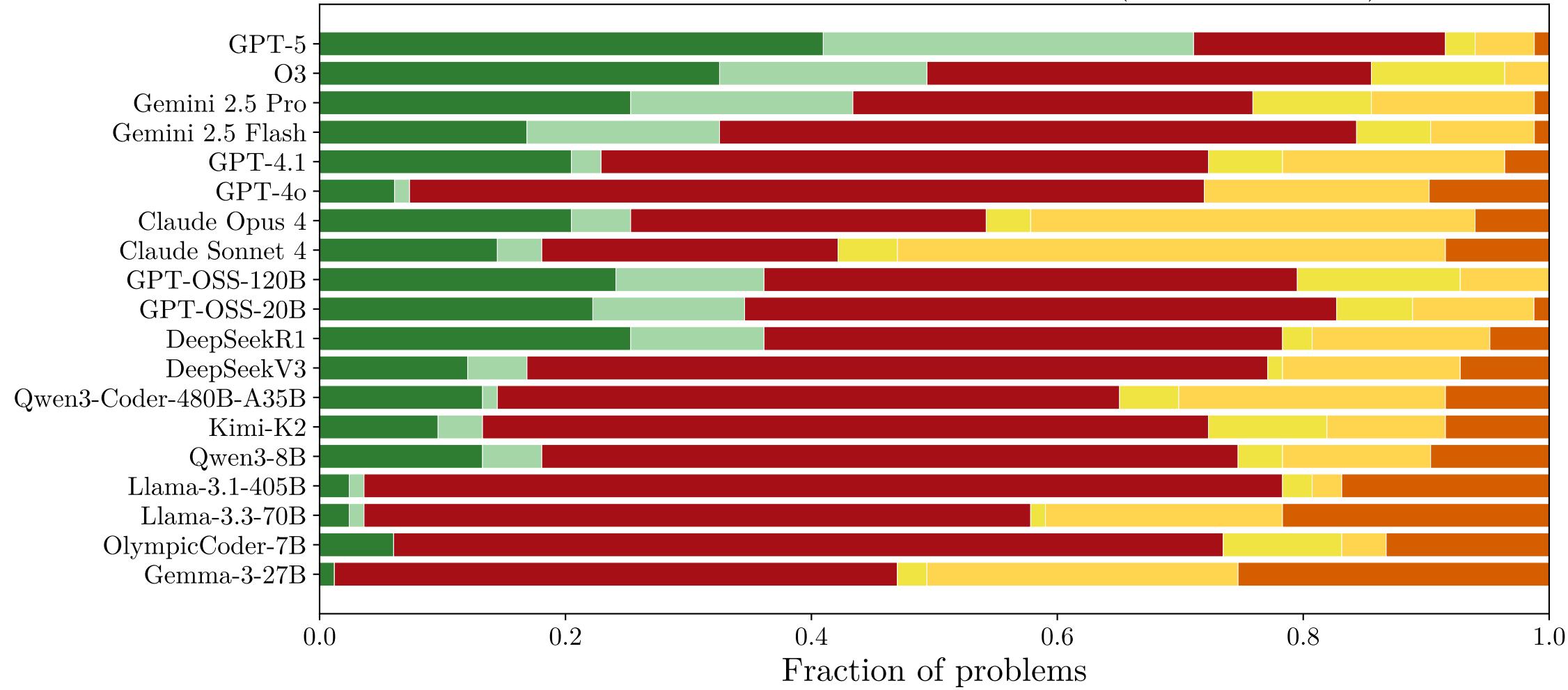


Editorial correctness breakdown by model (LLM-as-a-Judge)



Fraction of problems

LLM-judged category

Correct / Same as golden

Correct / Different from golden

Incorrect / Wrong algorithm

Incorrect / Correct algorithm but incorrect approach

Incorrect / Suboptimal but correct algorithm

Incorrect / Suboptimal and wrong algorithm