RedPenBench Results

Physics · 6th of June 2026

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

Physics | Reasoning OFF and ON (medium)

Mika leads both conditions — 93.1 reasoning-off and 94.0 reasoning-on, the highest physics score in the batch. Mika and Claude Opus 4.8 are the only models with zero flagged items across both conditions. PHYS-1008 (Energy & Rotation) flagged in 5 of 11 models and is recommended for item review.

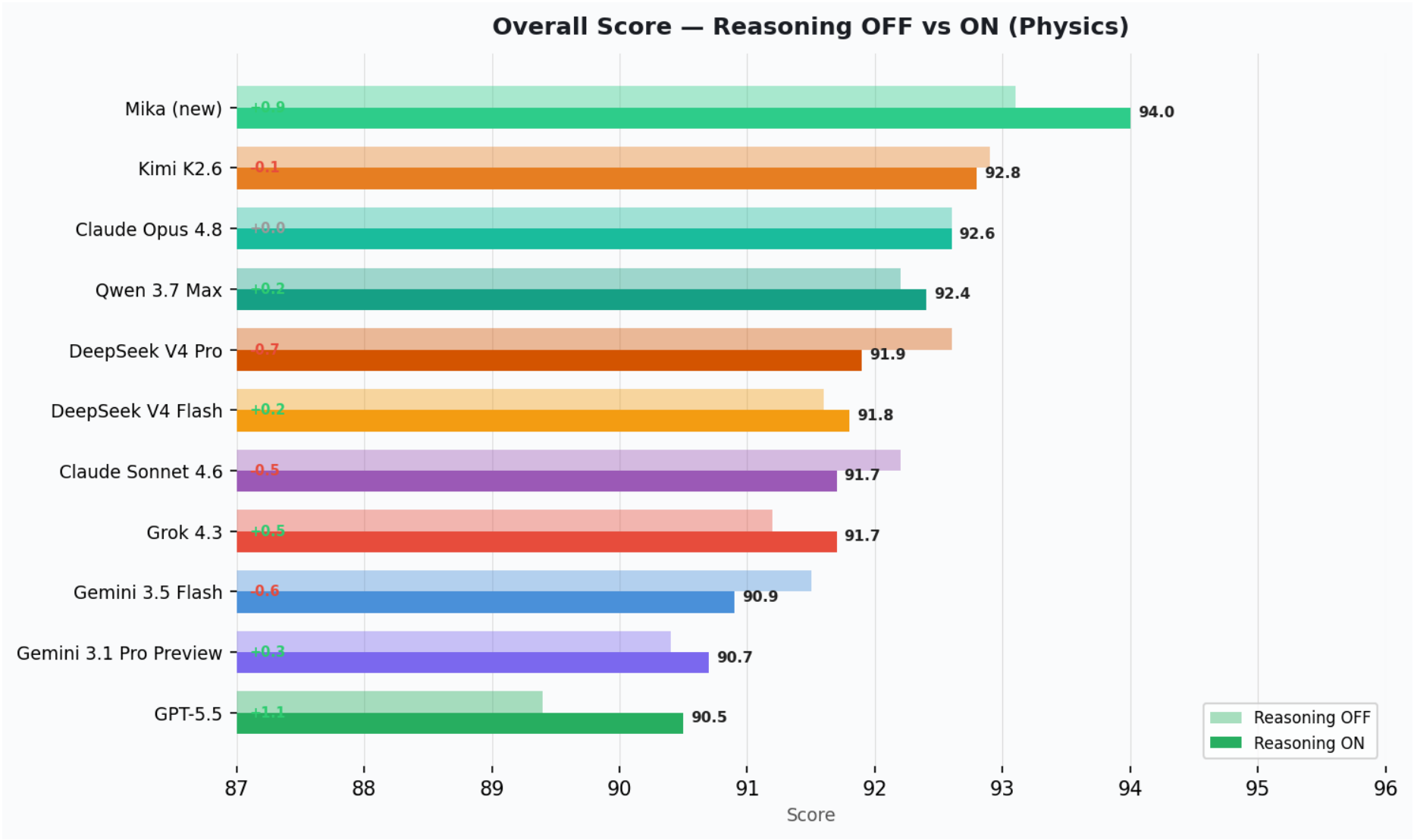
Summary

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

#	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	**94.0**	+0.90	92.8–93.4	93.7–94.3	Proprietary	Proprietary	160	130	0	0
2	Kimi K2.6	92.9	**92.8**	-0.10	92.4–93.4	91.8–93.7	$0.750	$0.810	76	227	1	2
3	Claude Opus 4.8	92.6	**92.6**	+0.00	92.2–93.1	92.1–93.0	$0.950	$1.180	61	64	0	0
4	Qwen 3.7 Max	92.2	**92.4**	+0.20	91.8–92.6	91.9–92.9	$0.470	$0.480	62	62	0	0
5	DeepSeek V4 Pro	92.6	**91.9**	-0.70	92.0–93.1	90.5–93.0	$0.240	$0.270	49	68	1	4
6	DeepSeek V4 Flash	91.6	**91.8**	+0.20	90.8–92.2	91.1–92.3	$0.015	$0.015	78	83	1	1
7	Grok 4.3	91.2	**91.7**	+0.50	90.4–91.9	91.3–92.1	$0.090	$0.120	118	176	1	0
8	Claude Sonnet 4.6	92.2	**91.7**	-0.50	91.7–92.7	90.6–92.5	$0.400	$1.130	38	50	1	1
9	Gemini 3.5 Flash	91.5	**90.9**	-0.60	90.9–92.0	90.1–91.5	$0.710	$0.660	153	144	0	1
10	Gemini 3.1 Pro Preview	90.4	**90.7**	+0.30	89.2–91.3	89.9–91.3	$0.810	$0.620	80	76	2	2
11	GPT-5.5	89.4	**90.5**	+1.10	87.8–90.9	89.3–91.5	$1.110	$1.070	42	44	5	5

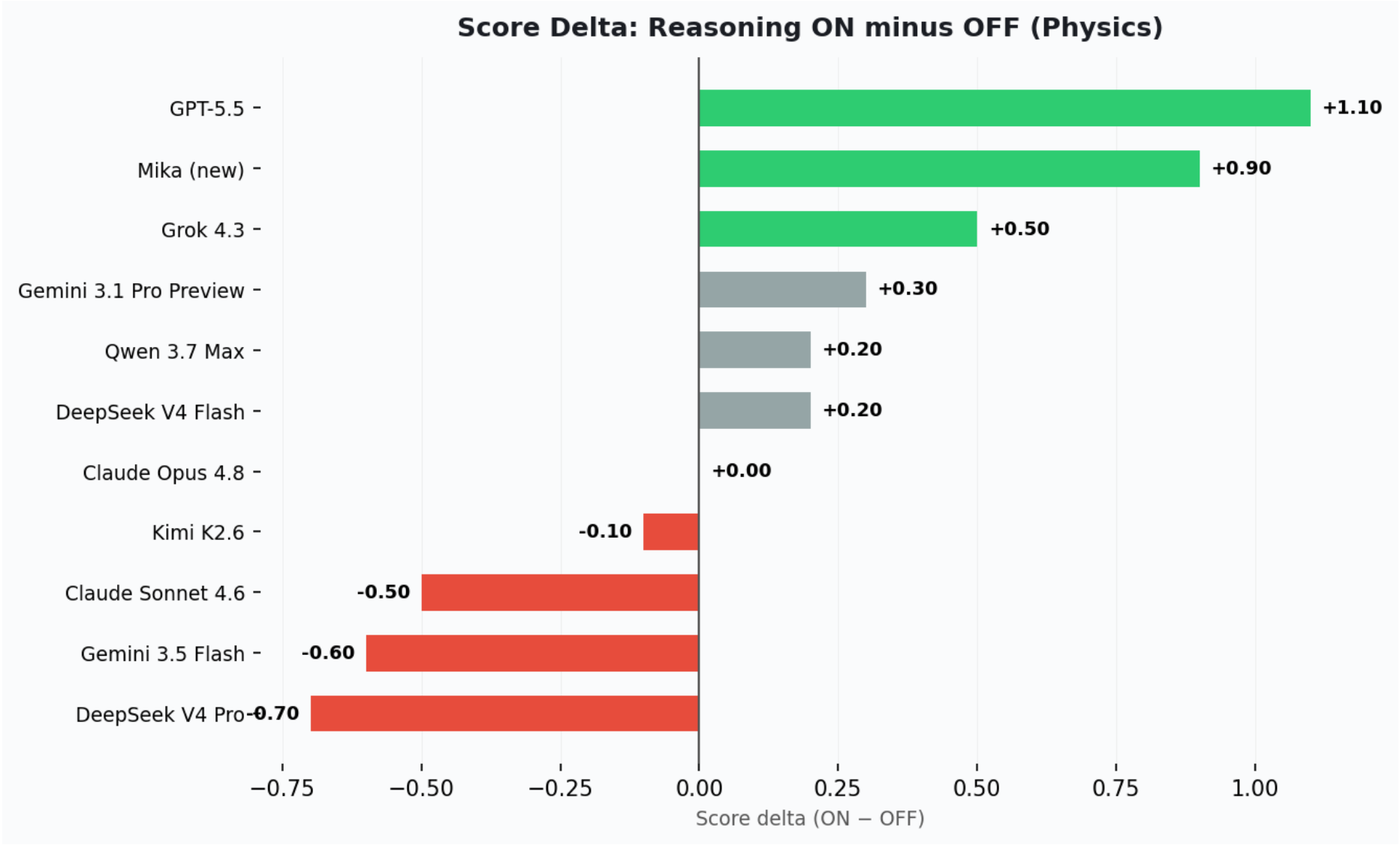
1. Overall Score — Reasoning OFF vs ON

Faded bars = reasoning off, solid = reasoning on. Delta labels on the left. Mika achieves the highest ON score (94.0) in the entire physics batch.


Overall Score — Reasoning OFF vs ON (Physics)
Mika (new)
+0.9
94.0
Kimi K2.6
-0.1
92.8
Claude Opus 4.8
+0.0
92.6
Qwen 3.7 Max
+0.2
92.4
DeepSeek V4 Pro
-0.7
91.9
DeepSeek V4 Flash
+0.2
91.8
Claude Sonnet 4.6
-0.5
91.7
Grok 4.3
+0.5
91.7
Gemini 3.5 Flash
-0.6
90.9
Gemini 3.1 Pro Preview
+0.3
90.7
GPT-5.5
+1.1
90.5
Reasoning OFF
Reasoning ON
87
88
89
90
91
92
93
94
95
96
Score

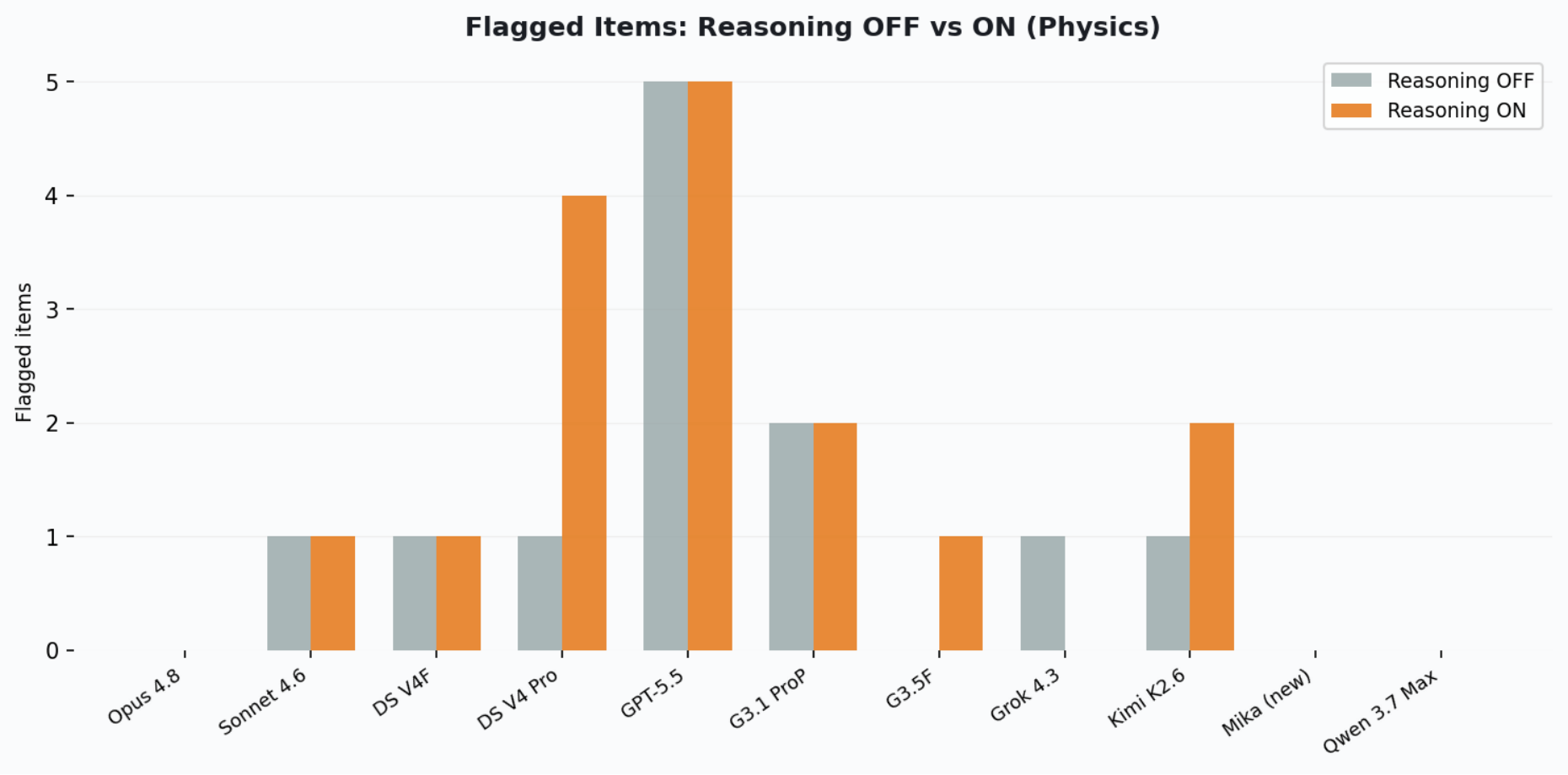
2. Impact of Reasoning — Score Delta

Mika gains the most (+0.9). GPT-5.5 also improves (+1.1) but from a much lower base. DeepSeek V4 Pro (−0.7) and Gemini 3.5 Flash (−0.6) regress — consistent with the pattern seen in maths. Kimi K2.6 and Opus are stable.


Score Delta: Reasoning ON minus OFF (Physics)
GPT-5.5
Mika (new)
Grok 4.3
Gemini 3.1 Pro Preview
Qwen 3.7 Max
DeepSeek V4 Flash
Claude Opus 4.8
Kimi K2.6
Claude Sonnet 4.6
Gemini 3.5 Flash
DeepSeek V4 Pro
+1.10
+0.90
+0.50
+0.30
+0.20
+0.20
+0.00
-0.10
-0.50
-0.60
-0.70
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Score delta (ON − OFF)

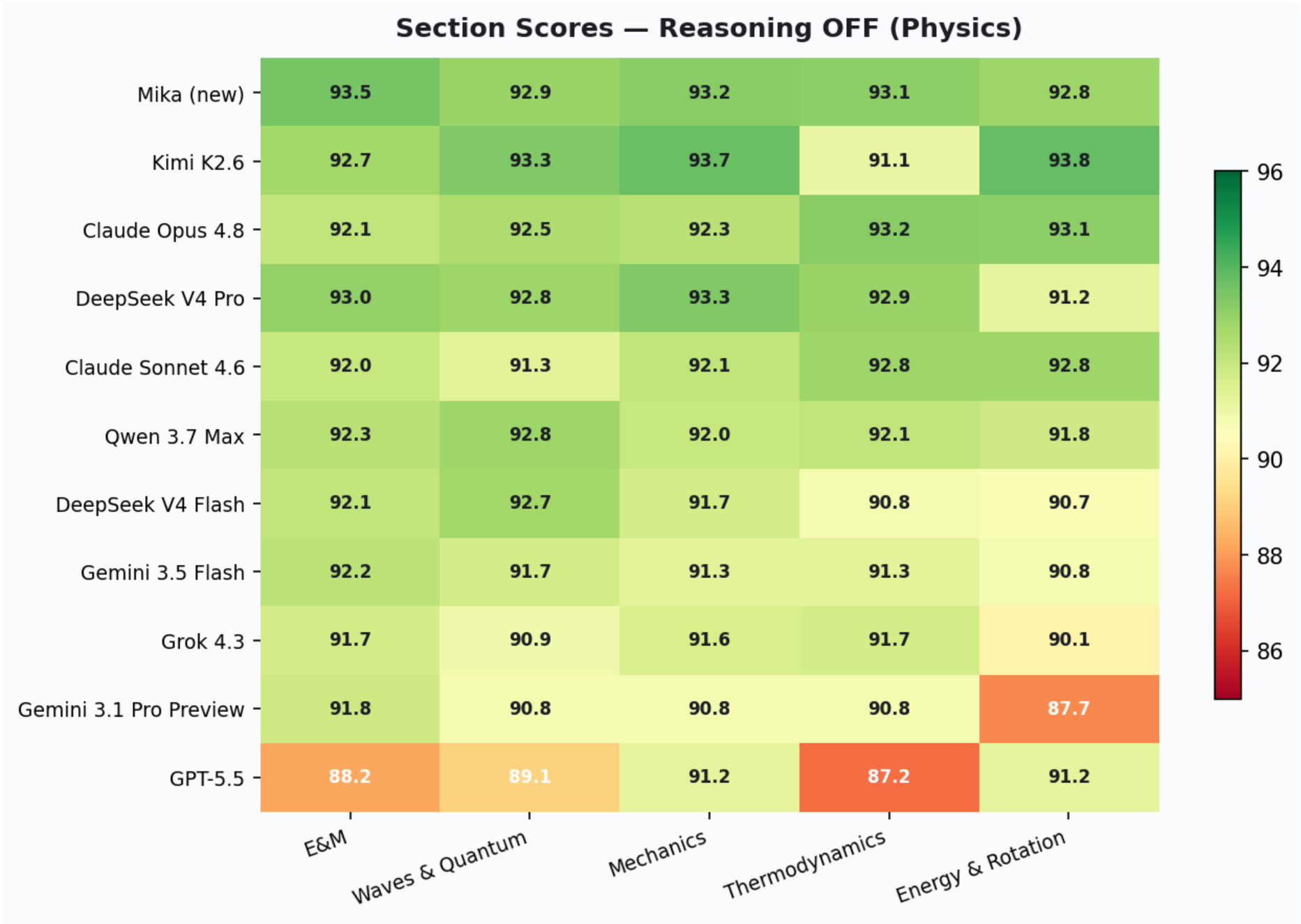
3. Flagged Items

GPT-5.5 flags 5 items in both conditions — the over-solving pattern is immune to reasoning. DeepSeek V4 Pro goes from 1 flag to 4 with reasoning on. Mika and Opus flag zero in both conditions.



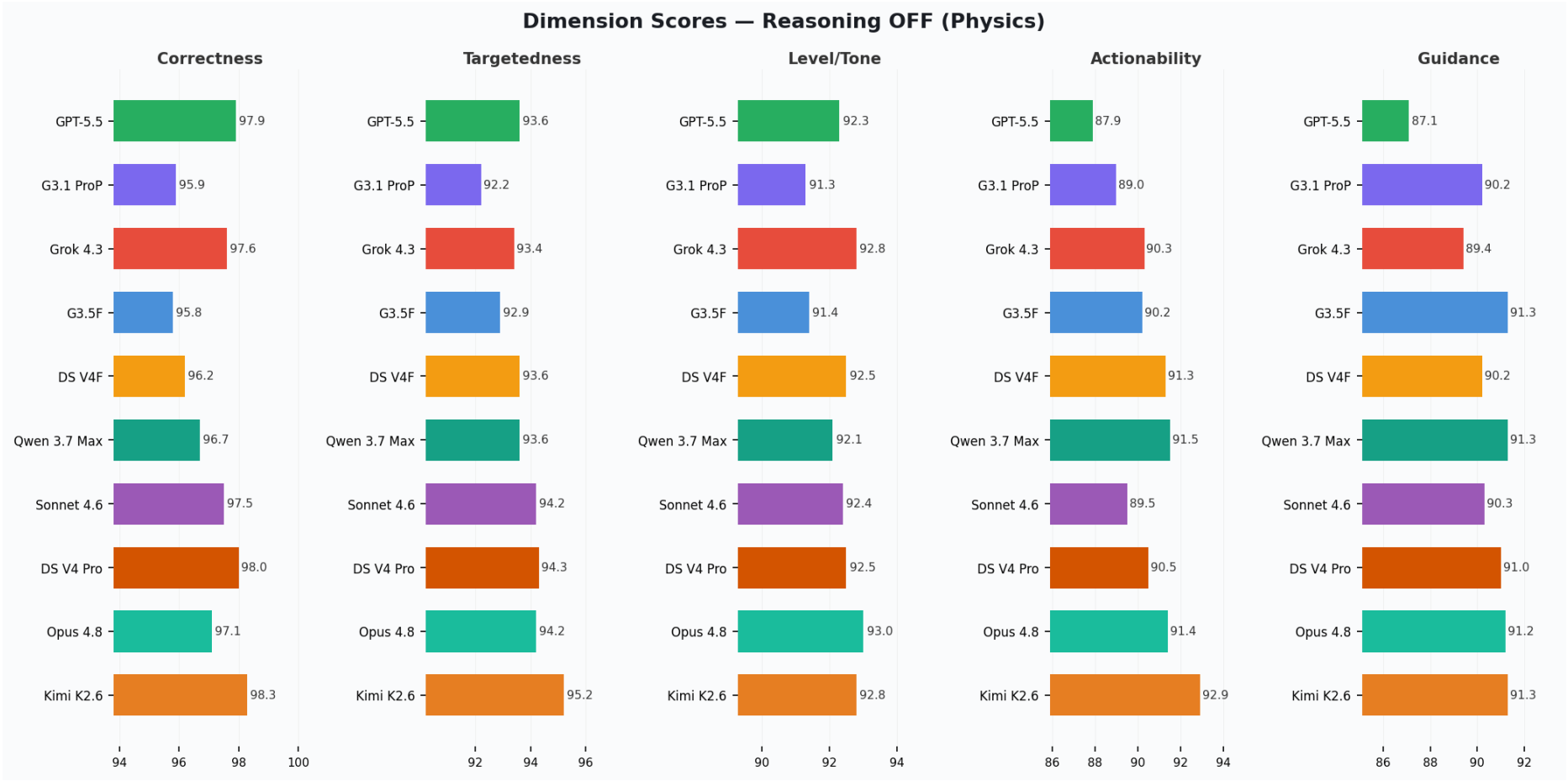
4. Section Scores

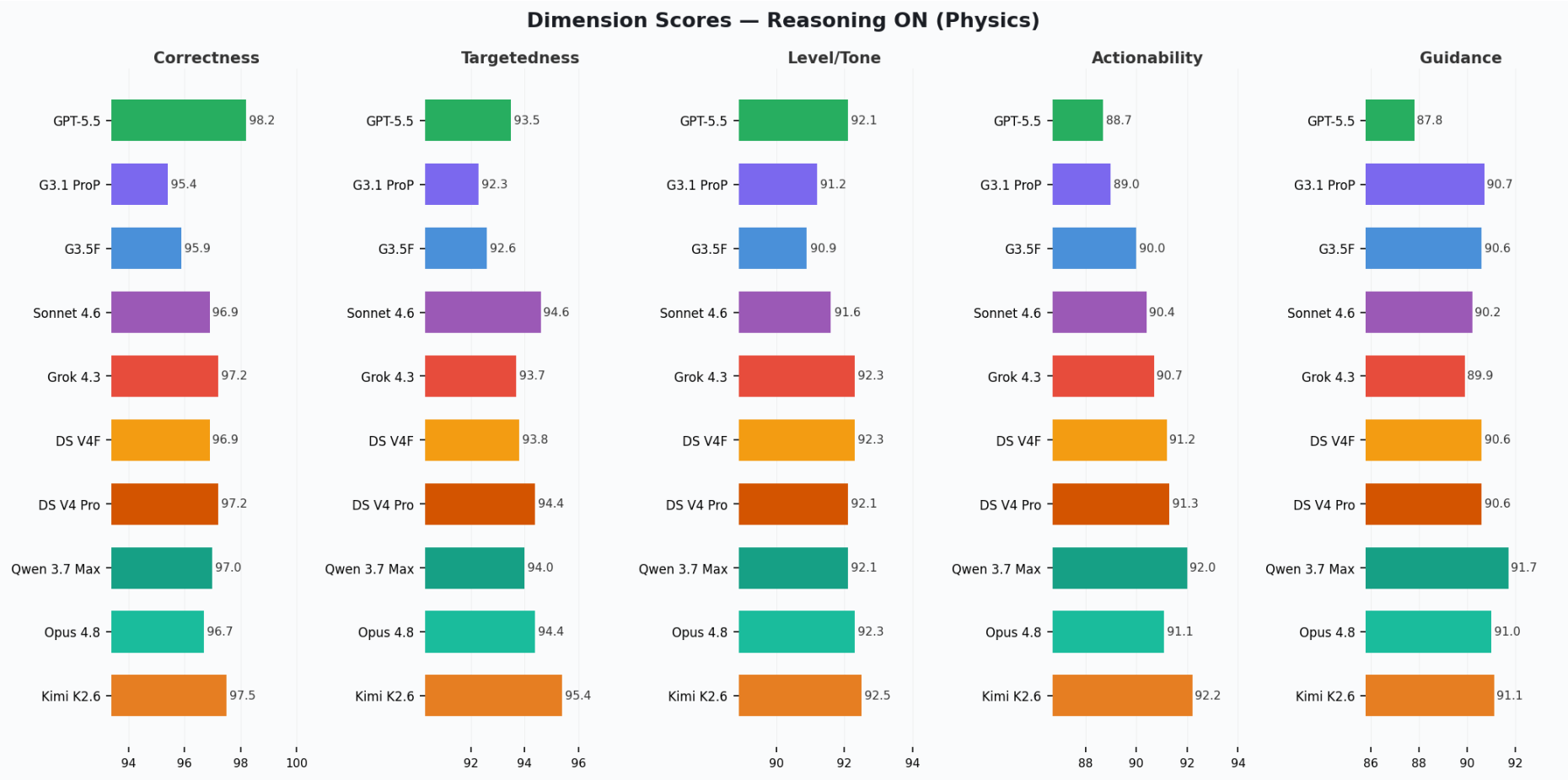
Energy & Rotation is the physics equivalent of Improper Integrals in maths — the section with the most flagged items (PHYS-1008 alone flagged in 5 models). Thermodynamics also proves difficult for Gemini Pro and DeepSeek Pro with reasoning on.


Section Scores — Reasoning OFF (Physics)
Mika (new)
93.5
92.9
93.2
93.1
92.8
Kimi K2.6
92.7
93.3
93.7
91.1
93.8
Claude Opus 4.8
92.1
92.5
92.3
93.2
93.1
DeepSeek V4 Pro
93.0
92.8
93.3
92.9
91.2
Claude Sonnet 4.6
92.0
91.3
92.1
92.8
92.8
Qwen 3.7 Max
92.3
92.8
92.0
92.1
91.8
DeepSeek V4 Flash
92.1
92.7
91.7
90.8
90.7
Gemini 3.5 Flash
92.2
91.7
91.3
91.3
90.8
Grok 4.3
91.7
90.9
91.6
91.7
90.1
Gemini 3.1 Pro Preview
91.8
90.8
90.8
90.8
87.7
GPT-5.5
88.2
89.1
91.2
87.2
91.2
E&M
Waves & Quantum
Mechanics
Thermodynamics
Energy & Rotation
96
94
92
90
88
86

5. Dimension Scores

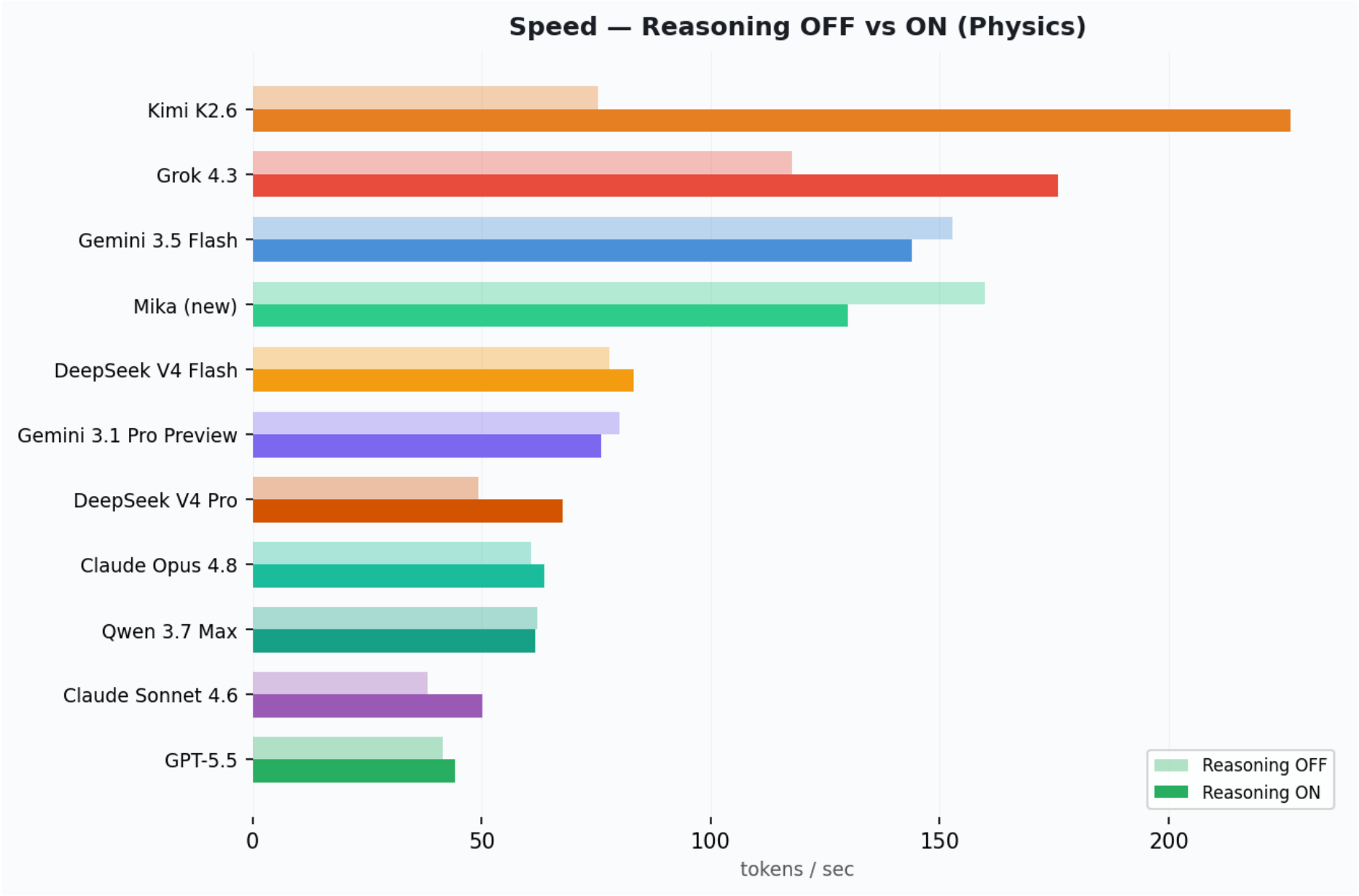
Guidance and Actionability remain the universally weakest dimensions — GPT-5.5's Guidance score (87.1 OFF, 87.8 ON) is the lowest in the physics batch. Correctness is strong across all models in both conditions.





6. Speed

Kimi K2.6 remains the fastest at 226.6 tok/s reasoning-on. Grok 4.3 jumps to 175.9 tok/s. Mika at 130 tok/s reasoning-on is the fastest top-tier scorer. GPT-5.5 and Sonnet 4.6 are the slowest.

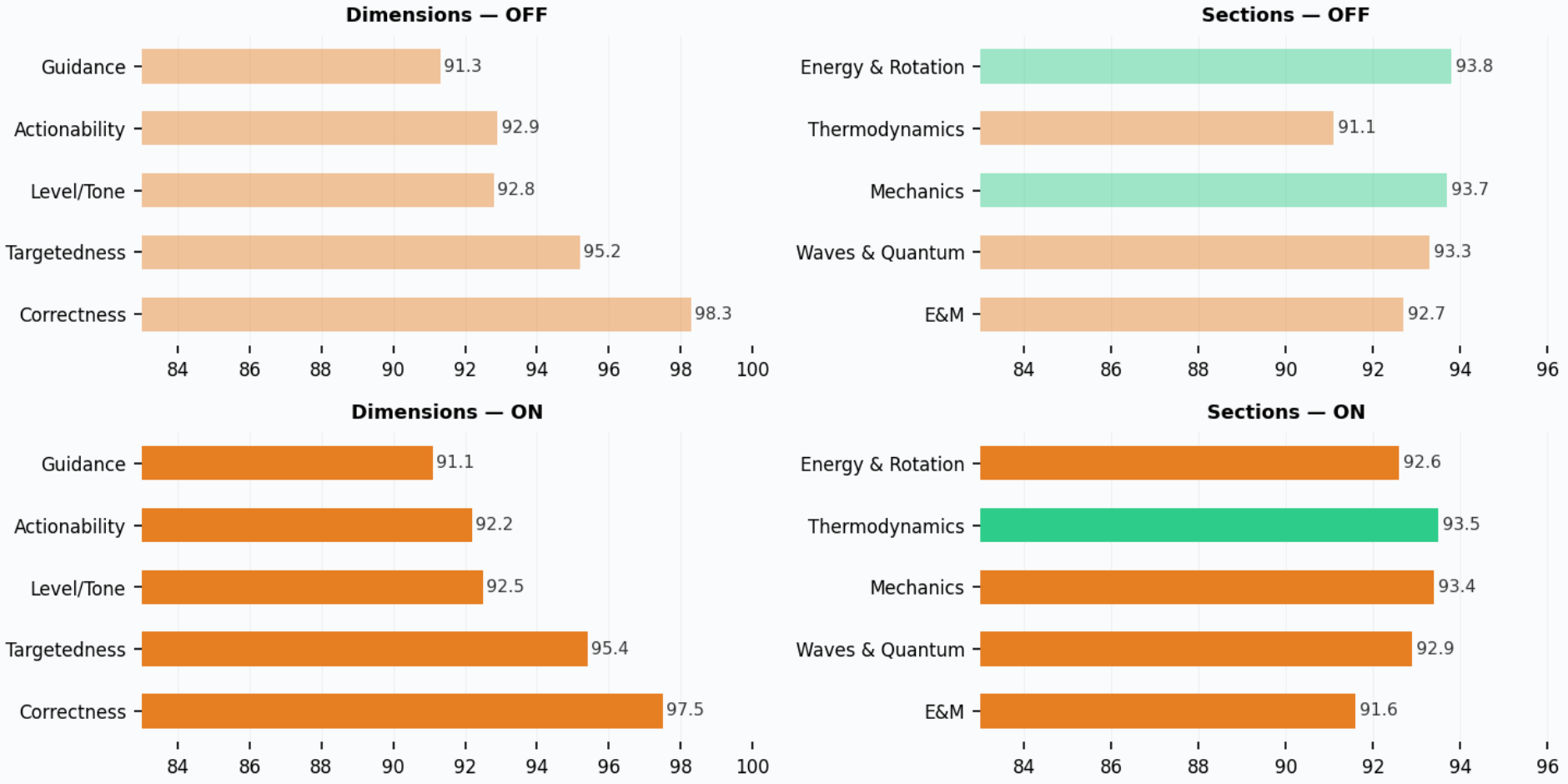

Speed — Reasoning OFF vs ON (Physics)
Kimi K2.6
Grok 4.3
Gemini 3.5 Flash
Mika (new)
DeepSeek V4 Flash
Gemini 3.1 Pro Preview
DeepSeek V4 Pro
Claude Opus 4.8
Qwen 3.7 Max
Claude Sonnet 4.6
GPT-5.5
Reasoning OFF
Reasoning ON
0
50
100
150
200
tokens / sec

7. Individual Model Cards

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

Kimi K2.6

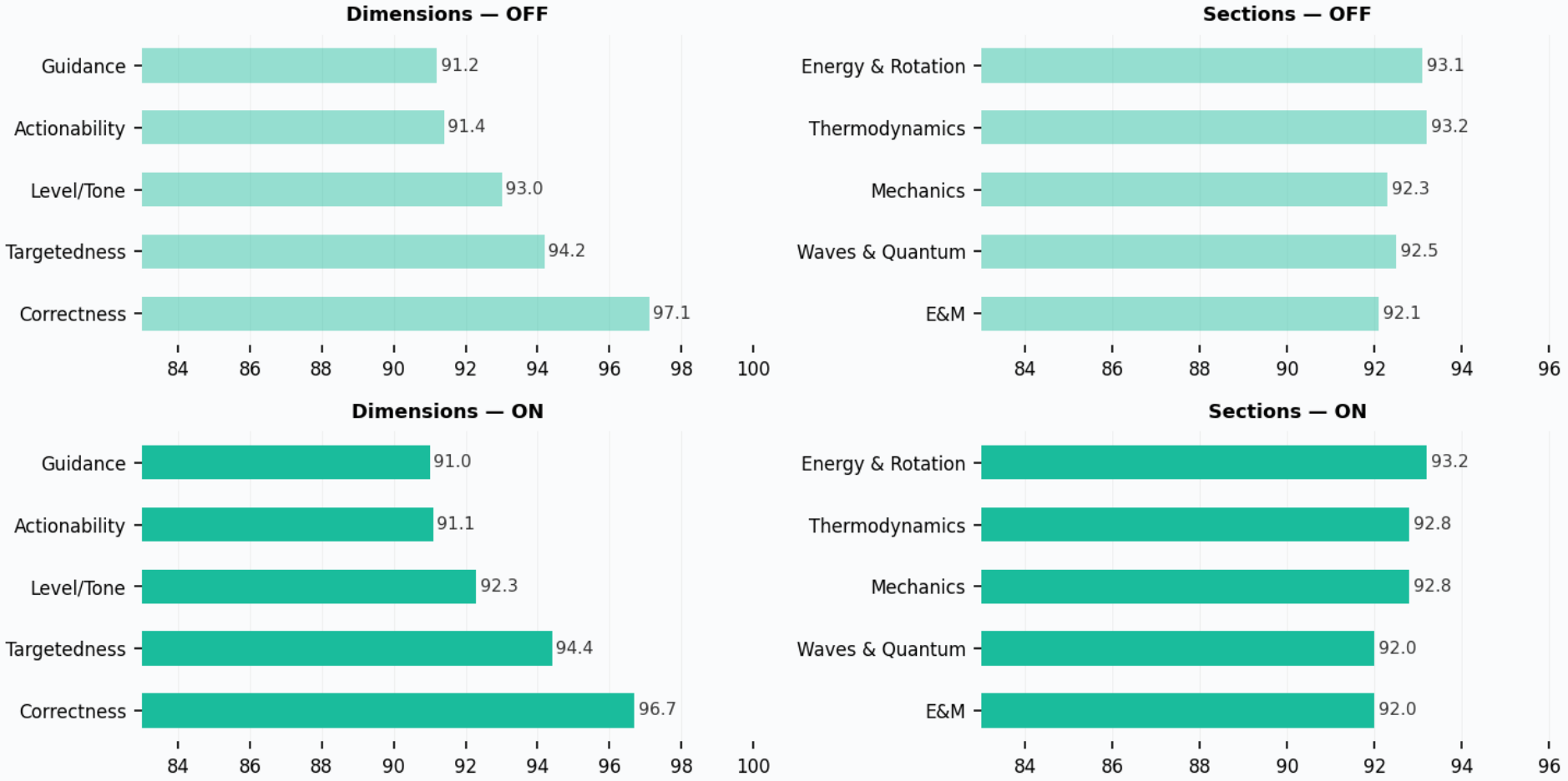
OFF **92.9** → ON **92.8** **-0.10**



Cost OFF: $0.750 | ON: $0.810 Speed OFF: 76 tok/s | ON: 227 tok/s Flags OFF: 1 | ON: 2

Claude Opus 4.8

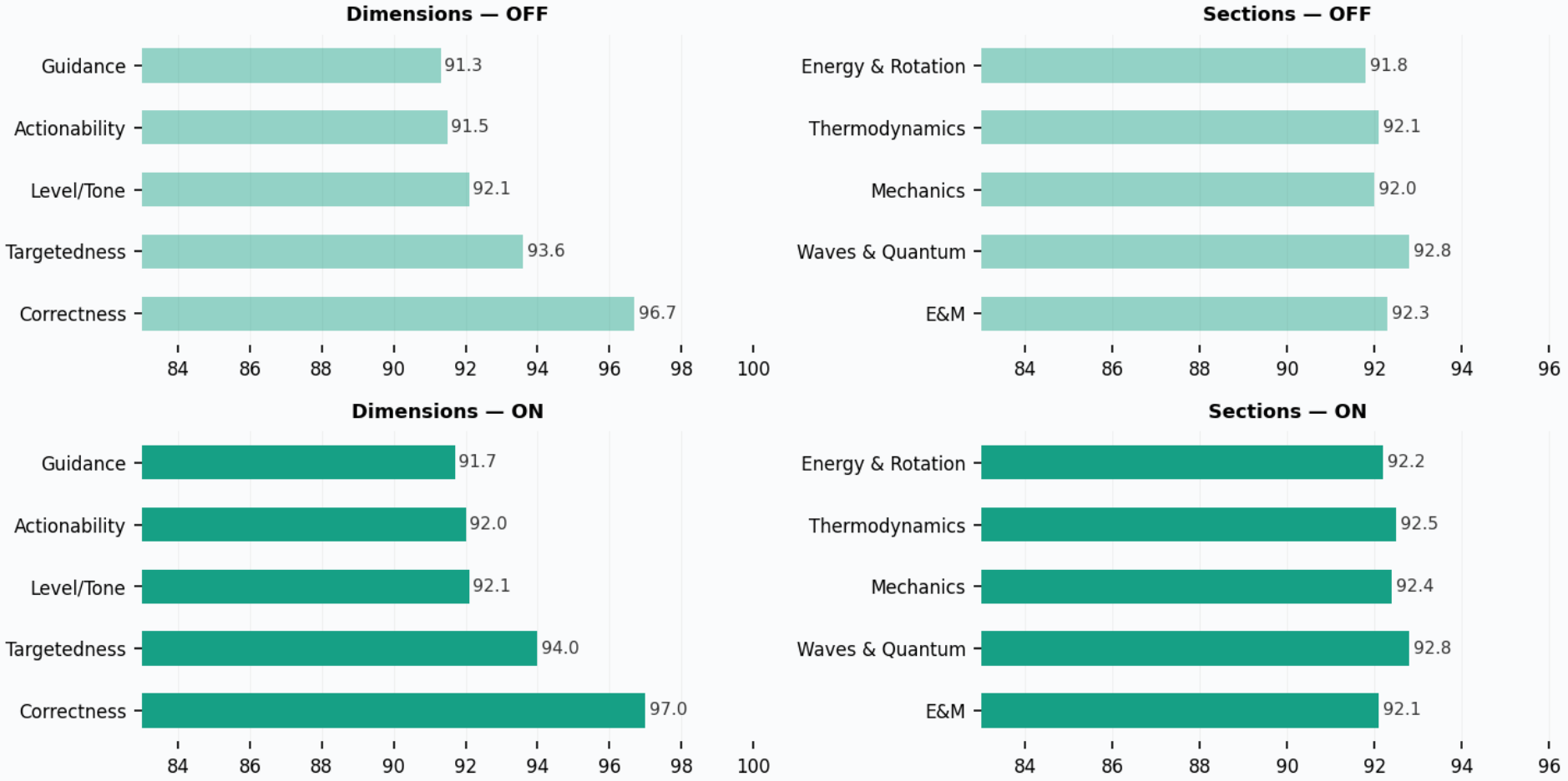
OFF **92.6** → ON **92.6** **+0.00**



Cost OFF: $0.950 | ON: $1.180 Speed OFF: 61 tok/s | ON: 64 tok/s Flags OFF: 0 | ON: 0

Qwen 3.7 Max

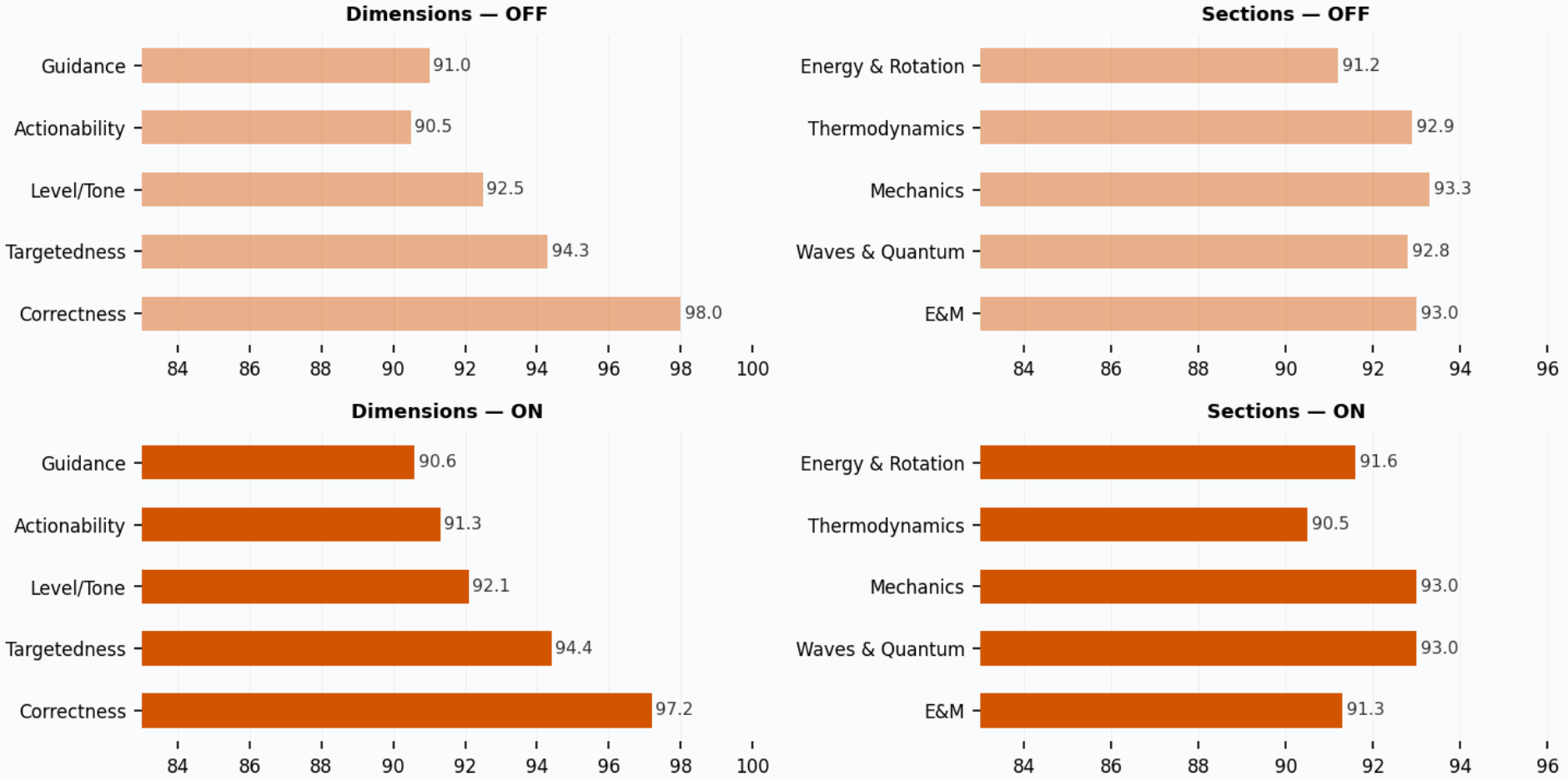
OFF **92.2** → ON **92.4** **+0.20**



Cost OFF: $0.470 | ON: $0.480 Speed OFF: 62 tok/s | ON: 62 tok/s Flags OFF: 0 | ON: 0

DeepSeek V4 Pro

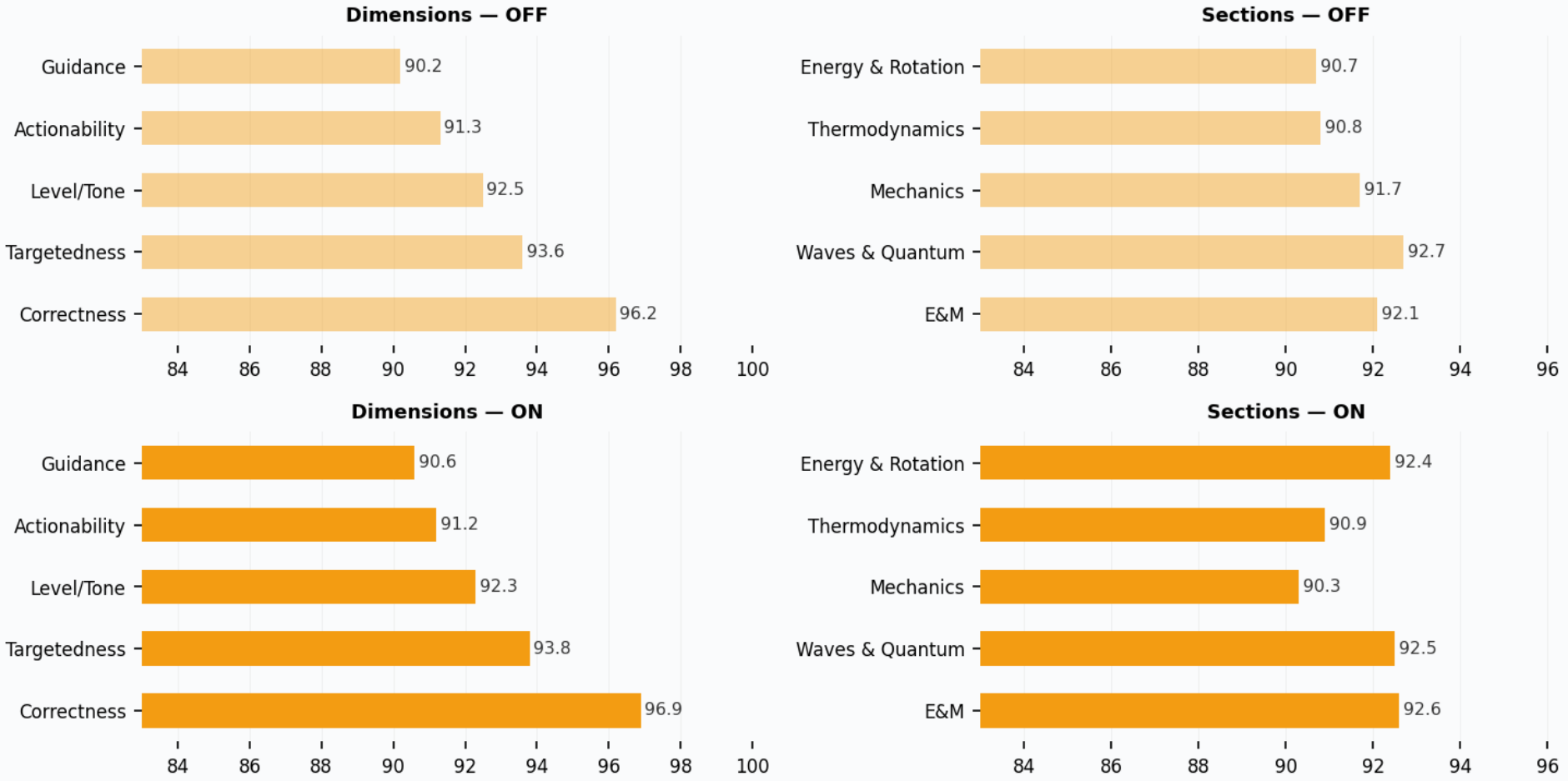
OFF **92.6** → ON **91.9** **-0.70**



Cost OFF: $0.240 | ON: $0.270 Speed OFF: 49 tok/s | ON: 68 tok/s Flags OFF: 1 | ON: 4

DeepSeek V4 Flash

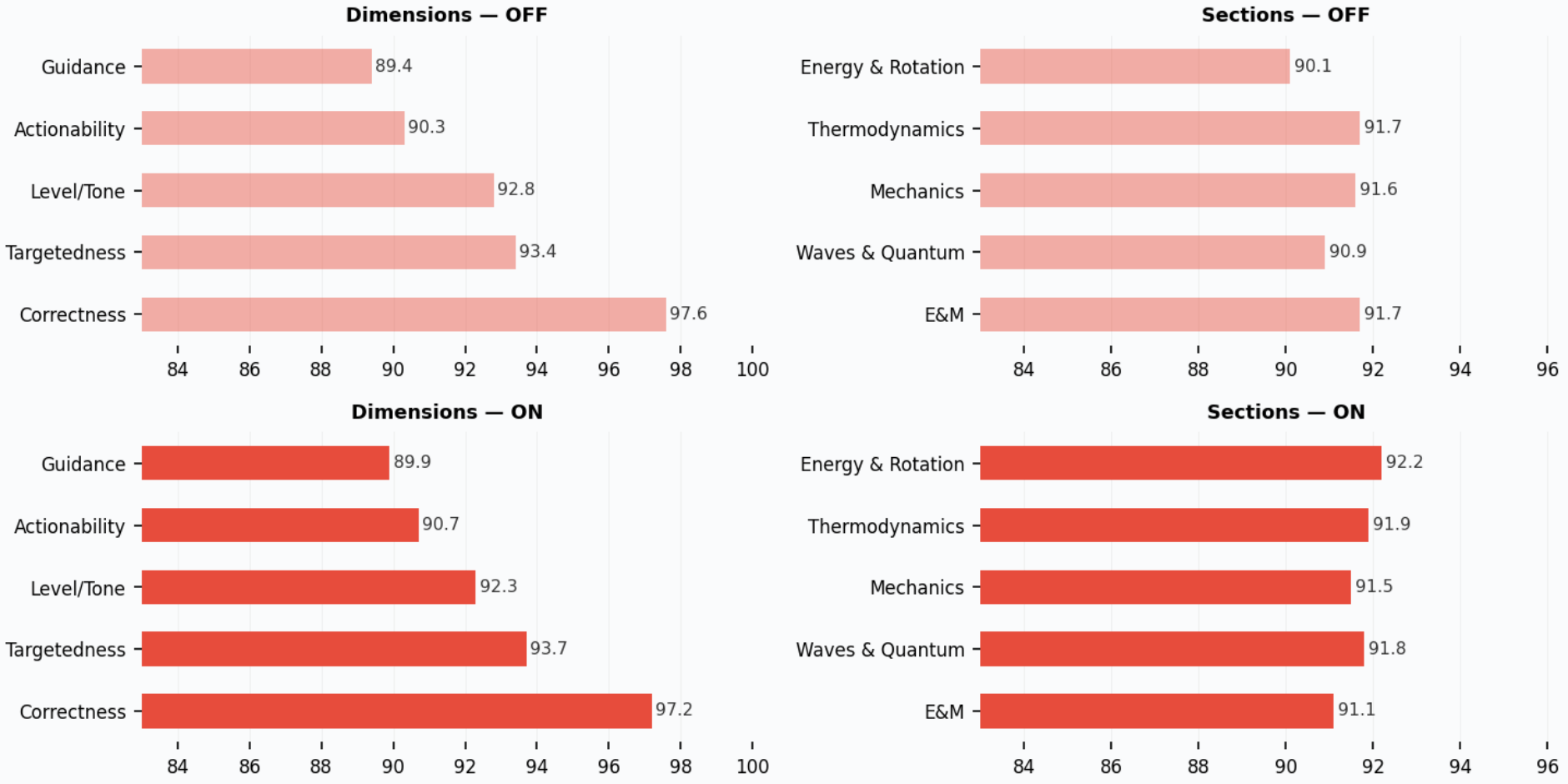
OFF **91.6** → ON **91.8** **+0.20**



Cost OFF: $0.015 | ON: $0.015 Speed OFF: 78 tok/s | ON: 83 tok/s Flags OFF: 1 | ON: 1

Grok 4.3

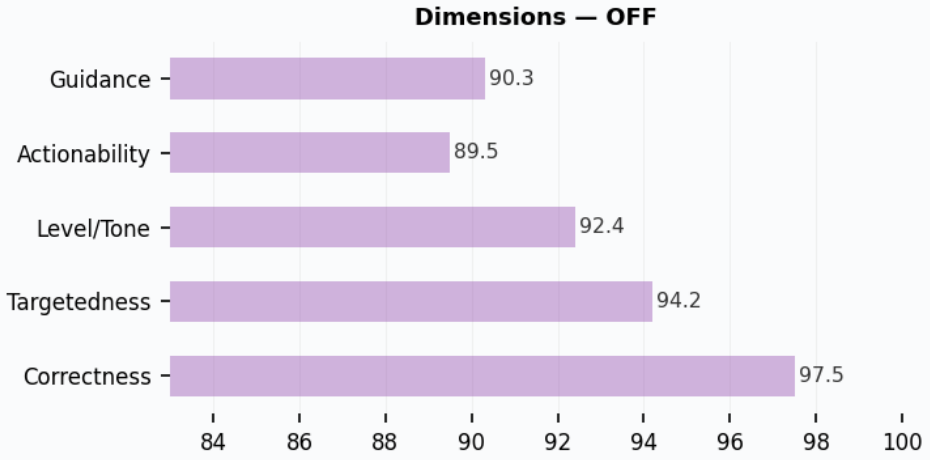
OFF **91.2** → ON **91.7** **+0.50**

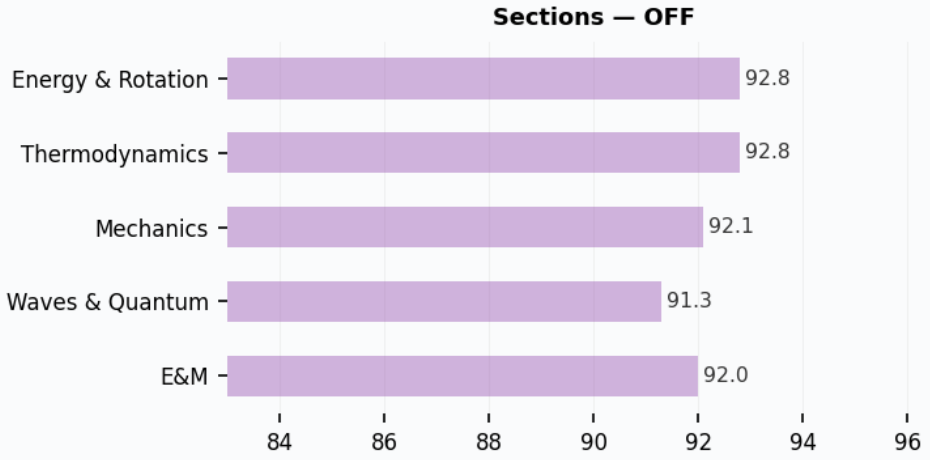


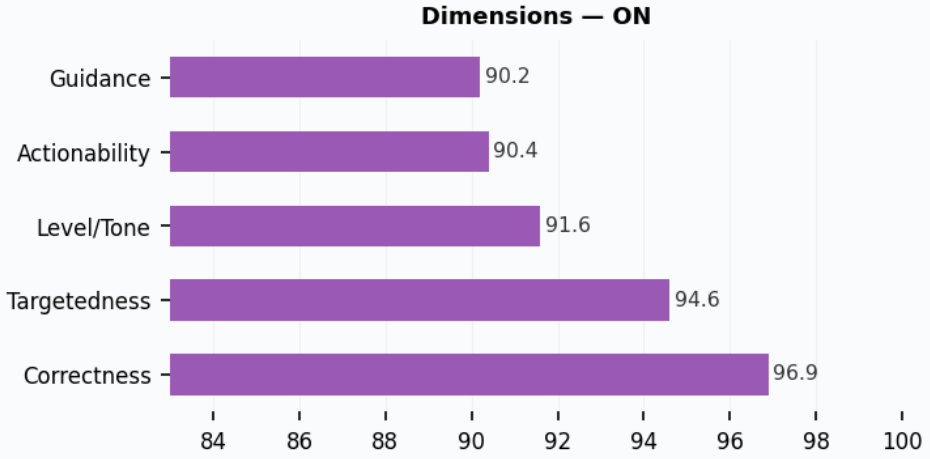
Cost OFF: $0.090 | ON: $0.120 Speed OFF: 118 tok/s | ON: 176 tok/s Flags OFF: 1 | ON: 0

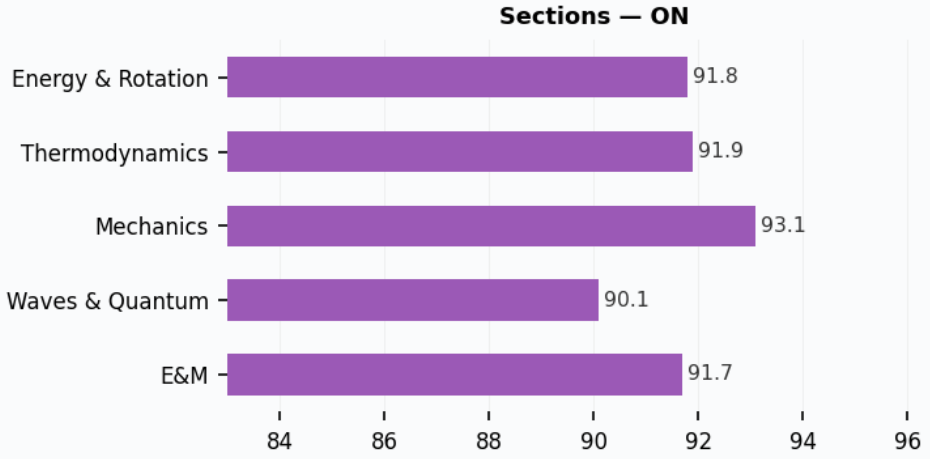
Claude Sonnet 4.6

OFF **92.2** → ON **91.7** **-0.50**





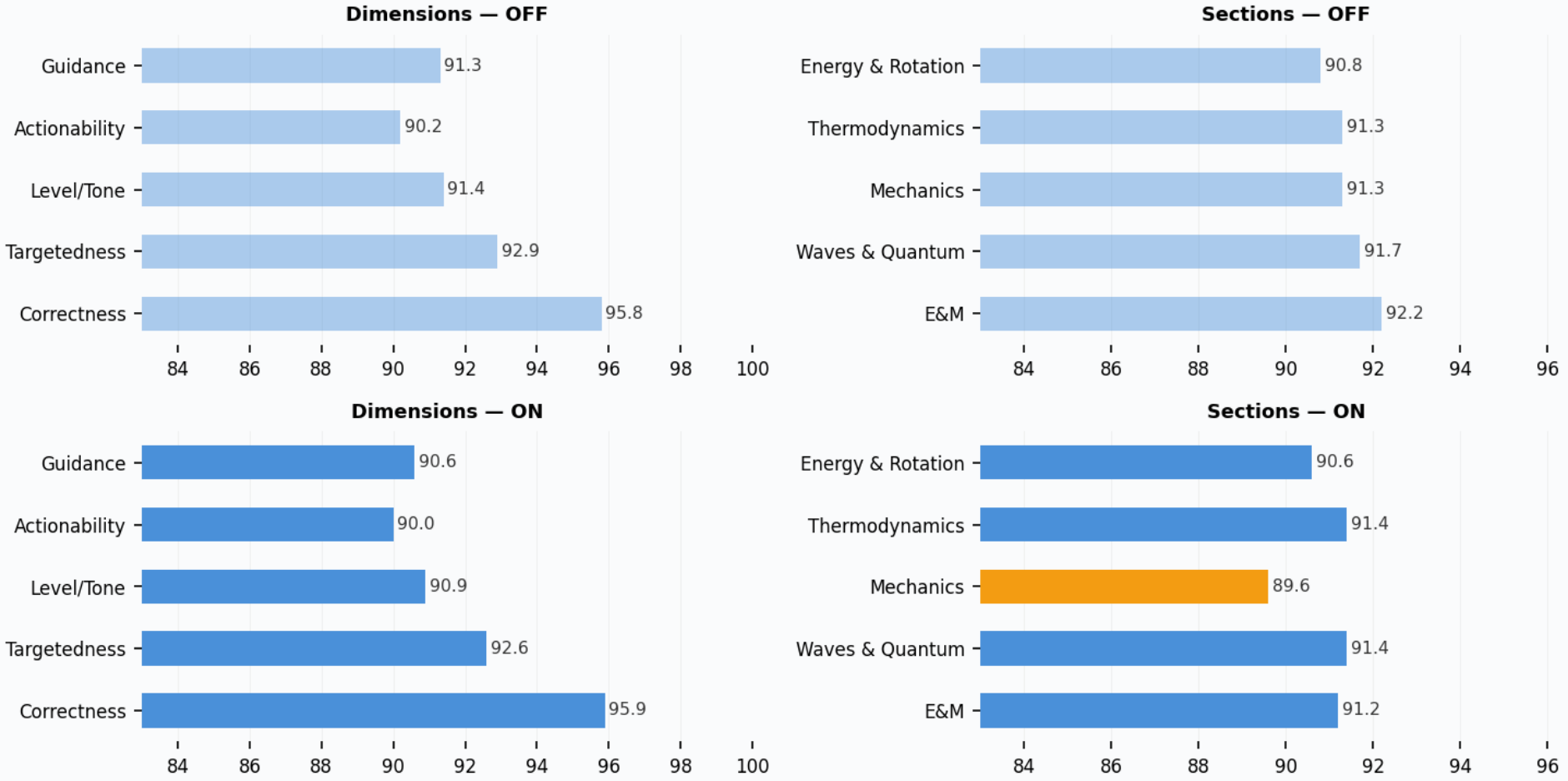




Cost OFF: $0.400 | ON: $1.130 Speed OFF: 38 tok/s | ON: 50 tok/s Flags OFF: 1 | ON: 1

Gemini 3.5 Flash

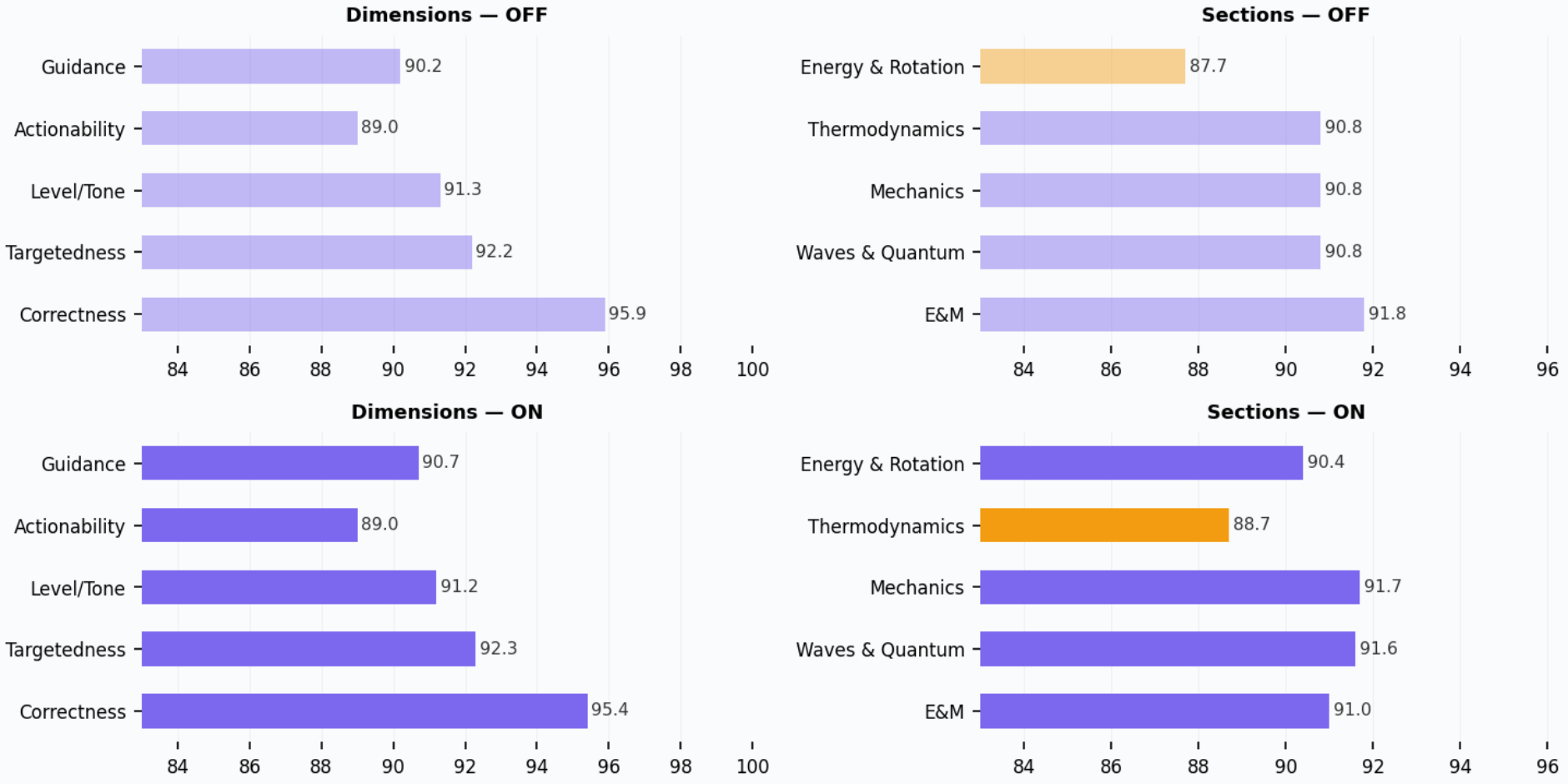
OFF **91.5** → ON **90.9** **-0.60**



Cost OFF: $0.710 | ON: $0.660 Speed OFF: 153 tok/s | ON: 144 tok/s Flags OFF: 0 | ON: 1

Gemini 3.1 Pro Preview

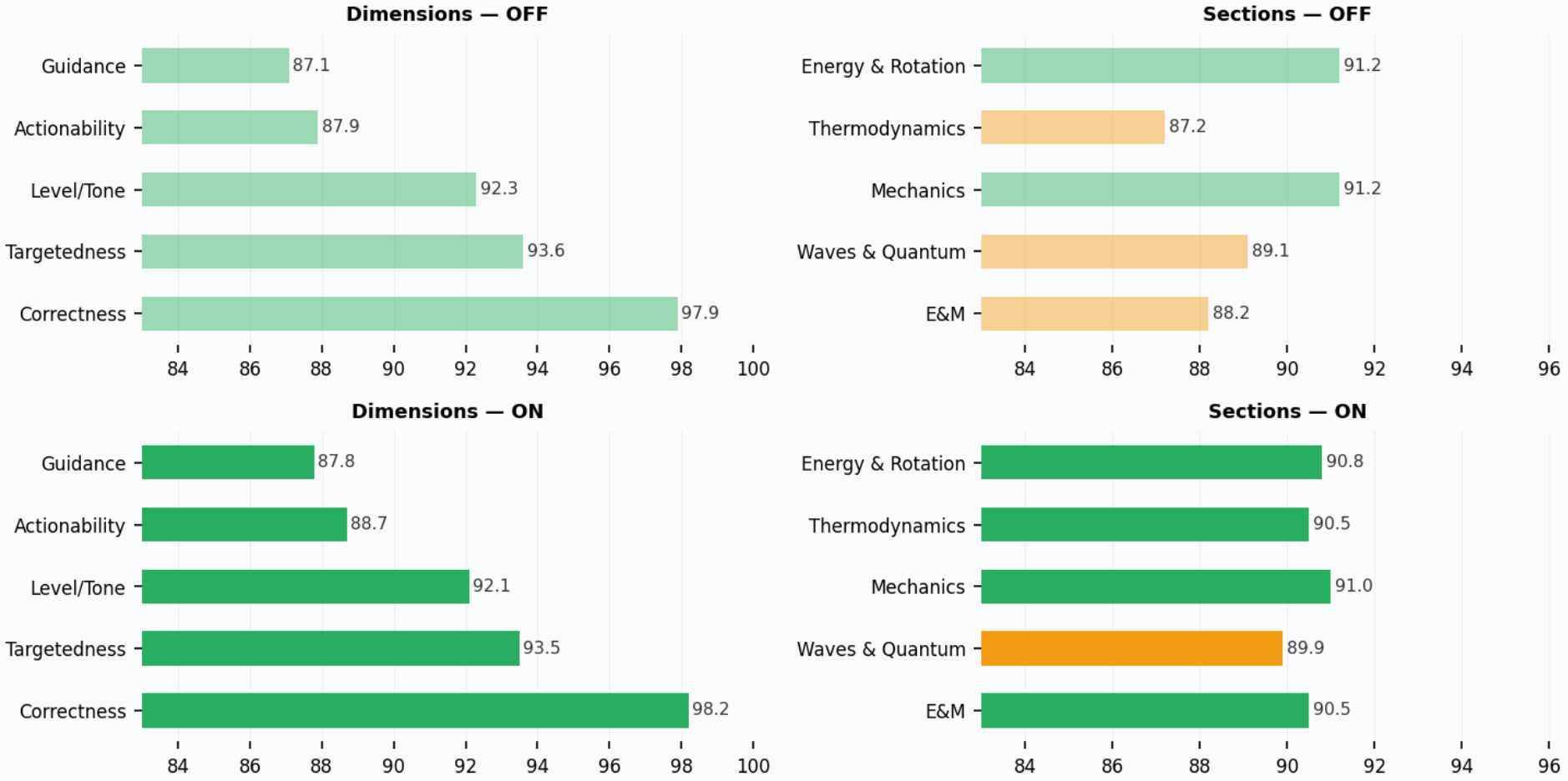
OFF **90.4** → ON **90.7** **+0.30**



Cost OFF: $0.810 | ON: $0.620 Speed OFF: 80 tok/s | ON: 76 tok/s Flags OFF: 2 | ON: 2

GPT-5.5

OFF **89.4** → ON **90.5** **+1.10**



Cost OFF: $1.110 | ON: $1.070 Speed OFF: 42 tok/s | ON: 44 tok/s Flags OFF: 5 | ON: 5

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: 2026-06-06 · RedPenBench v1 · Physics.