Per-Question Output Tokens with Difficulty
(Token lines: sorted by own tokens; Difficulty: sorted by difficulty)

Legend:
- Gemini 3 Pro (mean=20.2k)
- Gemini 3 Flash (mean=36.2k)
- GPT-5.2 High (mean=19.5k)
- GPT-5-mini High (mean=29.3k)
- Difficulty (sorted, 15 models)

X-axis: Question Rank (each line sorted independently)
Left Y-axis: Output Tokens (k)
Right Y-axis: Difficulty (1 - Accuracy)