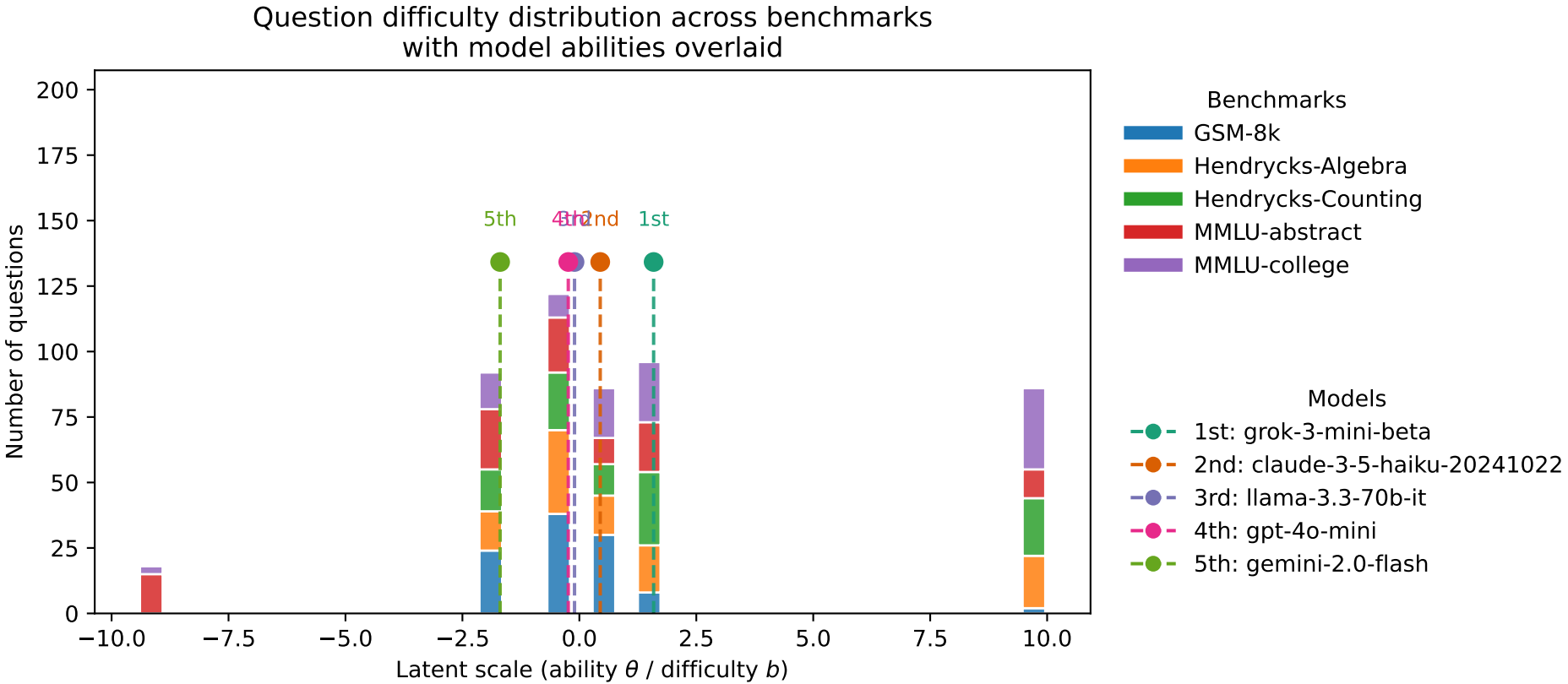
Question difficulty distribution across benchmarks
with model abilities overlaid
Number of questions
Latent scale (ability θ / difficulty b)
200
175
150
125
100
75
50
25
0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
5th
4th
3rd
2nd
1st
Benchmarks
GSM-8k
Hendrycks-Algebra
Hendrycks-Counting
MMLU-abstract
MMLU-college
Models
1st: grok-3-mini-beta
2nd: claude-3-5-haiku-20241022
3rd: llama-3.3-70b-it
4th: gpt-4o-mini
5th: gemini-2.0-flash