

Question difficulty distribution across benchmarks  
with model abilities overlaid

