RedPenBench Results

Biology · 6th of June 2026

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

Biology | Reasoning OFF

Mika leads with 94.1 — the highest biology score in the batch, zero flagged items, and the only model to score above 93 in any section across any subject. Kimi K2.6 is the only other model with zero flags (92.9). Biology is the hardest subject in this benchmark — no reasoning-off model besides Mika and Kimi scores above 92.

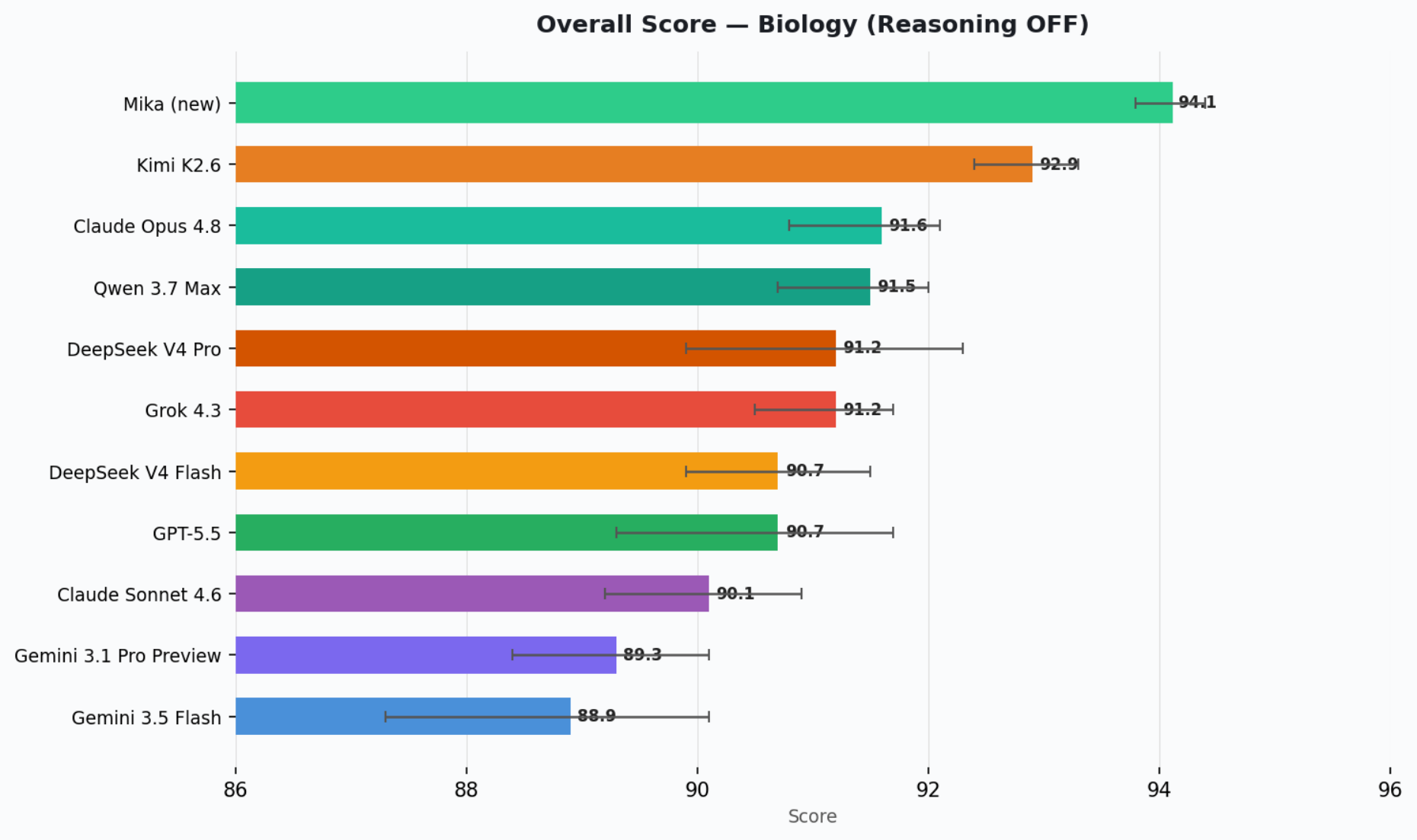
Summary

All 11 models ranked by score. Mika row highlighted. Note: Biology is reasoning-off only — no reasoning-on condition was run for this subject.

#	Model	Score	CI	Cost	Speed (tok/s)	Flagged
1	Mika (new)	**94.1**	93.8–94.4	Proprietary	170	0
2	Kimi K2.6	**92.9**	92.4–93.3	$0.540	210	0
3	Claude Opus 4.8	**91.6**	90.8–92.1	$0.960	56	1
4	Qwen 3.7 Max	**91.5**	90.7–92.0	$0.470	52	2
5	Grok 4.3	**91.2**	90.5–91.7	$0.080	146	2
6	DeepSeek V4 Pro	**91.2**	89.9–92.3	$0.180	48	2
7	GPT-5.5	**90.7**	89.3–91.7	$0.830	41	**3**
8	DeepSeek V4 Flash	**90.7**	89.9–91.5	$0.011	96	2
9	Claude Sonnet 4.6	**90.1**	89.2–90.9	$0.350	35	**5**
10	Gemini 3.1 Pro Preview	**89.3**	88.4–90.1	$0.770	76	**5**
11	Gemini 3.5 Flash	**88.9**	87.3–90.1	$0.590	138	**4**

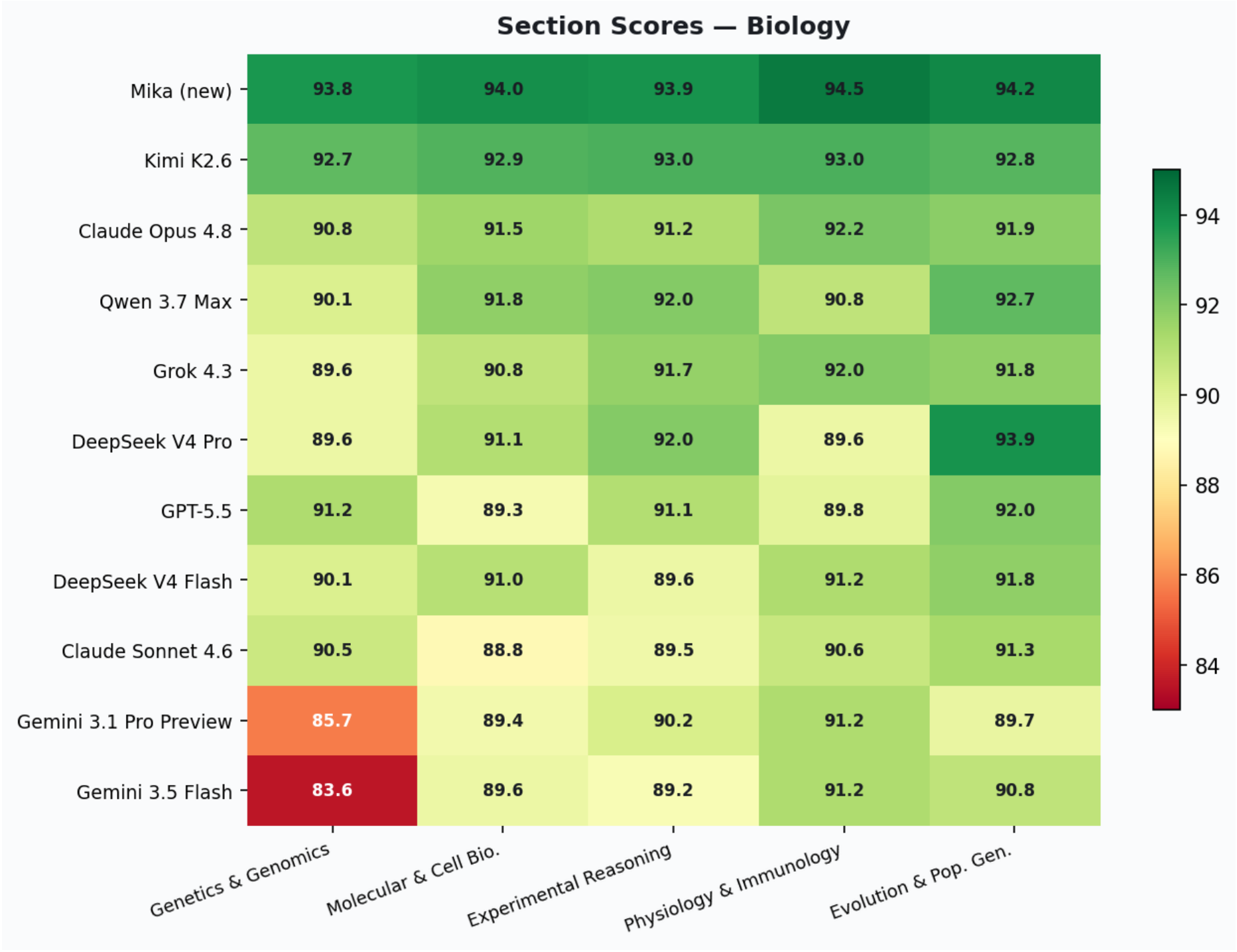
1. Overall Scores

Biology scores are uniformly lower than maths and physics. Only Mika (94.1) and Kimi (92.9) break above 92. GPT-5.5 performs relatively better in biology than physics — scoring above Sonnet 4.6 here for the first time.



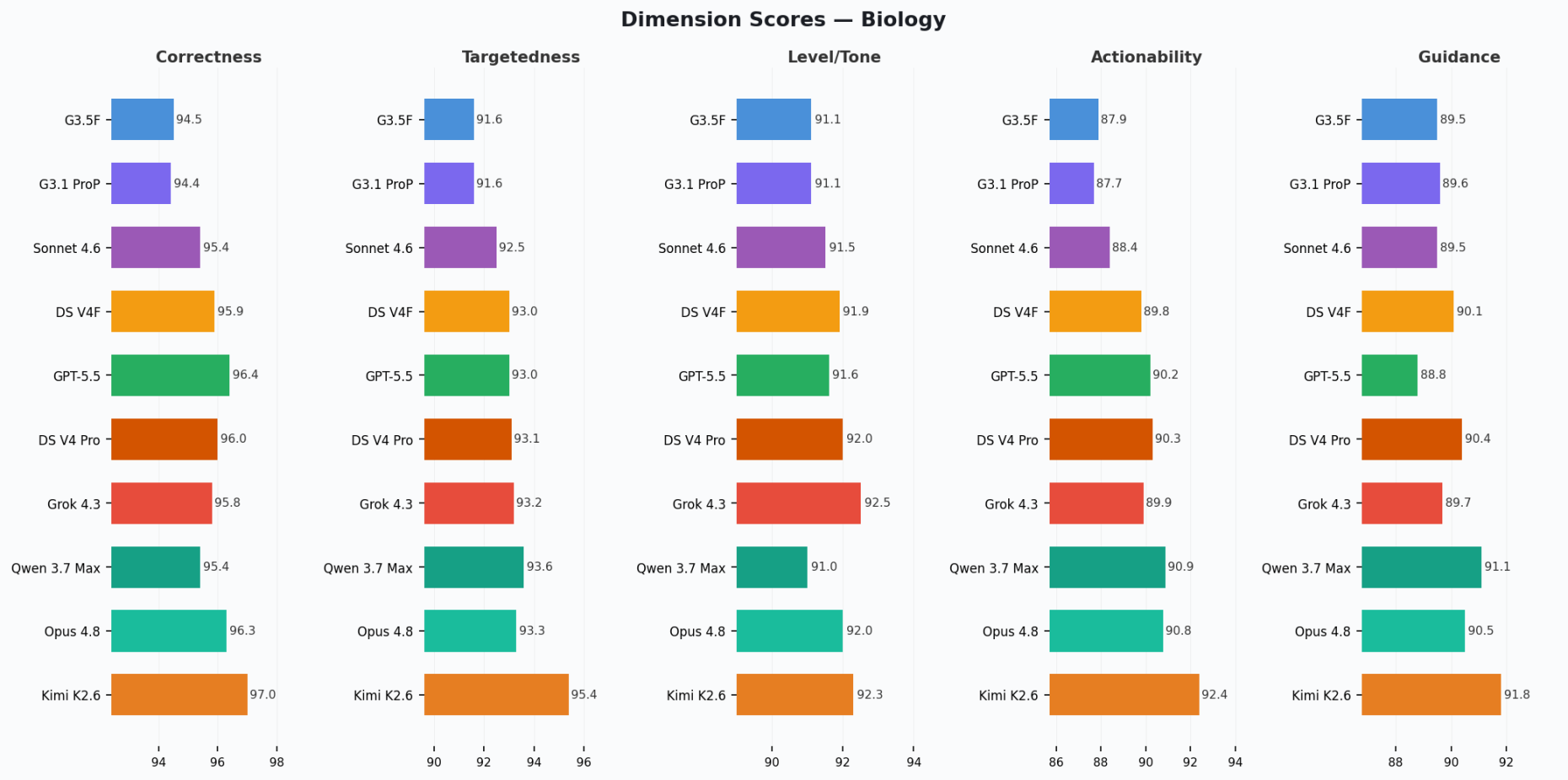
2. Section Scores

Genetics & Genomics is the weakest section for almost every model — the biology equivalent of Improper Integrals (maths) and Energy & Rotation (physics). Physiology & Immunology and Evolution tend to be strongest. Mika is the only model without a section below 93.8.


Section Scores — Biology
Mika (new)
Kimi K2.6
Claude Opus 4.8
Qwen 3.7 Max
Grok 4.3
DeepSeek V4 Pro
GPT-5.5
DeepSeek V4 Flash
Claude Sonnet 4.6
Gemini 3.1 Pro Preview
Gemini 3.5 Flash
93.8 94.0 93.9 94.5 94.2
92.7 92.9 93.0 93.0 92.8
90.8 91.5 91.2 92.2 91.9
90.1 91.8 92.0 90.8 92.7
89.6 90.8 91.7 92.0 91.8
89.6 91.1 92.0 89.6 93.9
91.2 89.3 91.1 89.8 92.0
90.1 91.0 89.6 91.2 91.8
90.5 88.8 89.5 90.6 91.3
85.7 89.4 90.2 91.2 89.7
83.6 89.6 89.2 91.2 90.8
Genetics & Genomics
Molecular & Cell Bio.
Experimental Reasoning
Physiology & Immunology
Evolution & Pop. Gen.
94
92
90
88
86
84

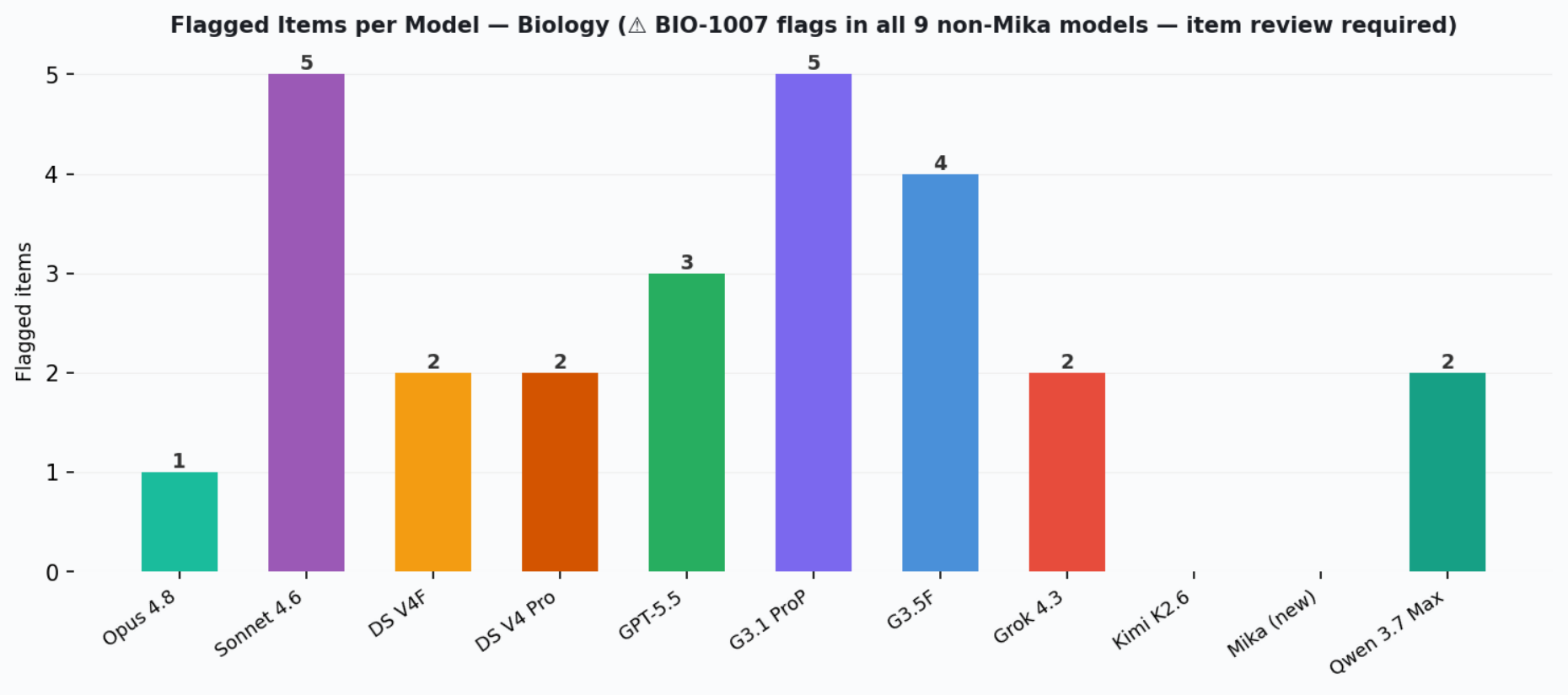
3. Dimension Scores

Actionability and Guidance are universally weakest — the same pattern seen across all three subjects. Biology shows slightly lower Actionability than maths or physics, suggesting that tutoring misconceptions in biological contexts requires more concrete next-step guidance than models currently provide.



4. Flagged Items

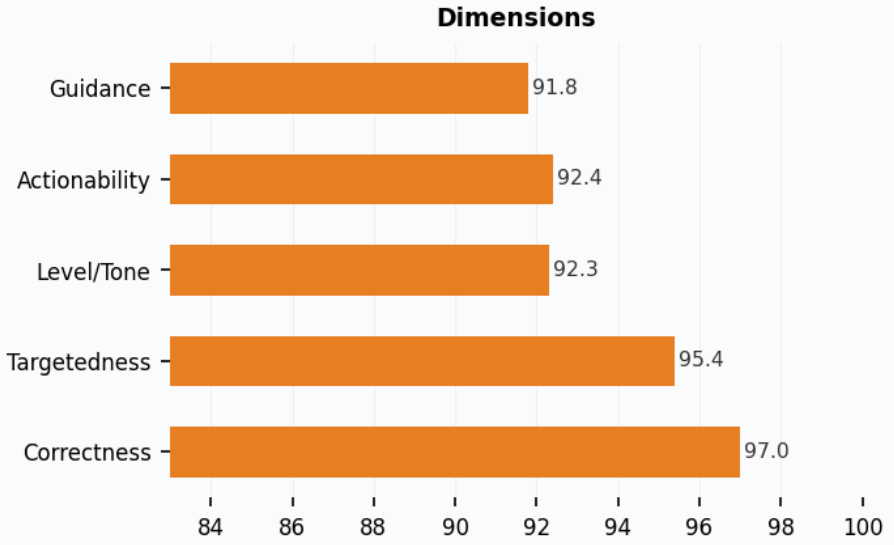
Total: 26 flags across 11 models. Gemini 3.1 Pro and Sonnet 4.6 lead with 5 each. Mika and Kimi have zero flags.

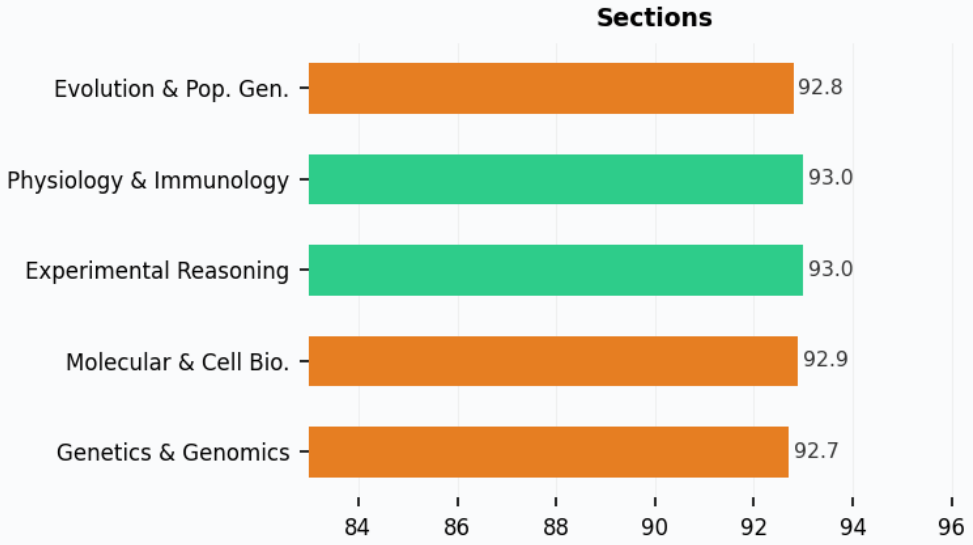


5. Individual Model Cards

Kimi K2.6

Score: **92.9** CI 92.4–93.3 ✓ **None**

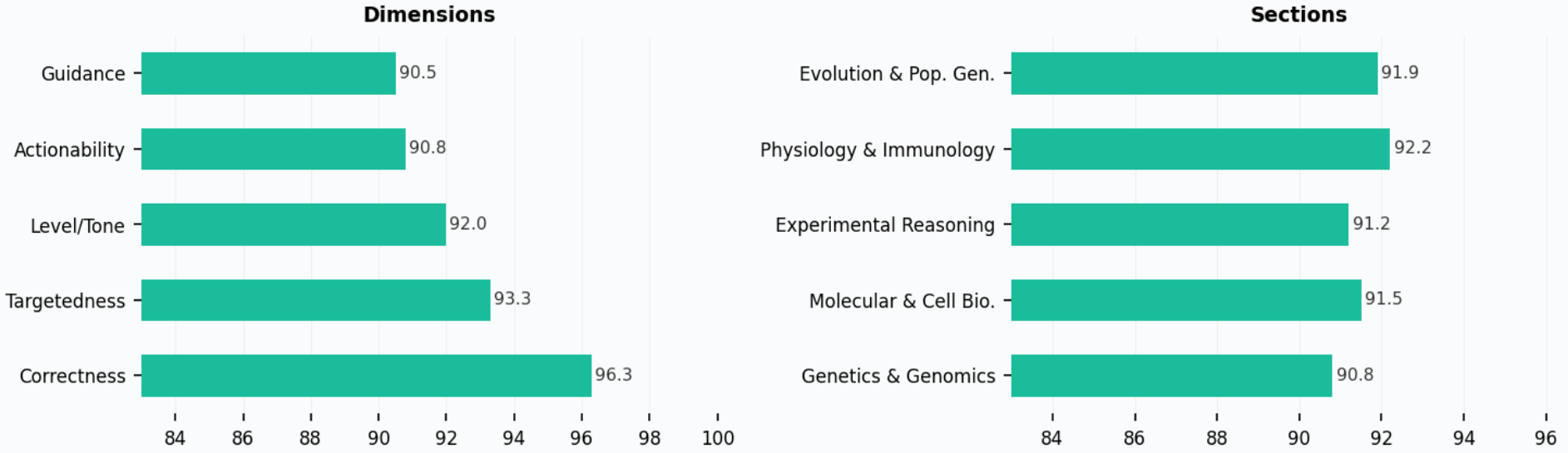




Cost: $0.540 Speed: 210 tok/s Flagged: 0

Claude Opus 4.8

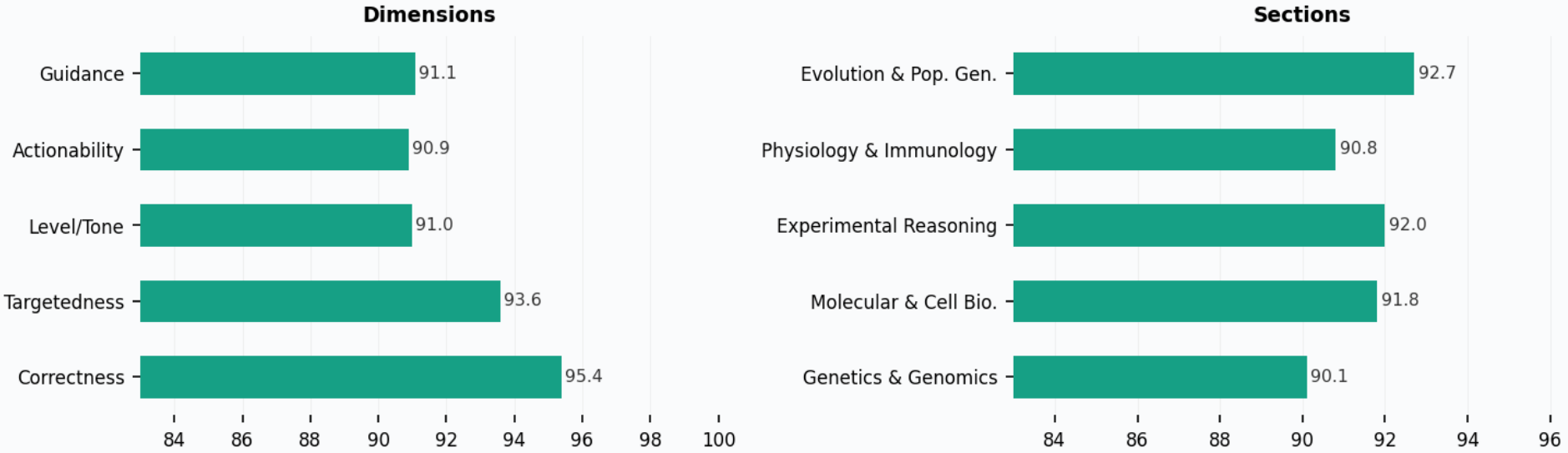
Score: **91.6** CI 90.8–92.1 ⚠ **1 item**



Cost: $0.960 Speed: 56 tok/s Flagged: 1

Qwen 3.7 Max

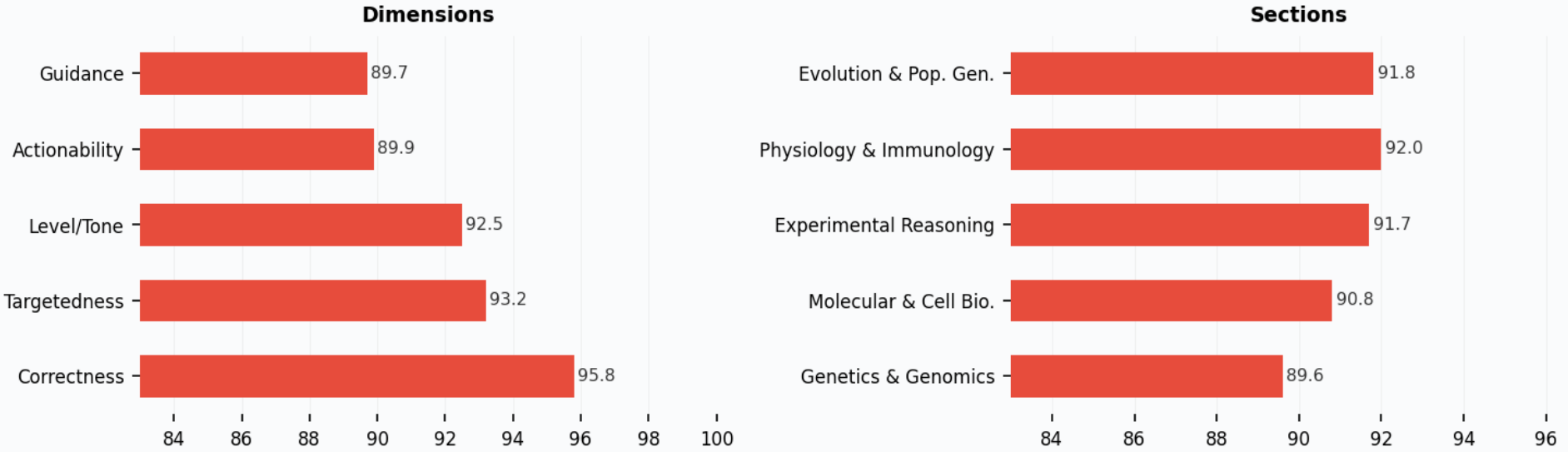
Score: **91.5** CI 90.7–92.0 ⚠ **2 items**



Cost: $0.470 Speed: 52 tok/s Flagged: 2

Grok 4.3

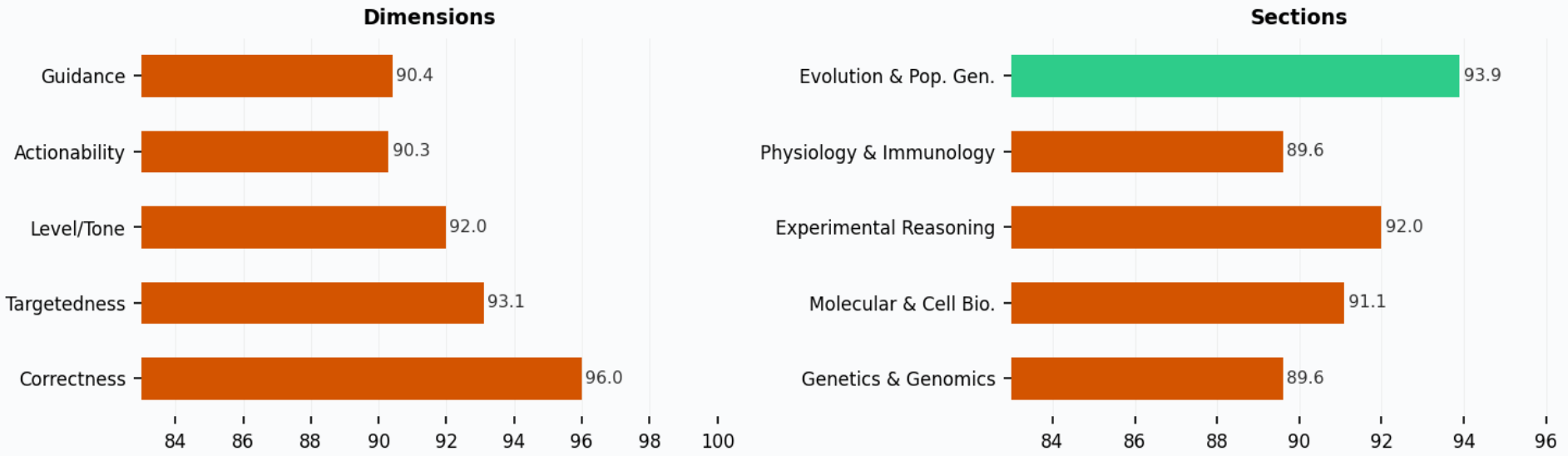
Score: **91.2** CI 90.5–91.7 ⚠ **2 items**



Cost: $0.080 Speed: 146 tok/s Flagged: 2

DeepSeek V4 Pro

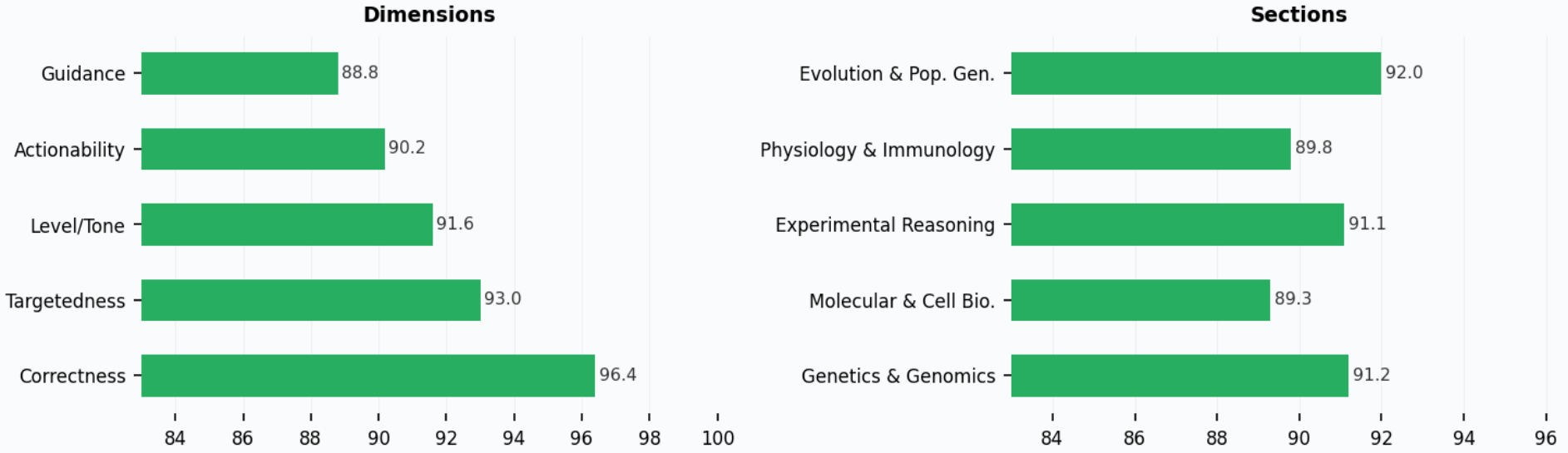
Score: **91.2** CI 89.9–92.3 ⚠ **2 items**



Cost: $0.180 Speed: 48 tok/s Flagged: 2

GPT-5.5

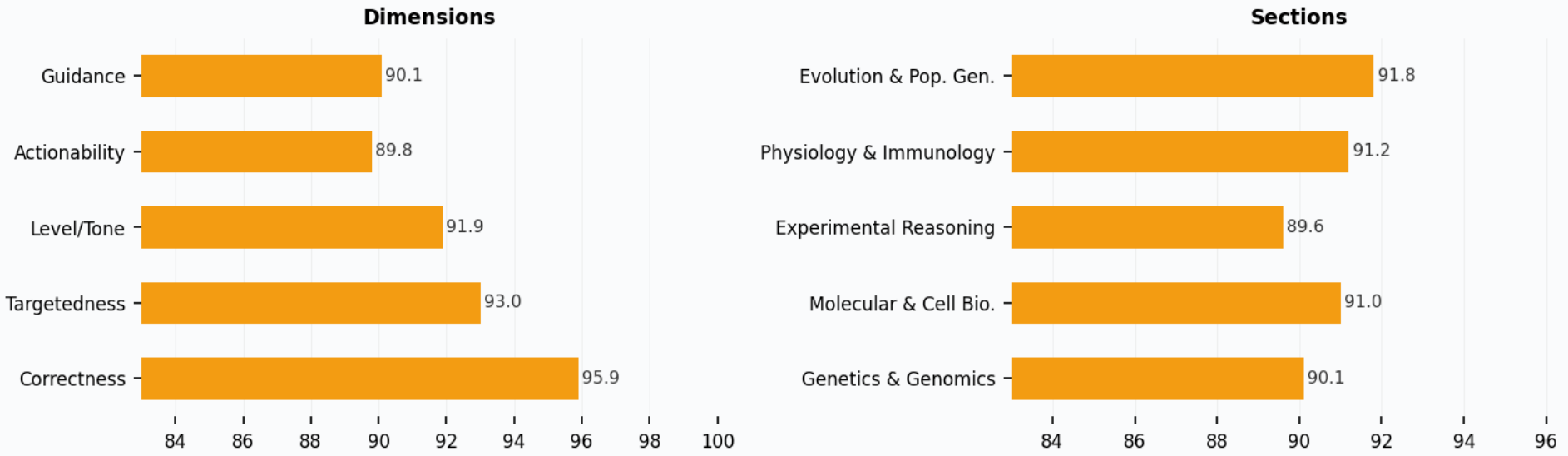
Score: **90.7** CI 89.3–91.7 ⚠ **3 items**



Cost: $0.830 Speed: 41 tok/s Flagged: 3

DeepSeek V4 Flash

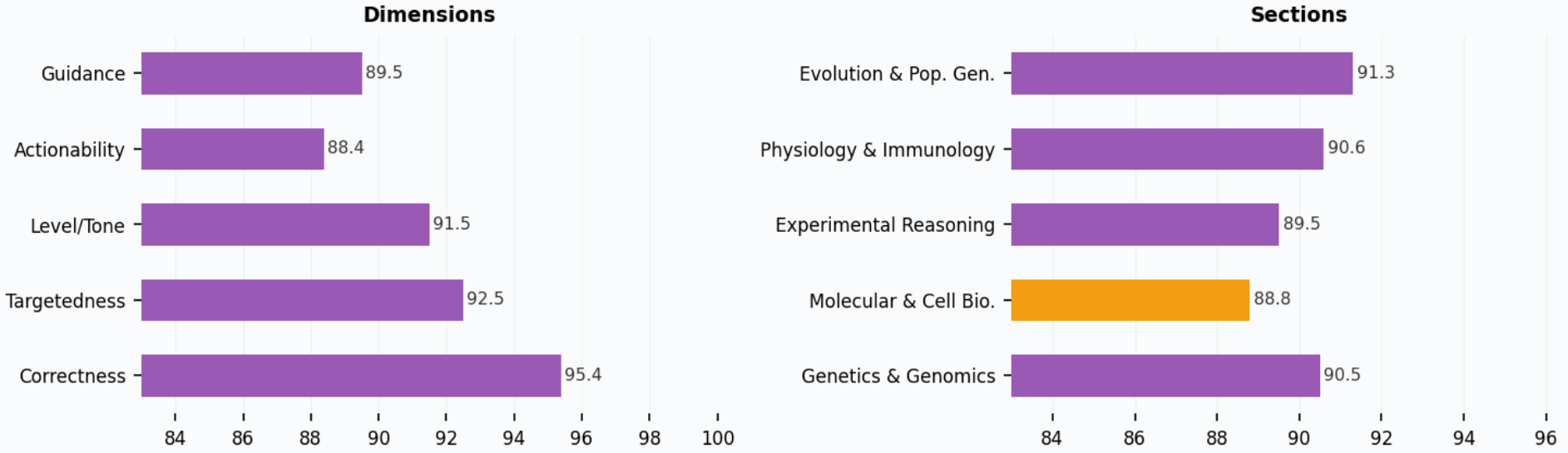
Score: **90.7** CI 89.9–91.5 ⚠ **2 items**



Cost: $0.011 Speed: 96 tok/s Flagged: 2

Claude Sonnet 4.6

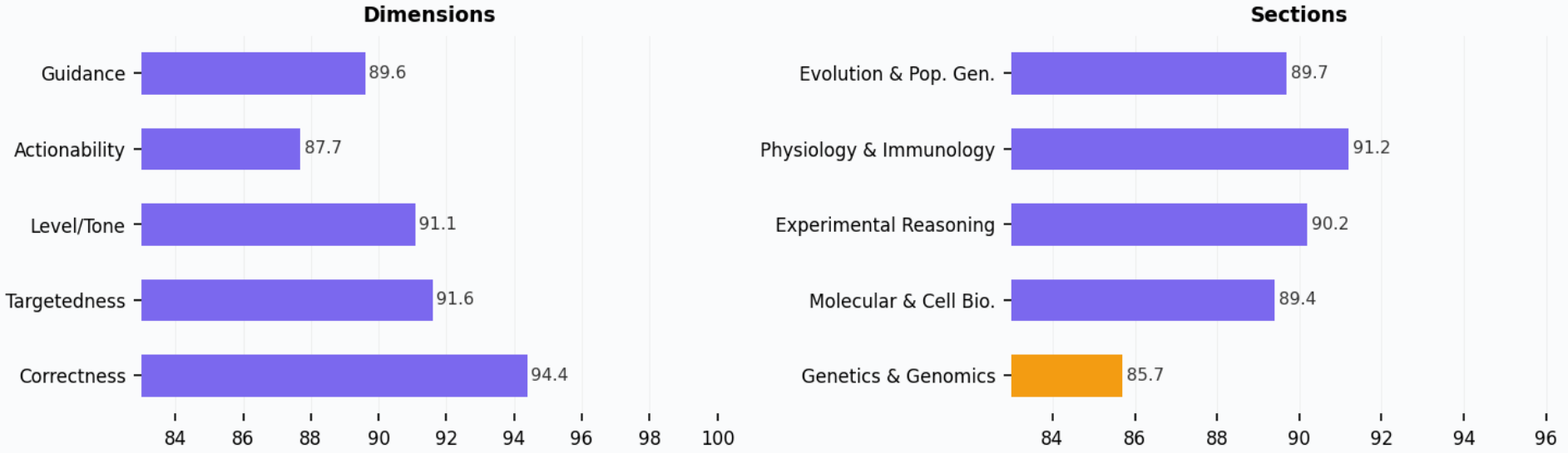
Score: **90.1** CI 89.2–90.9 ⚠ **5 items**



Cost: $0.350 Speed: 35 tok/s Flagged: 5

Gemini 3.1 Pro Preview

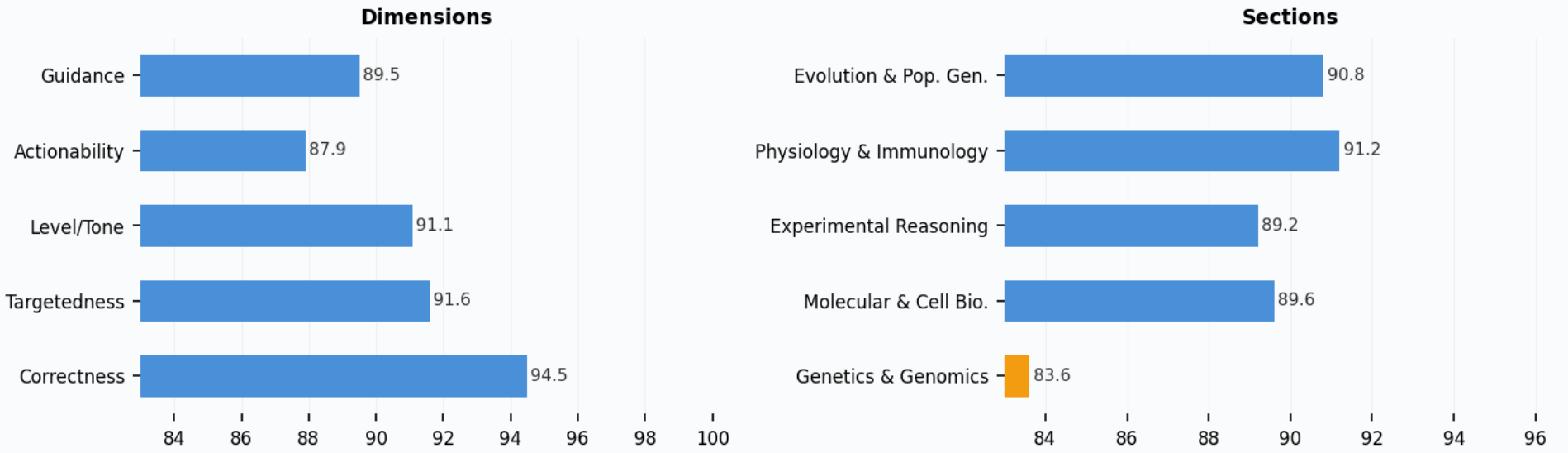
Score: **89.3** CI 88.4–90.1 ⚠ **5 items**



Cost: $0.770 Speed: 76 tok/s Flagged: 5

Gemini 3.5 Flash

Score: **88.9** CI 87.3–90.1 ⚠ **4 items**



Cost: $0.590 Speed: 138 tok/s Flagged: 4

All scores are LLM-judged by openai/o4-mini using judge-v0.3-100pt. Reasoning OFF only — no reasoning-on condition run for Biology. Results should be validated against human ratings before publication. Mika cost is proprietary and not disclosed. Report generated: 2026-06-06 · RedPenBench v1 · Biology.