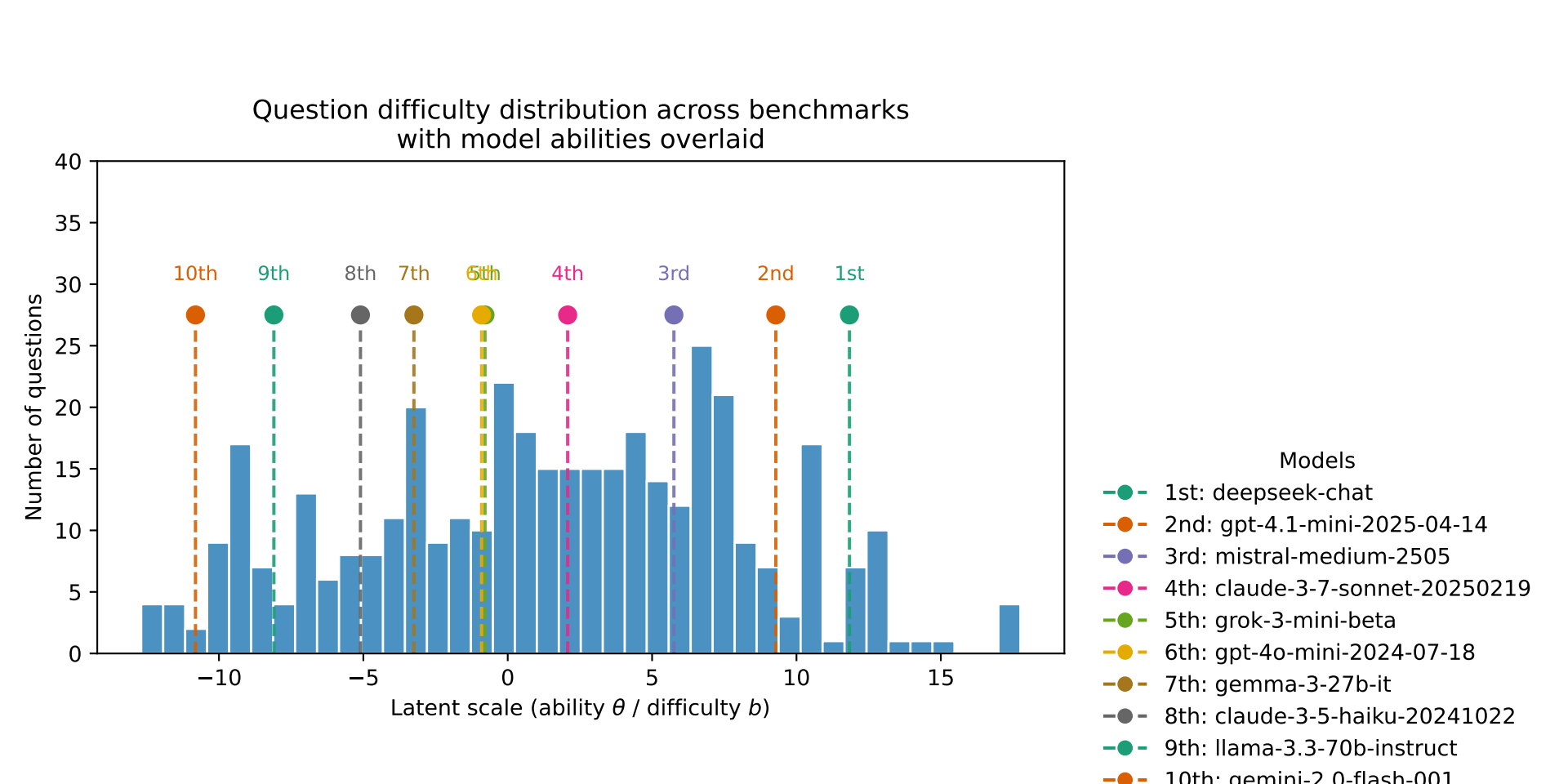
Question difficulty distribution across benchmarks
with model abilities overlaid
Number of questions
Latent scale (ability θ / difficulty b)
10th
9th
8th
7th
6th
5th
4th
3rd
2nd
1st
Models
1st: deepseek-chat
2nd: gpt-4.1-mini-2025-04-14
3rd: mistral-medium-2505
4th: claude-3-7-sonnet-20250219
5th: grok-3-mini-beta
6th: gpt-4o-mini-2024-07-18
7th: gemma-3-27b-it
8th: claude-3-5-haiku-20241022
9th: llama-3.3-70b-instruct
10th: gemini-2.0-flash-001