F Difficulty-Based Performance Analysis

Llama 3.1	Easy	Med	Hard
Visual	6.61	6.36	4.64
Data	9.39	8.05	6.53
Code	6.12	4.94	3.47
Avg	**7.37**	**6.45**	**4.88**

Gemma 2	Easy	Med	Hard
Visual	6.43	5.91	4.11
Data	9.39	8.31	6.94
Code	6.12	4.94	3.67
Avg	**7.31**	**6.39**	**4.91**

Qwen2.5	Easy	Med	Hard
Visual	8.39	7.73	6.43
Data	5.92	5.84	5.31
Code	7.82	7.14	5.99
Avg	**7.38**	**6.90**	**5.91**

Claude 3.5	Easy	Med	Hard
Visual	7.32	6.48	7.14
Data	8.98	7.92	7.96
Code	7.55	5.84	7.14
Avg	**7.95**	**6.75**	**7.41**

Deepseek	Easy	Med	Hard
Visual	6.96	6.82	7.32
Data	9.59	8.31	7.96
Code	6.94	6.36	7.14
Avg	**7.83**	**7.16**	**7.47**

Gemini 2.0	Easy	Med	Hard
Visual	7.14	6.70	7.50
Data	8.16	7.79	7.35
Code	6.94	5.71	5.51
Avg	**7.41**	**6.74**	**6.79**

GPT-4o	Easy	Med	Hard
Visual	6.79	6.25	5.54
Data	8.57	7.79	7.55
Code	6.94	5.45	5.92
Avg	**7.43**	**6.50**	**6.34**

O1-High	Easy	Med	Hard
Visual	7.86	7.27	7.50
Data	9.80	8.70	8.57
Code	8.16	7.01	7.14
Avg	**8.61**	**7.66**	**7.74**

O1-High with Additional Context

	Easy	Med	Hard
Visual	8.75	7.50	7.32
Data	9.80	8.70	8.37
Code	8.16	6.62	6.53
Average	8.90	7.61	7.41

Table 16: Detailed breakdown of O1-High + Additional Context scores by difficulty level.

Claude 3.7 Sonnet Detailed Scores by Difficulty

	Easy	Med	Hard
Visual	8.39	8.75	9.64
Data	9.59	9.87	10.00
Code	9.39	9.09	10.00
Average	9.12	9.24	9.88

Table 17: Breakdown of Claude-3.7 Sonnet with 25 Iterations (no additional context).

	Easy	Med	Hard
Visual	9.11	9.20	8.75
Data	10.00	9.87	9.18
Code	9.80	9.61	8.78
Average	9.63	9.56	8.90

Table 18: Breakdown of Claude-3.7 Sonnet with Additional Context and 25 Iterations.