RedPenBench Results

5th of June 2026

Judge: openai/o4-mini (judge-v0.3-100pt) · 20 items × 3 runs · 11 models · Scores are LLM-judged

All models evaluated with Reasoning OFF and Reasoning ON (medium)

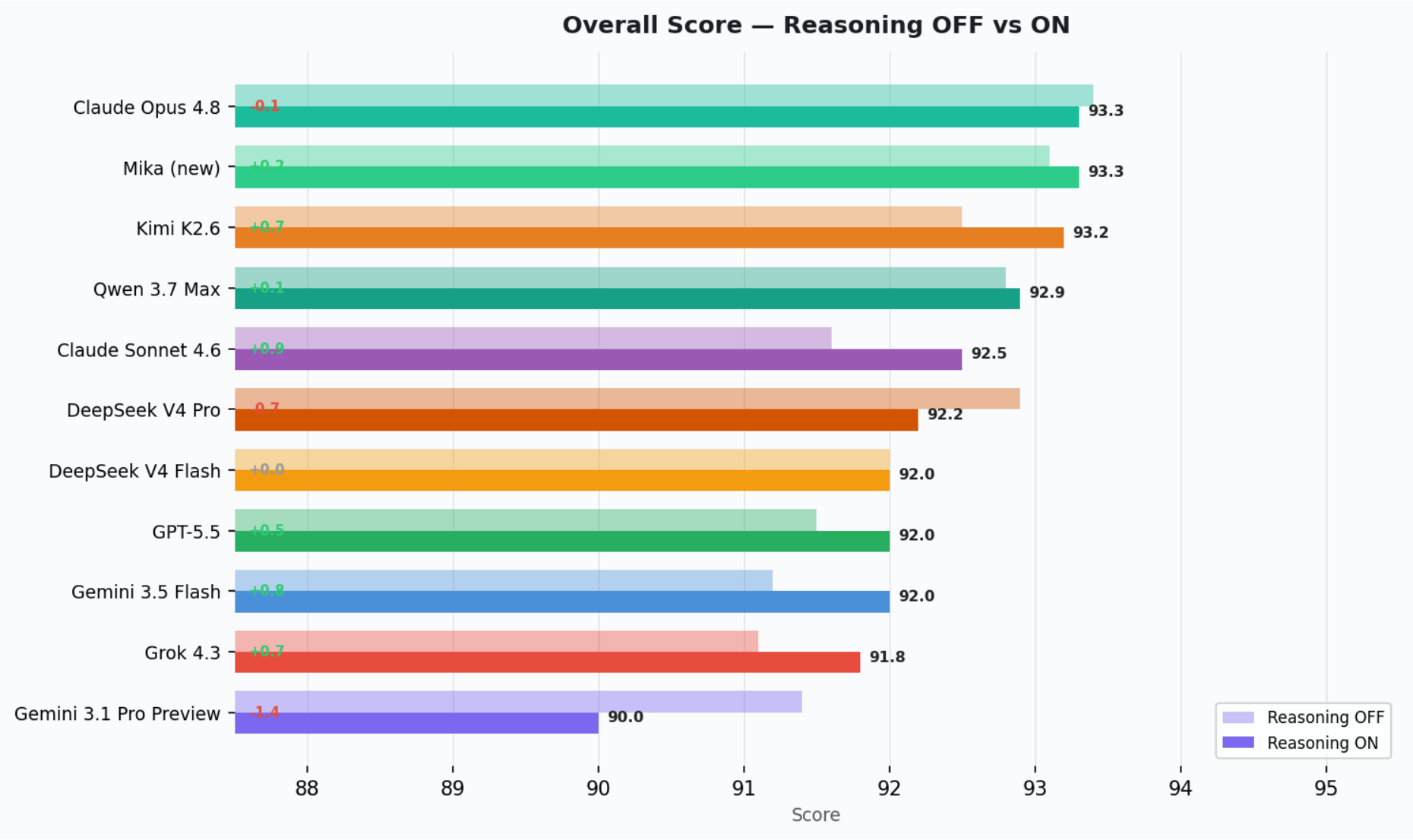
Summary

All 11 models ranked by Reasoning ON score. Mika row highlighted. Delta = ON minus OFF score.

#	Model	OFF Score	ON Score	Delta	CI OFF	CI ON	Cost OFF	Cost ON	Speed OFF	Speed ON	Flags OFF	Flags ON
1	Mika (new)	93.1	**93.3**	+0.20	92.8–93.4	92.9–93.7	Proprietary	Proprietary	170	186	0	0
2	Claude Opus 4.8	93.4	**93.3**	-0.10	92.9–93.8	92.9–93.7	$1.060	$1.190	65	66	1	0
3	Kimi K2.6	92.5	**93.2**	+0.70	91.1–93.5	92.6–93.7	$1.010	$0.700	255	257	2	1
4	Qwen 3.7 Max	92.8	**92.9**	+0.10	92.4–93.2	92.5–93.3	$0.470	$0.470	69	72	1	0
5	Claude Sonnet 4.6	91.6	**92.5**	+0.90	90.4–93.0	92.0–93.0	$0.470	$1.370	43	60	2	0
6	DeepSeek V4 Pro	92.9	**92.2**	-0.70	92.4–93.4	91.5–92.9	$0.580	$0.260	82	76	0	2
7	Gemini 3.5 Flash	91.2	**92.0**	+0.80	89.8–92.3	91.2–92.5	$0.970	$0.680	154	151	2	1
8	GPT-5.5	91.5	**92.0**	+0.50	90.6–92.4	91.1–92.8	$1.470	$1.200	52	52	3	2
9	DeepSeek V4 Flash	92.0	**92.0**	+0.00	91.2–92.6	91.4–92.4	$0.016	$0.016	89	119	2	1
10	Grok 4.3	91.1	**91.8**	+0.70	89.8–92.0	91.0–92.5	$0.370	$0.120	124	186	1	1
11	Gemini 3.1 Pro Preview	91.4	**90.0**	-1.40	90.4–92.1	88.6–91.2	$1.160	$0.660	84	80	1	3

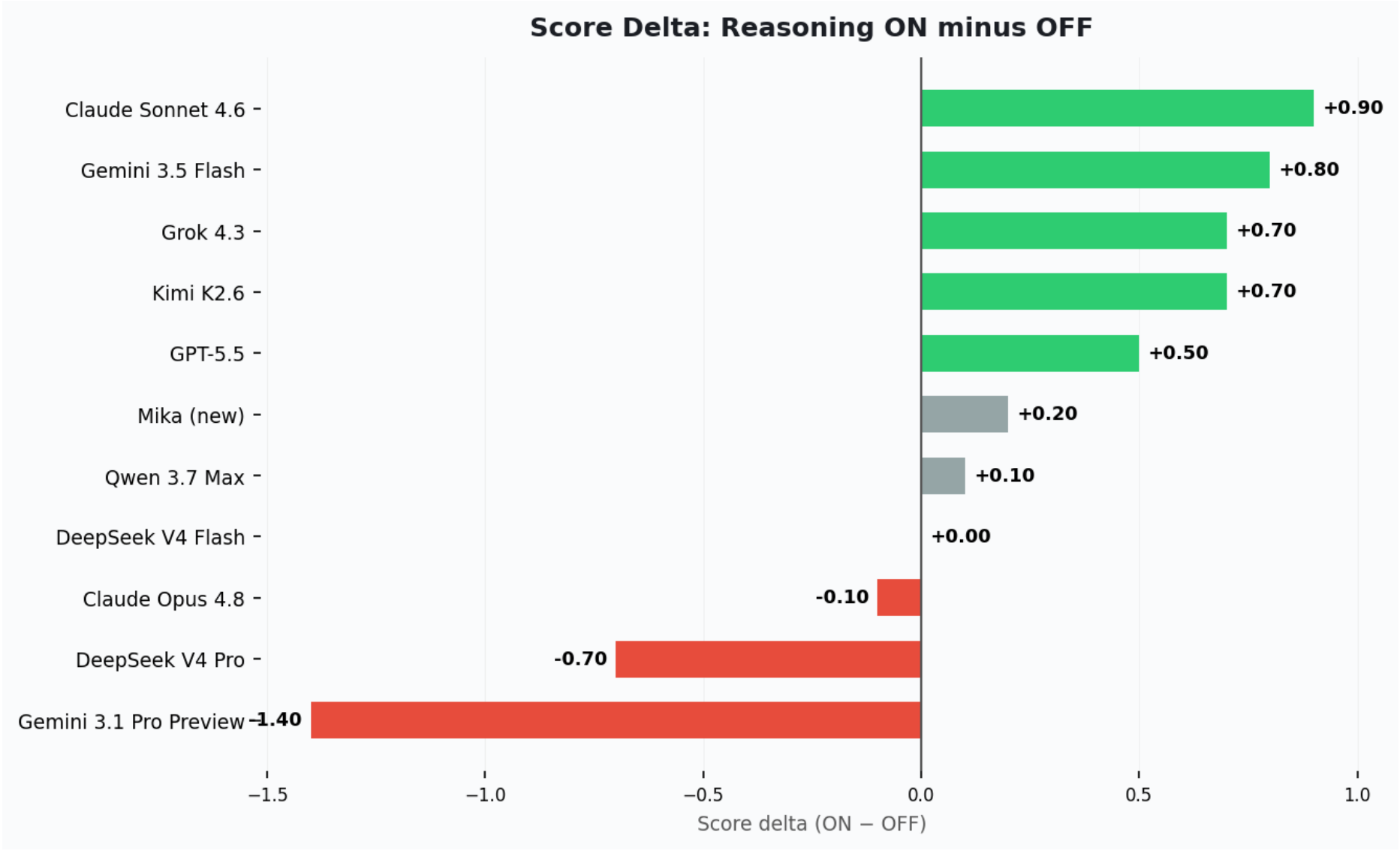
1. Overall Score — Reasoning OFF vs ON

Side-by-side comparison of all models. Faded bars = reasoning off, solid = reasoning on. Delta labels shown on left.


Overall Score — Reasoning OFF vs ON
Claude Opus 4.8
-0.1
93.3
Mika (new)
+0.2
93.3
Kimi K2.6
+0.7
93.2
Qwen 3.7 Max
+0.1
92.9
Claude Sonnet 4.6
+0.9
92.5
DeepSeek V4 Pro
-0.7
92.2
DeepSeek V4 Flash
+0.0
92.0
GPT-5.5
+0.5
92.0
Gemini 3.5 Flash
+0.8
92.0
Grok 4.3
+0.7
91.8
Gemini 3.1 Pro Preview
-1.4
90.0
Reasoning OFF
Reasoning ON
88
89
90
91
92
93
94
95
Score

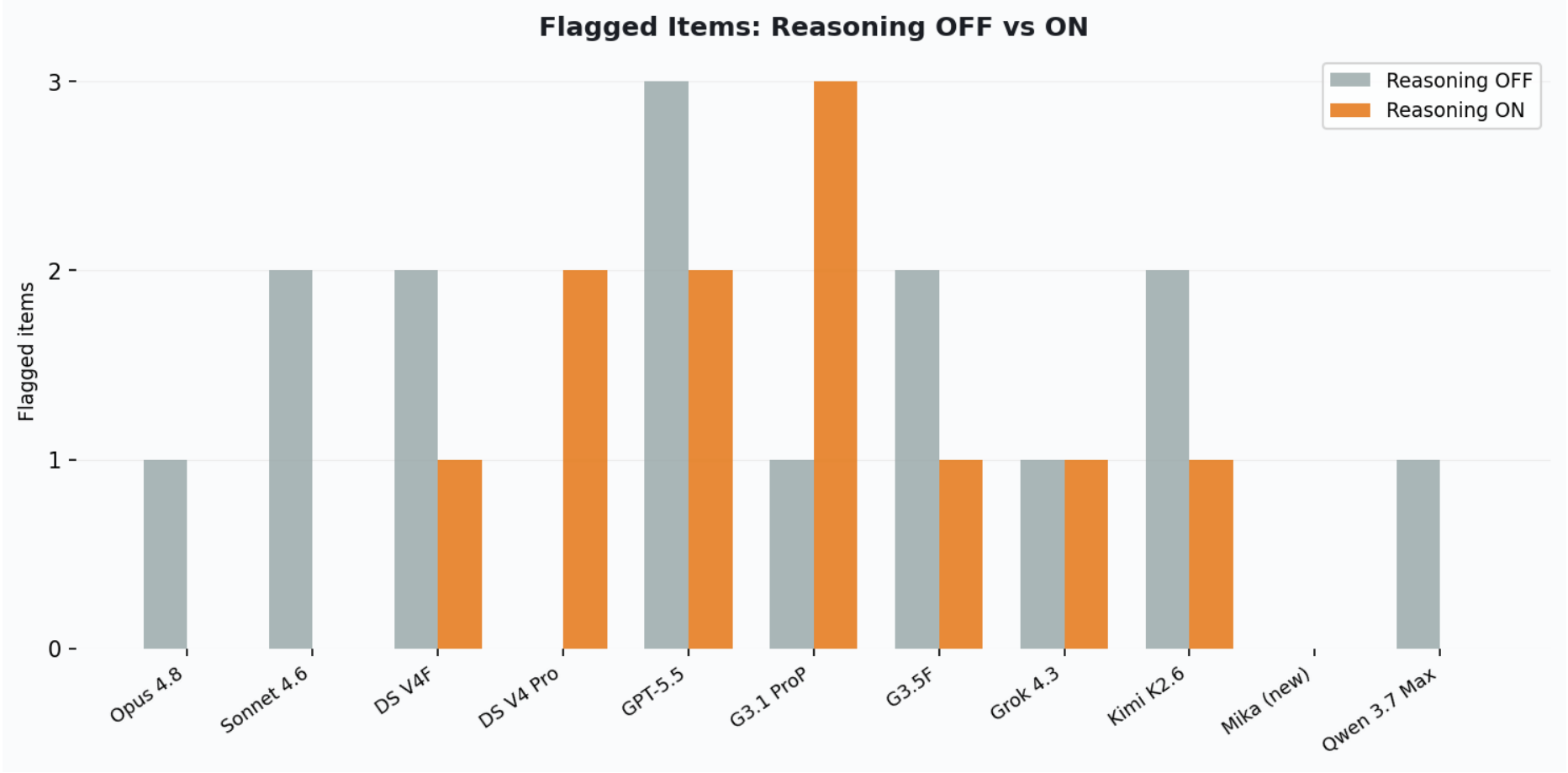
2. Impact of Reasoning — Score Delta

Gemini 3.1 Pro Preview shows the largest regression (−1.4). Claude Sonnet 4.6 shows the largest gain (+0.9), with reasoning fixing its systematic MATH-1017 numerical error. Mika, Opus, and Qwen are the most reasoning-stable models.


Score Delta: Reasoning ON minus OFF
Claude Sonnet 4.6
Gemini 3.5 Flash
Grok 4.3
Kimi K2.6
GPT-5.5
Mika (new)
Qwen 3.7 Max
DeepSeek V4 Flash
Claude Opus 4.8
DeepSeek V4 Pro
Gemini 3.1 Pro Preview
+0.90
+0.80
+0.70
+0.70
+0.50
+0.20
+0.10
+0.00
-0.10
-0.70
-1.40
−1.5
−1.0
−0.5
0.0
0.5
1.0
Score delta (ON − OFF)

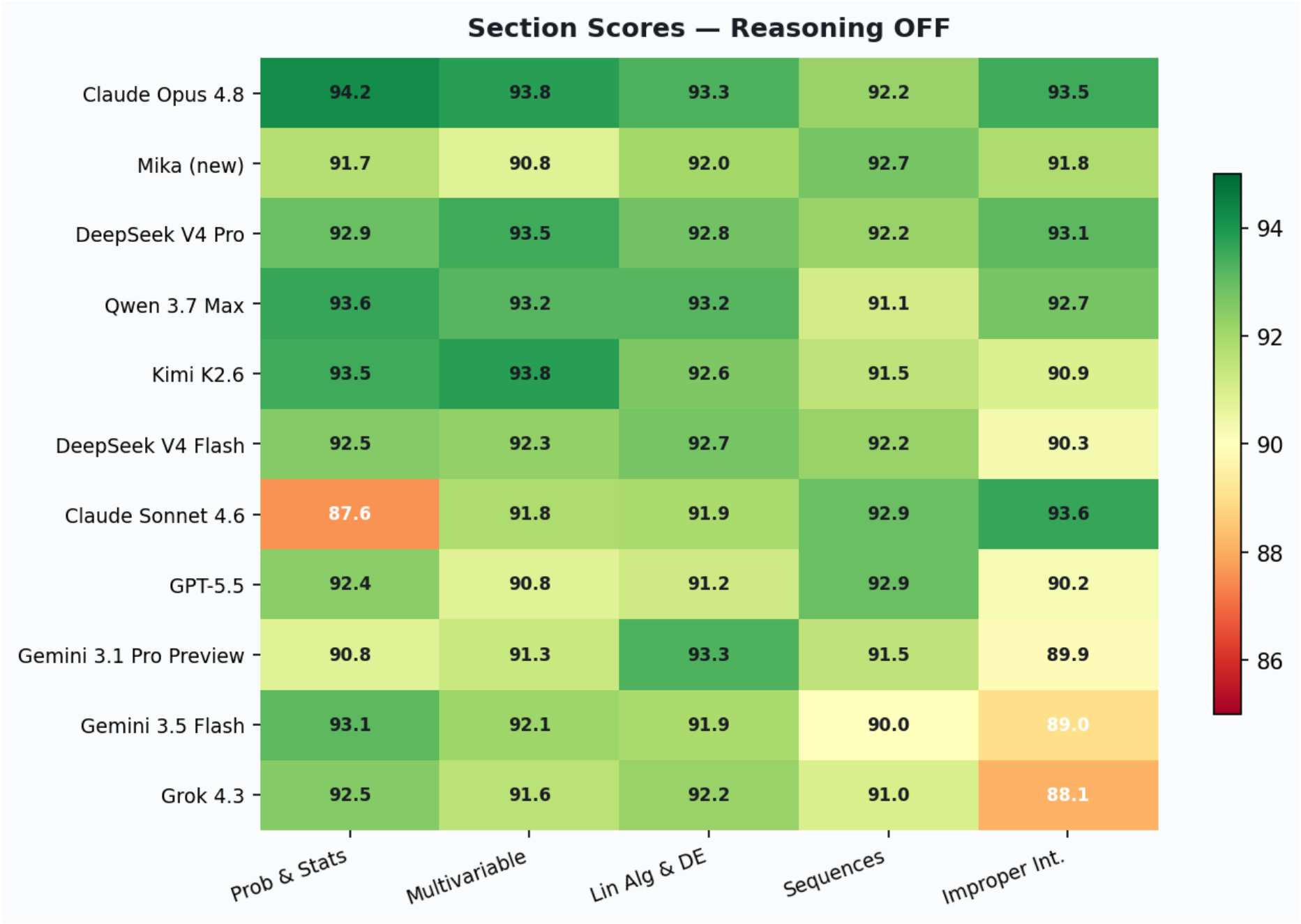
3. Flagged Items — Reasoning OFF vs ON

Flagged = item average below 88 or run variance ≥ 10. Total flags drop from 17 (OFF) to 11 (ON). DeepSeek V4 Pro loses its zero-flag status; Sonnet 4.6, Qwen, and Opus gain it.



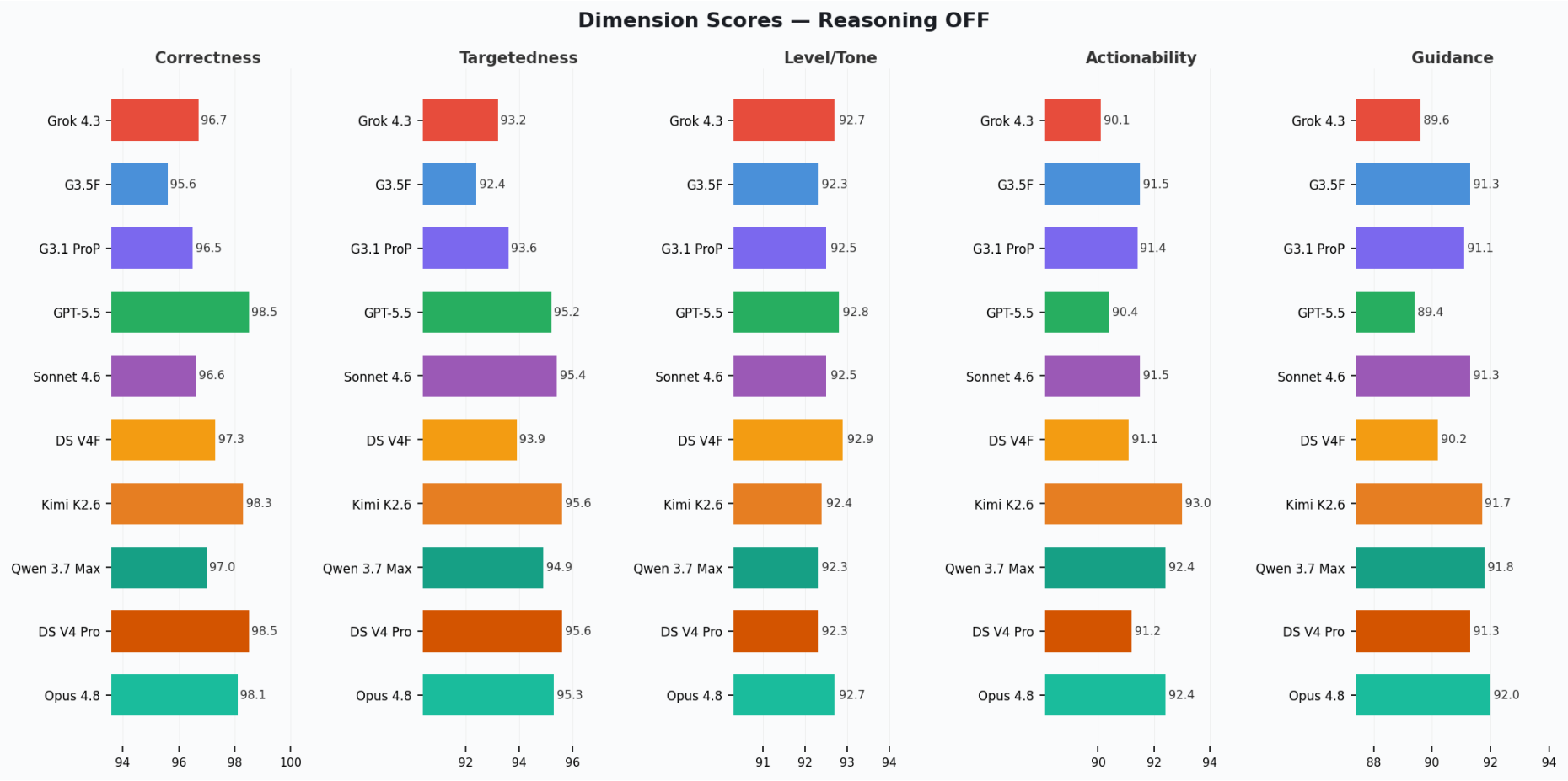
4. Section Scores

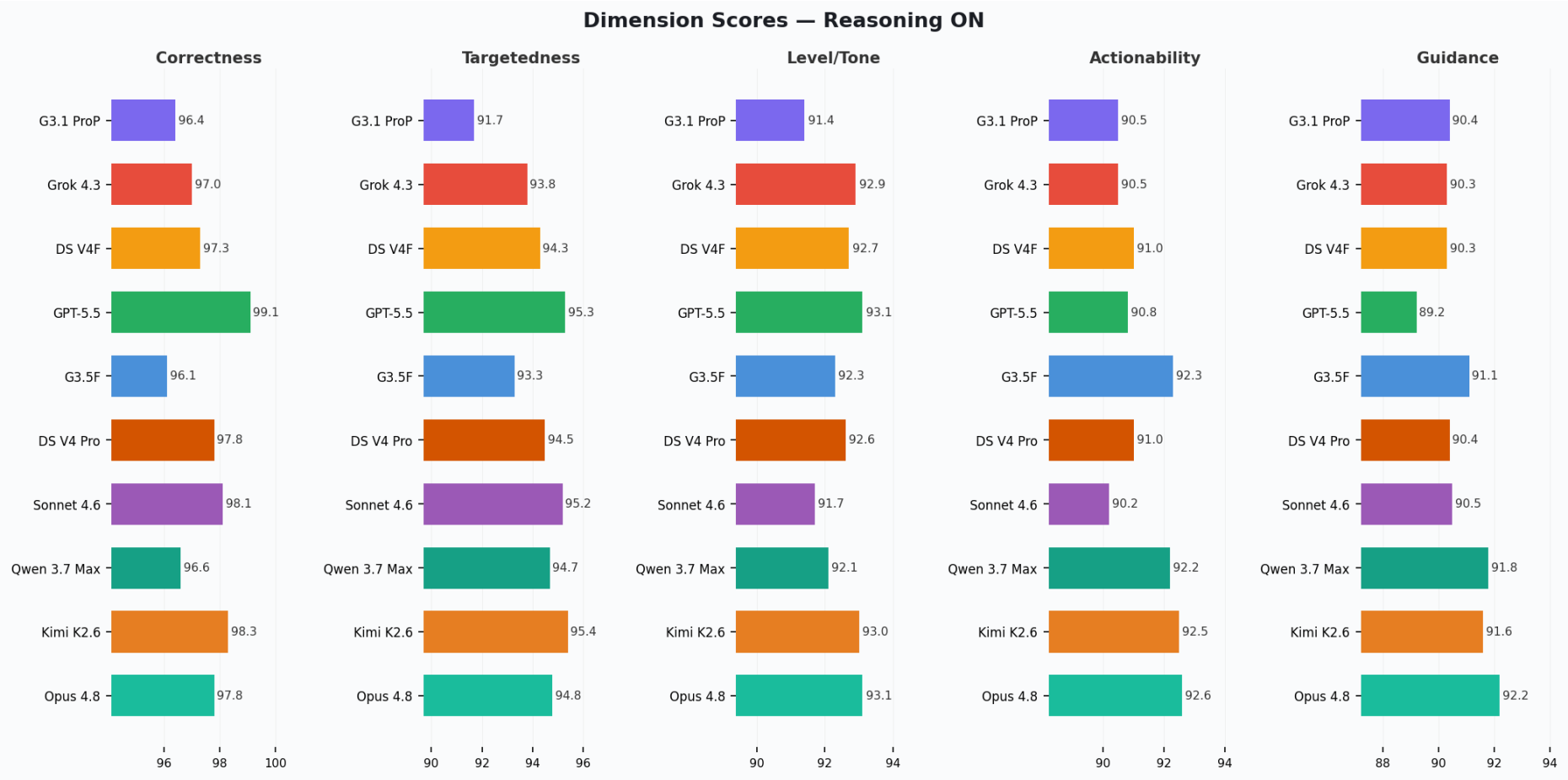
Improper Integrals & Limits improves most consistently with reasoning on — it was the weakest section for most models reasoning-off.


Section Scores — Reasoning OFF
Claude Opus 4.8
Mika (new)
DeepSeek V4 Pro
Qwen 3.7 Max
Kimi K2.6
DeepSeek V4 Flash
Claude Sonnet 4.6
GPT-5.5
Gemini 3.1 Pro Preview
Gemini 3.5 Flash
Grok 4.3
94.2 93.8 93.3 92.2 93.5
91.7 90.8 92.0 92.7 91.8
92.9 93.5 92.8 92.2 93.1
93.6 93.2 93.2 91.1 92.7
93.5 93.8 92.6 91.5 90.9
92.5 92.3 92.7 92.2 90.3
87.6 91.8 91.9 92.9 93.6
92.4 90.8 91.2 92.9 90.2
90.8 91.3 93.3 91.5 89.9
93.1 92.1 91.9 90.0 89.0
92.5 91.6 92.2 91.0 88.1
Prob & Stats
Multivariable
Lin Alg & DE
Sequences
Improper Int.
94
92
90
88
86

5. Dimension Scores

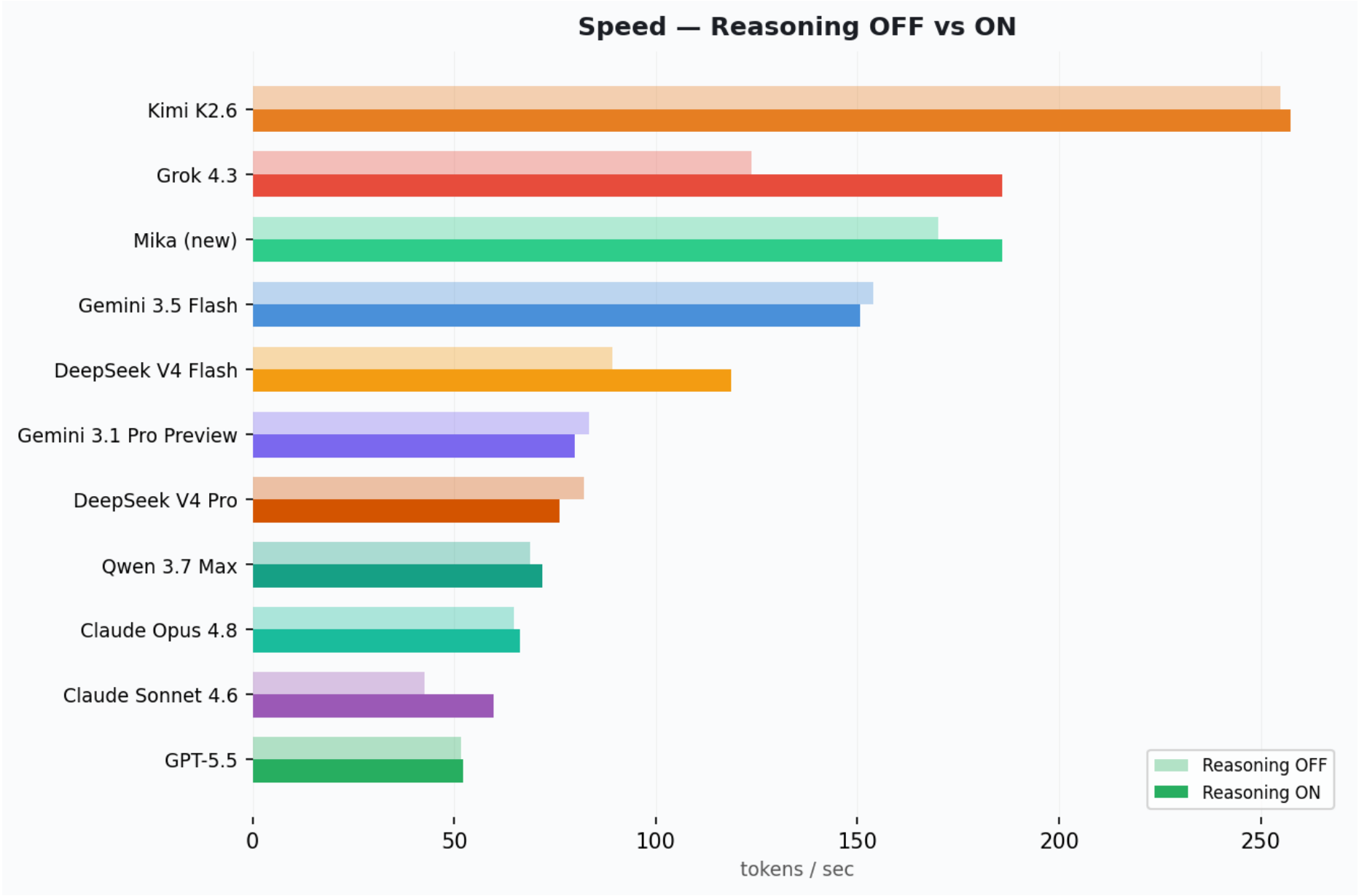
Correctness improves broadly with reasoning on. Guidance and Actionability remain the universally weakest dimensions — reasoning does not fix the scaffolding gap.





6. Speed

Kimi K2.6 is fastest in both conditions (255/257 tok/s). Grok 4.3 and Mika both jump to 186 tok/s with reasoning on. Claude Sonnet 4.6 and Opus 4.8 are the slowest. Mika cost is proprietary.

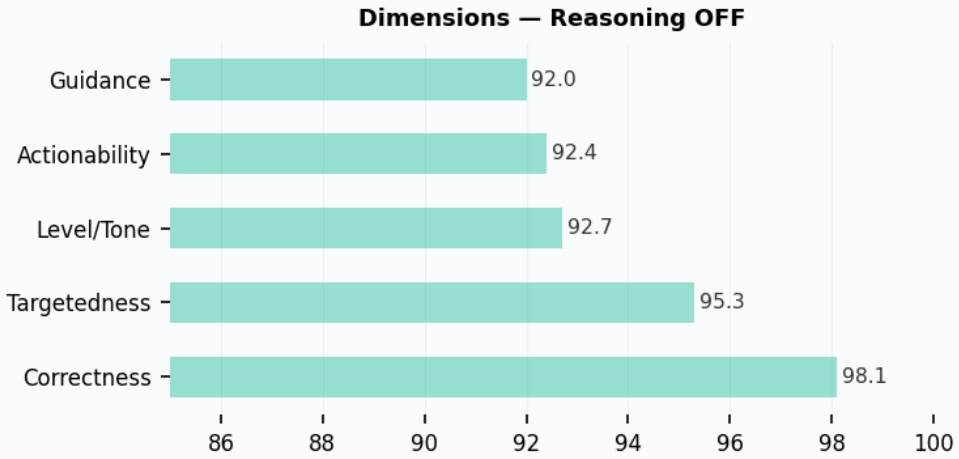

Speed — Reasoning OFF vs ON
Kimi K2.6
Grok 4.3
Mika (new)
Gemini 3.5 Flash
DeepSeek V4 Flash
Gemini 3.1 Pro Preview
DeepSeek V4 Pro
Qwen 3.7 Max
Claude Opus 4.8
Claude Sonnet 4.6
GPT-5.5
Reasoning OFF
Reasoning ON
0
50
100
150
200
250
tokens / sec

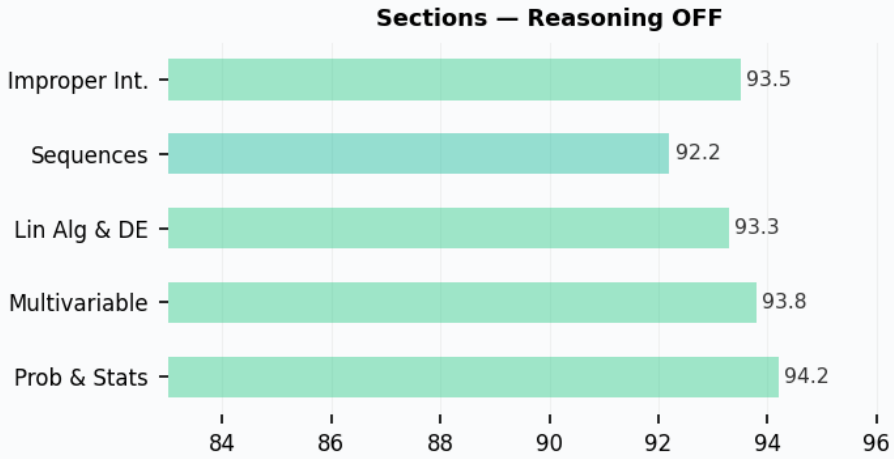
7. Individual Model Cards

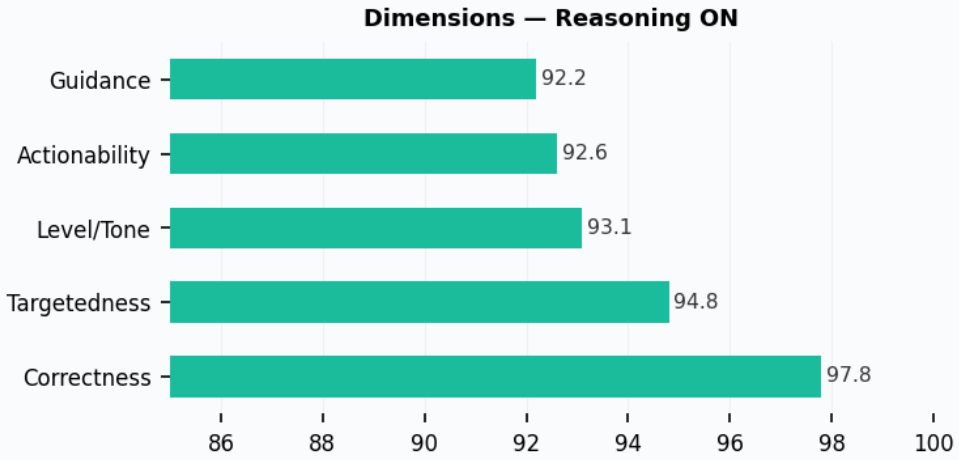
Each card shows dimension and section scores for both reasoning conditions.

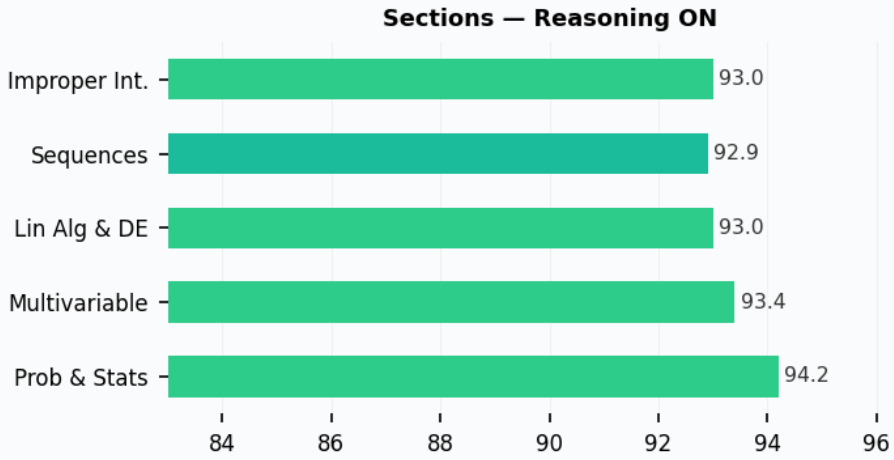
Claude Opus 4.8

OFF **93.4** → ON **93.3** **-0.10**





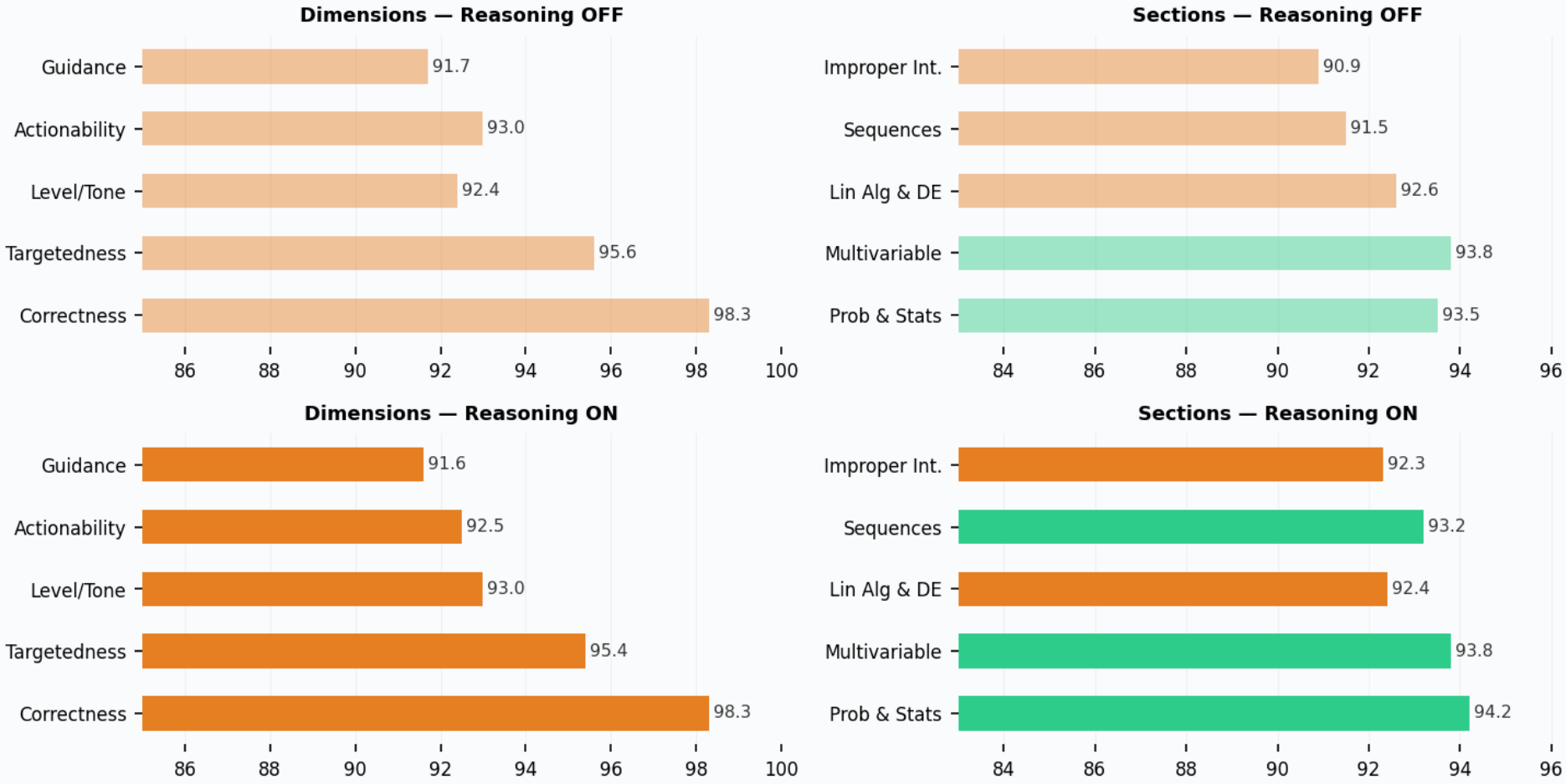




Cost OFF: $1.060 | ON: $1.190 Speed OFF: 65 tok/s | ON: 66 tok/s Flags OFF: 1 | ON: 0

Kimi K2.6

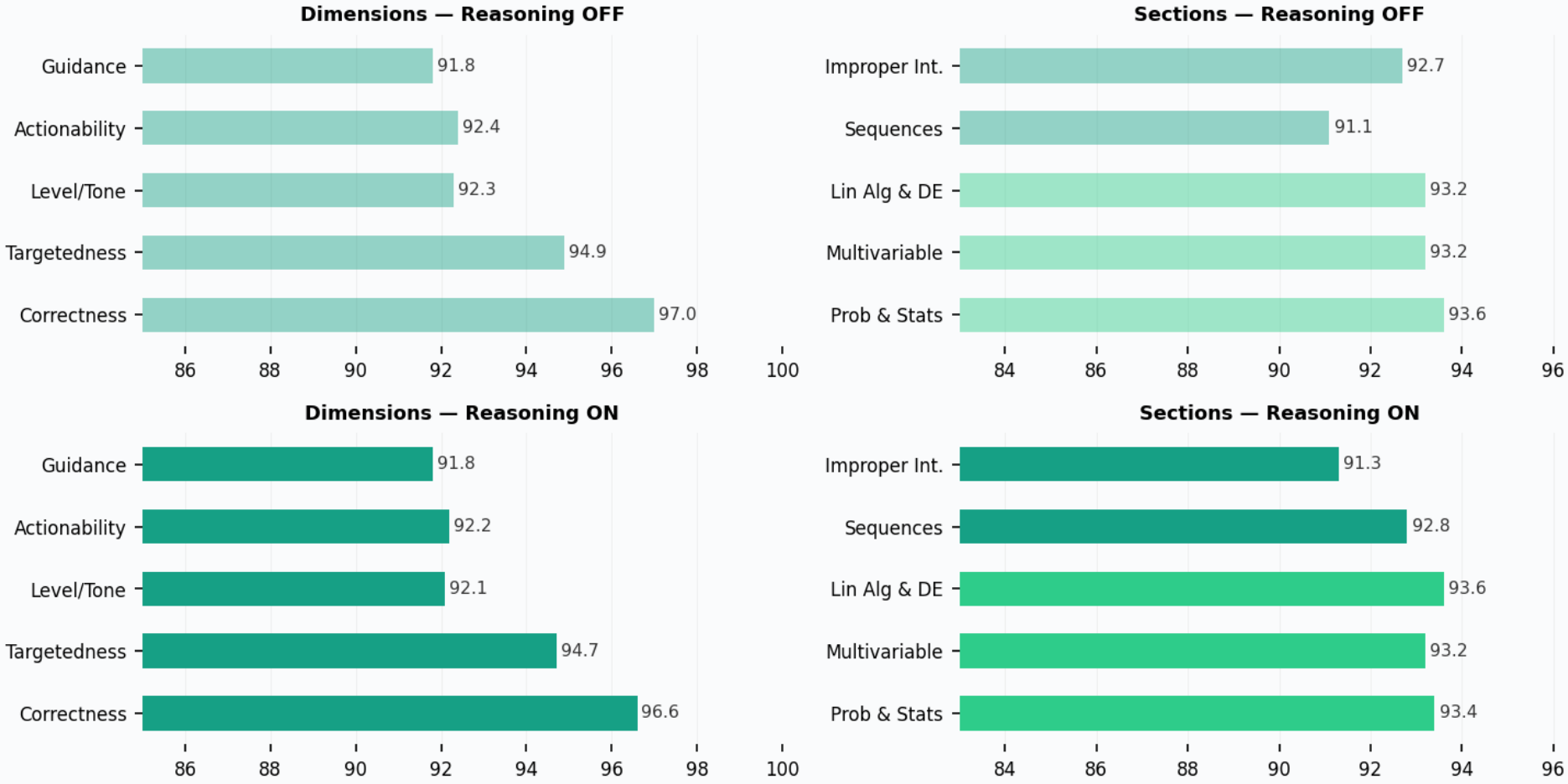
OFF **92.5** → ON **93.2** **+0.70**



Cost OFF: $1.010 | ON: $0.700 Speed OFF: 255 tok/s | ON: 257 tok/s Flags OFF: 2 | ON: 1

Qwen 3.7 Max

OFF **92.8** → ON **92.9** **+0.10**



Cost OFF: $0.470 | ON: $0.470 Speed OFF: 69 tok/s | ON: 72 tok/s Flags OFF: 1 | ON: 0

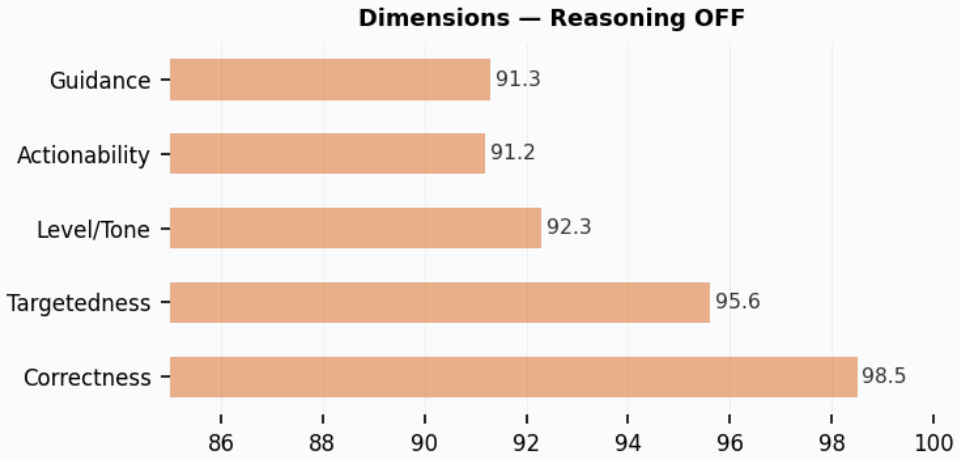
Claude Sonnet 4.6

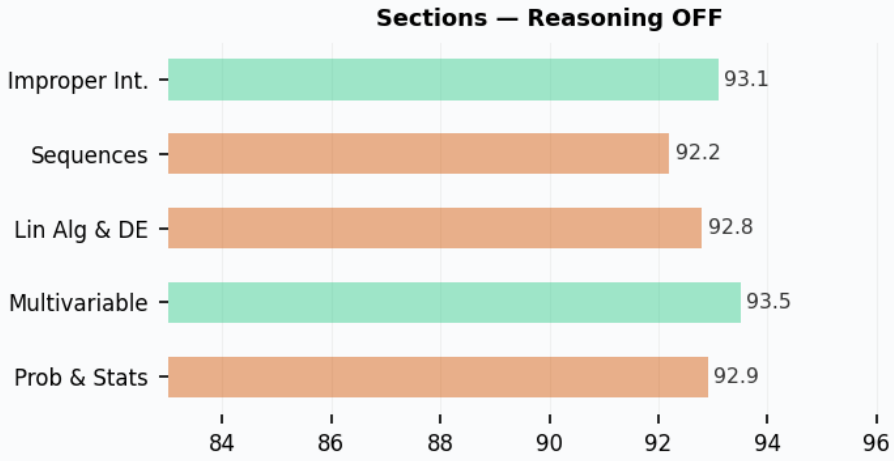
OFF **91.6** → ON **92.5** **+0.90**

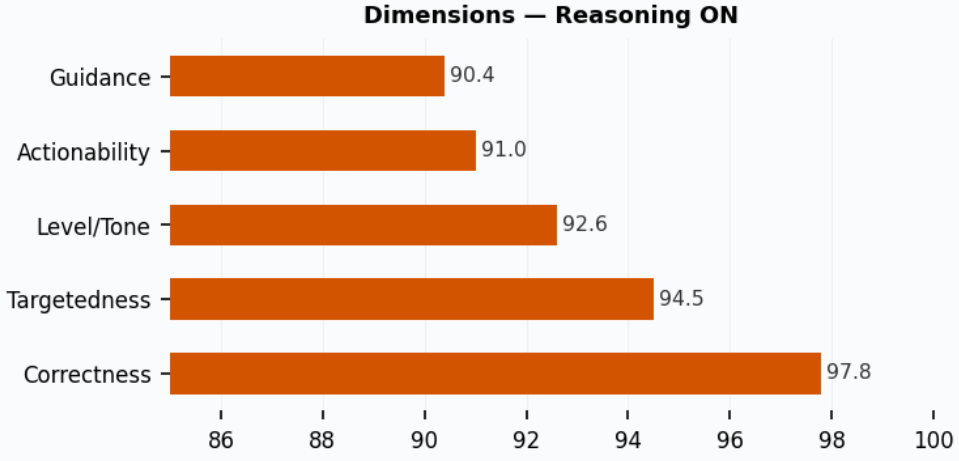
Cost OFF: $0.470 | ON: $1.370 Speed OFF: 43 tok/s | ON: 60 tok/s Flags OFF: 2 | ON: 0

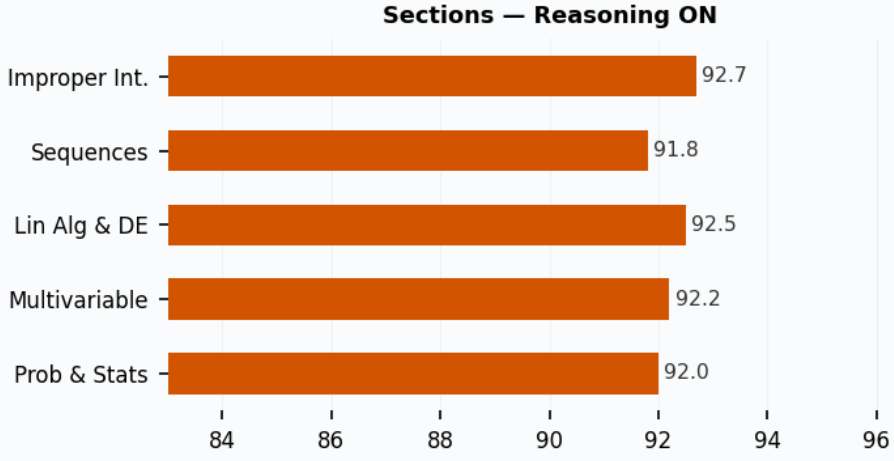
DeepSeek V4 Pro

OFF **92.9** → ON **92.2** **-0.70**





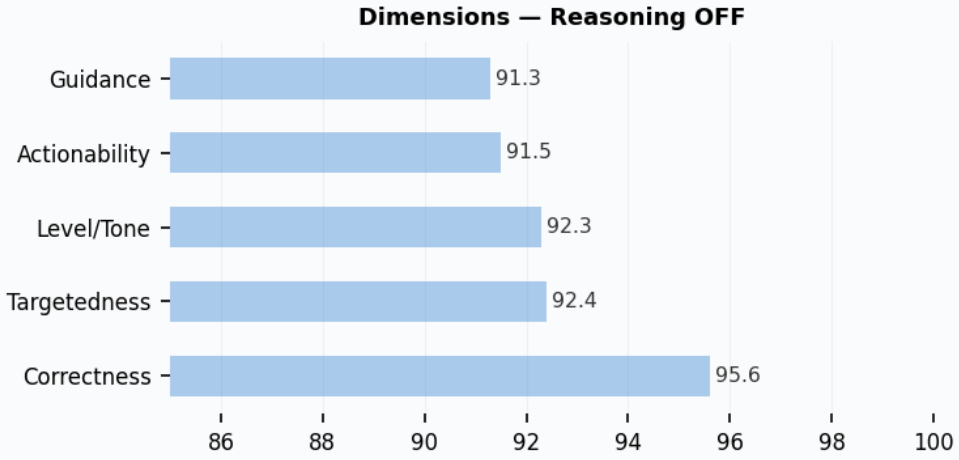


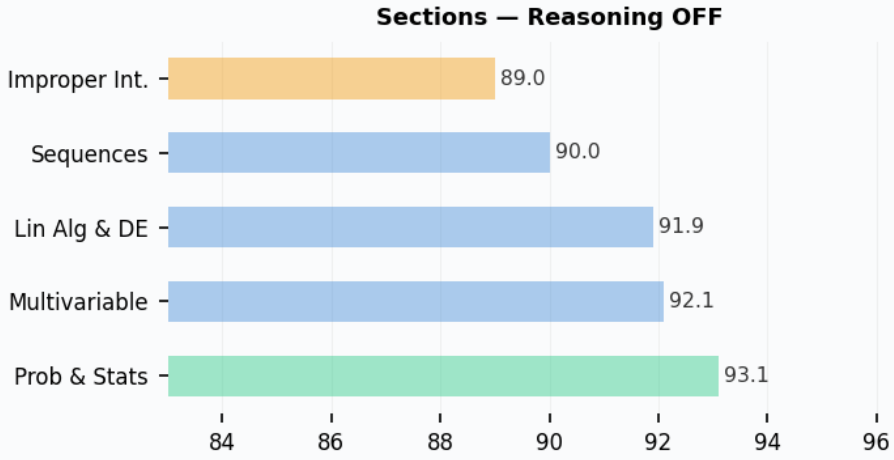


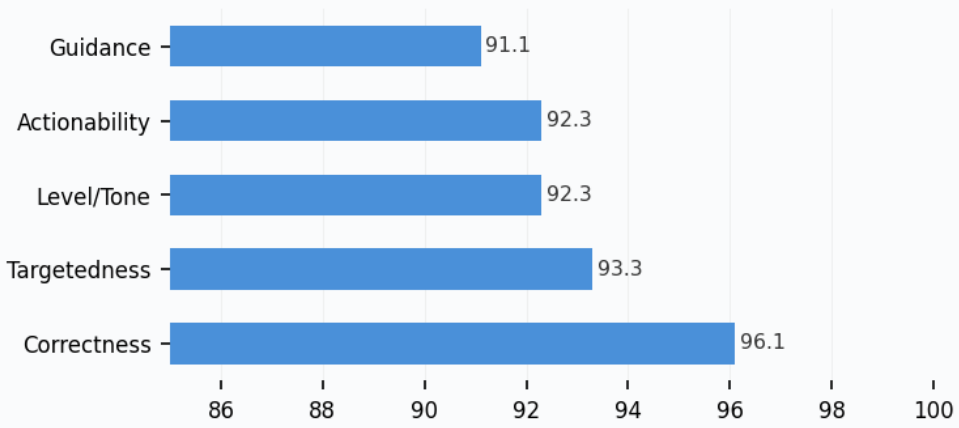
Cost OFF: $0.580 | ON: $0.260 Speed OFF: 82 tok/s | ON: 76 tok/s Flags OFF: 0 | ON: 2

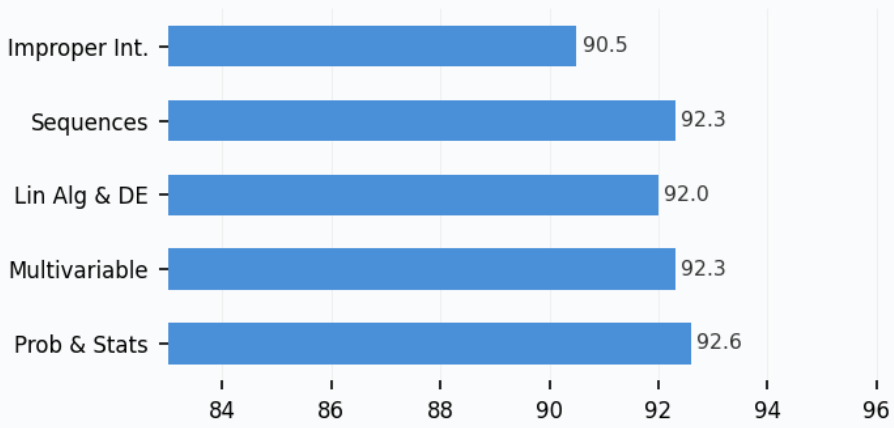
Gemini 3.5 Flash

OFF **91.2** → ON **92.0** **+0.80**





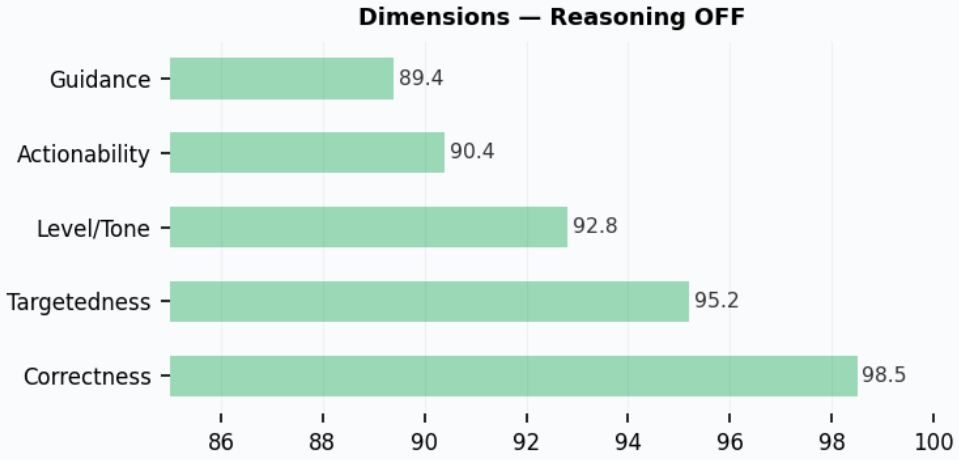


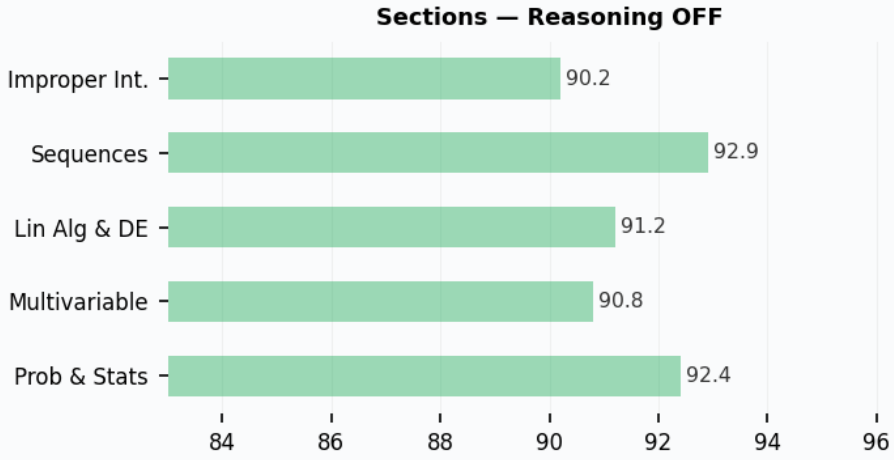


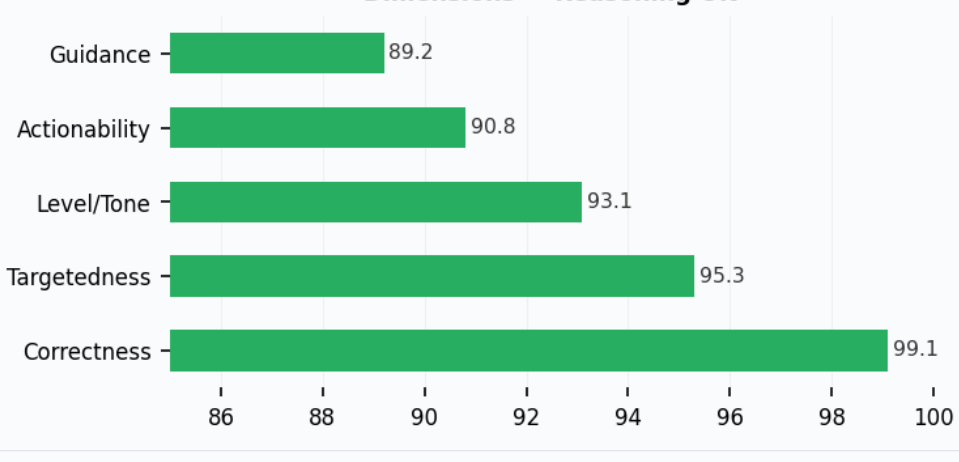
Cost OFF: $0.970 | ON: $0.680 Speed OFF: 154 tok/s | ON: 151 tok/s Flags OFF: 2 | ON: 1

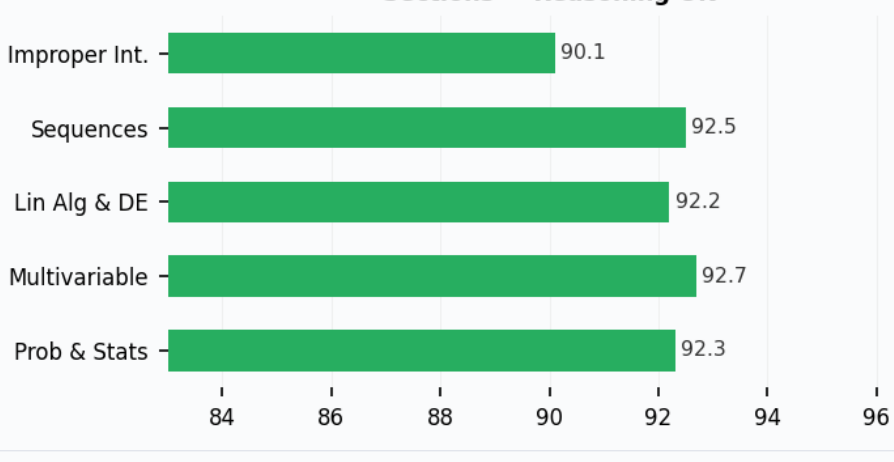
GPT-5.5

OFF **91.5** → ON **92.0** **+0.50**





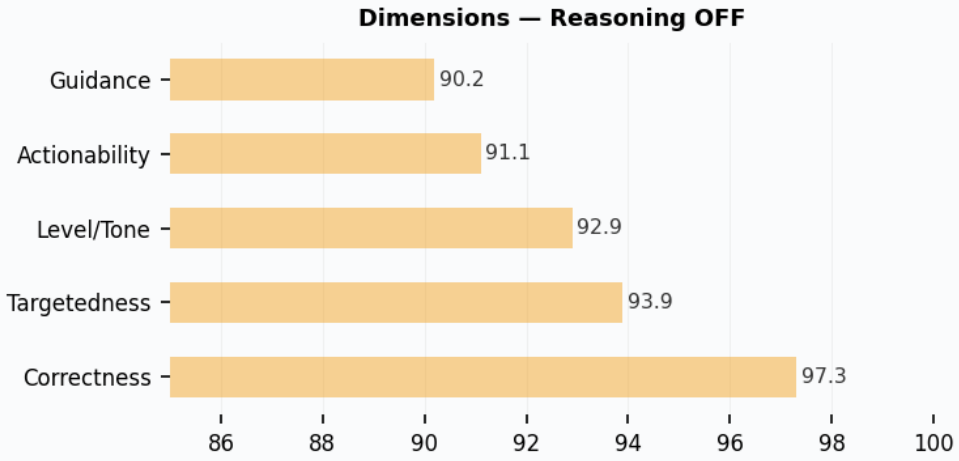


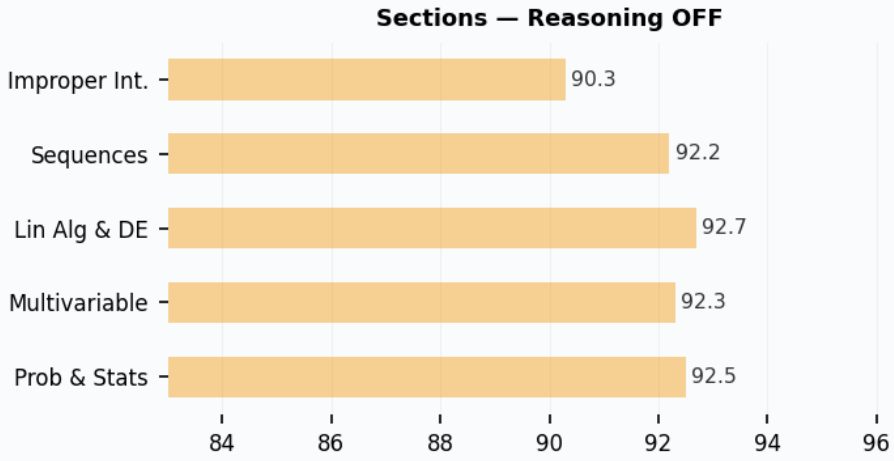


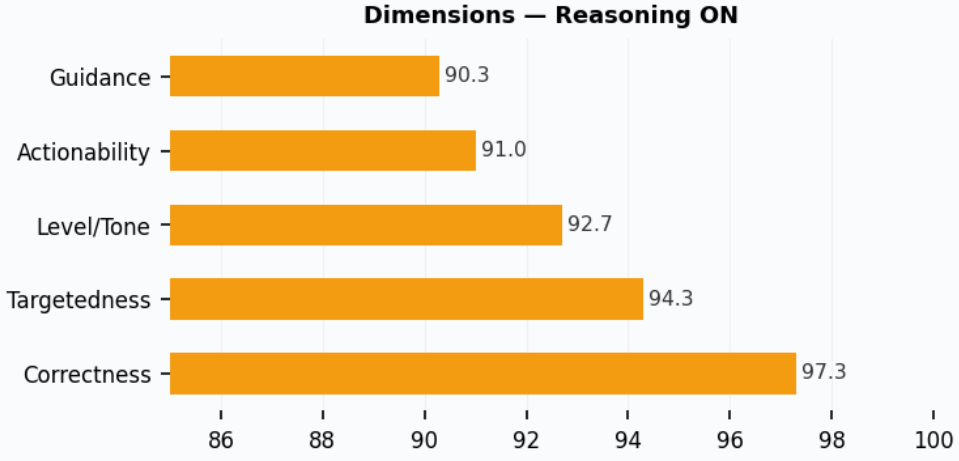
Cost OFF: $1.470 | ON: $1.200 Speed OFF: 52 tok/s | ON: 52 tok/s Flags OFF: 3 | ON: 2

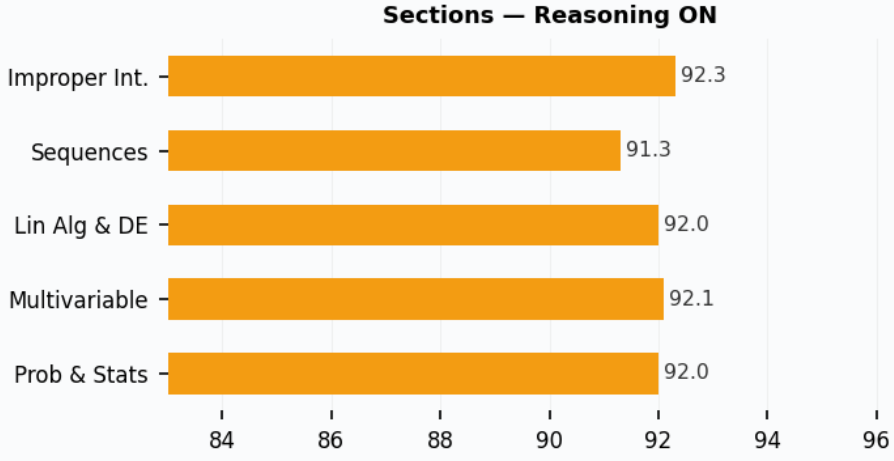
DeepSeek V4 Flash

OFF **92.0** → ON **92.0** **+0.00**





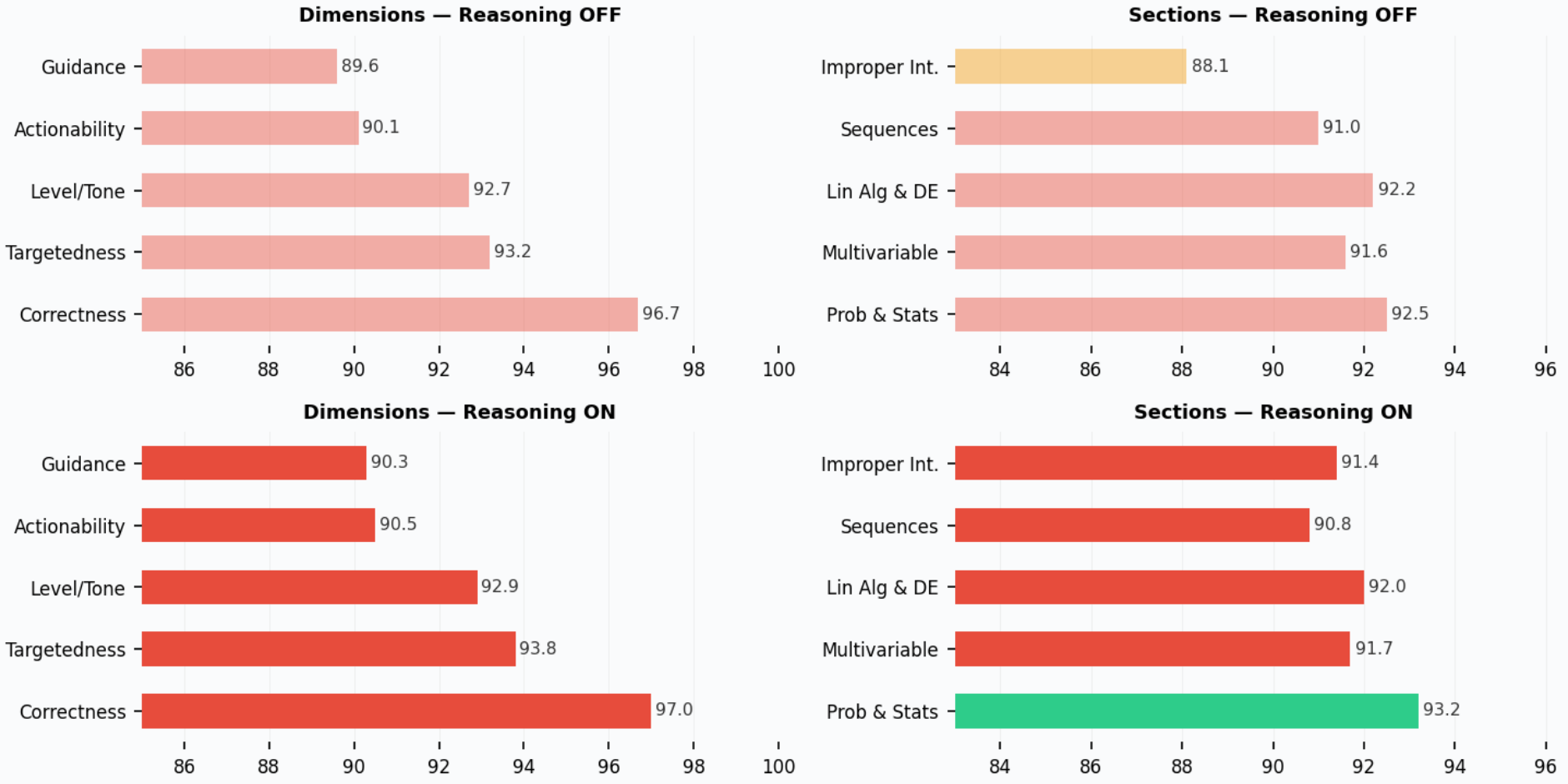




Cost OFF: $0.016 | ON: $0.016 Speed OFF: 89 tok/s | ON: 119 tok/s Flags OFF: 2 | ON: 1

Grok 4.3

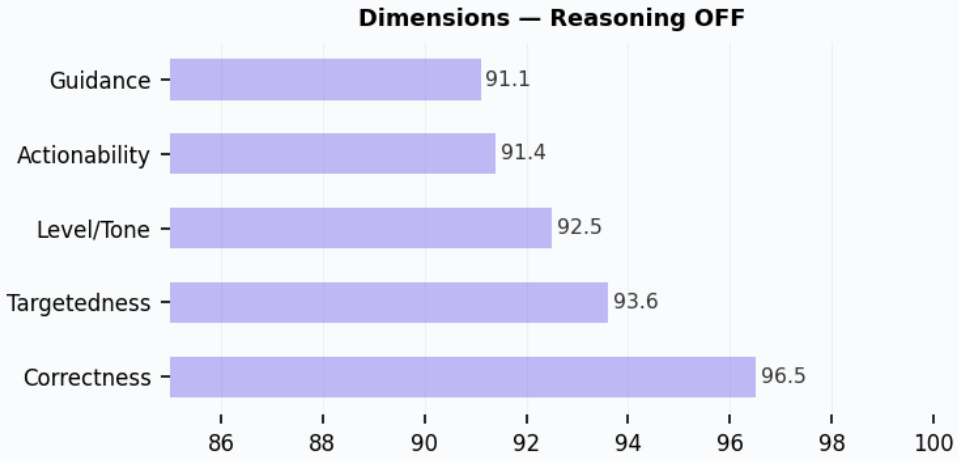
OFF **91.1** → ON **91.8** **+0.70**

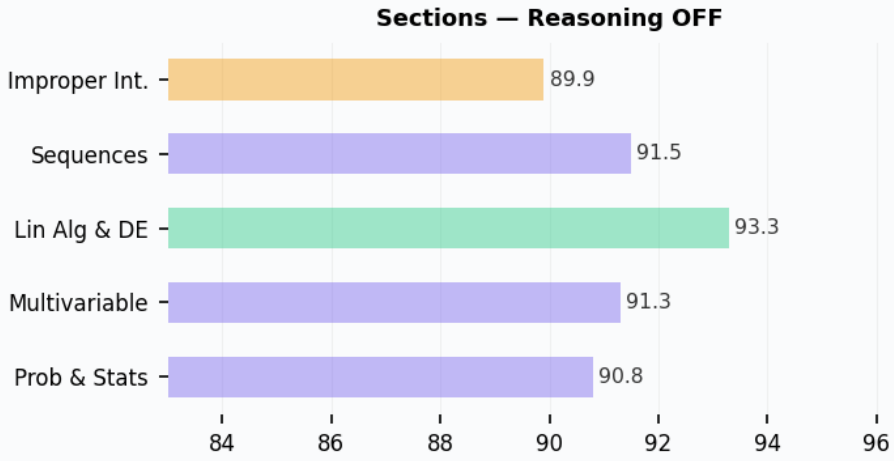


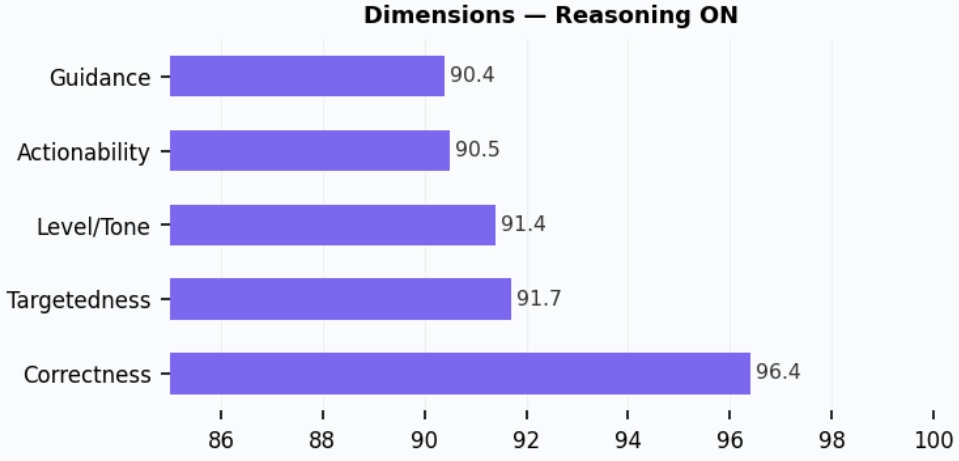
Cost OFF: $0.370 | ON: $0.120 Speed OFF: 124 tok/s | ON: 186 tok/s Flags OFF: 1 | ON: 1

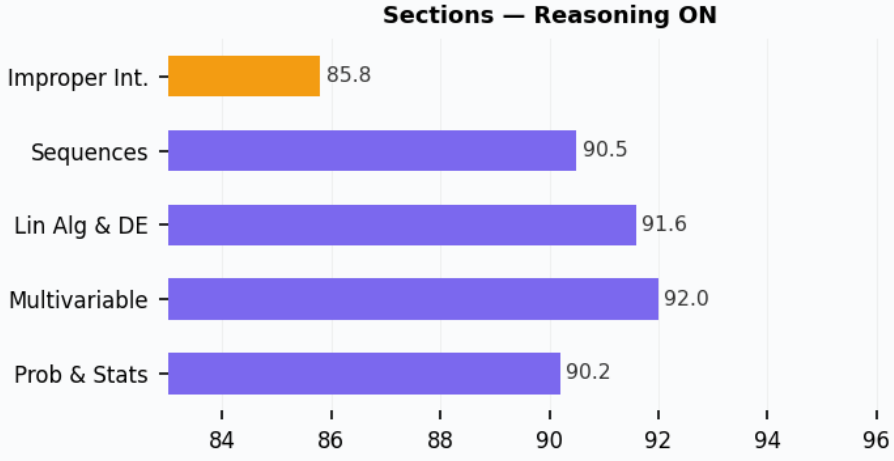
Gemini 3.1 Pro Preview

OFF **91.4** → ON **90.0** **-1.40**









Cost OFF: $1.160 | ON: $0.660 Speed OFF: 84 tok/s | ON: 80 tok/s Flags OFF: 1 | ON: 3

All scores are LLM-judged by openai/o4-mini using judge-v0.3-100pt. All models evaluated with reasoning disabled (OFF) and enabled at medium level (ON). Results should be validated against human ratings before publication. Mika cost is proprietary and not disclosed. Report generated: 2026-06-05 · RedPenBench v1.