# RedPenBench Results

Chemistry · 9th of June 2026

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

Chemistry | Reasoning OFF

**Mika leads with 94.0** — the highest chemistry score and zero flagged items. Kimi K2.6 (93.1) and Claude Opus 4.8 (92.8) are the closest competitors. Chemistry is the cleanest subject in this benchmark — 7 of 11 models have zero flagged items, and the total flag count (6) is the lowest across all four subjects.

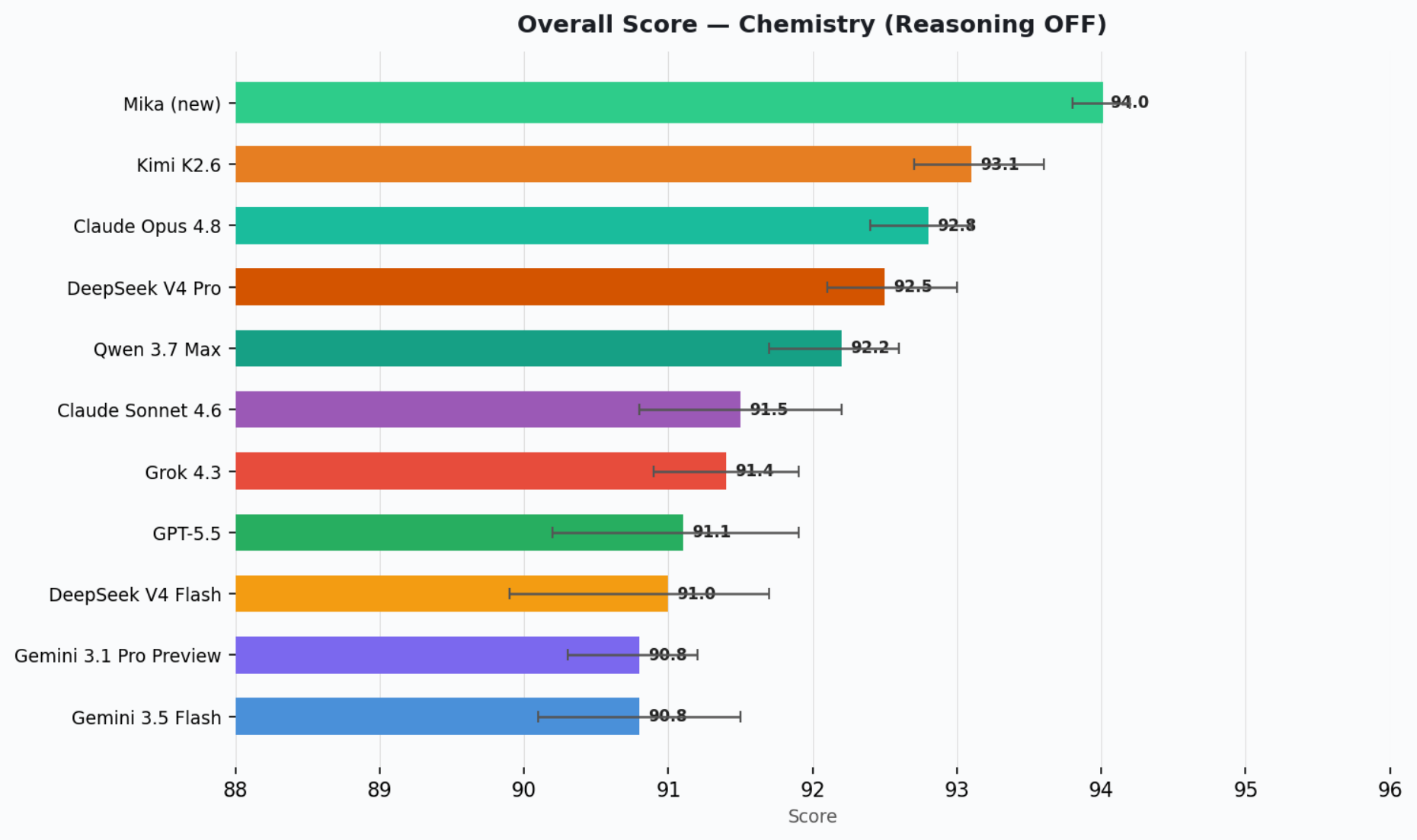
## Summary

All 11 models ranked by score. Mika row highlighted. Reasoning OFF only — no reasoning-on condition run for Chemistry.

| # | Model | Score | CI | Cost | Speed (tok/s) | Flagged |
|---|---|---|---|---|---|---|
| 1 | Mika (new) | **94.0** | 93.8–94.2 | Proprietary | Proprietary | 0 |
| 2 | Kimi K2.6 | **93.1** | 92.7–93.6 | $0.660 | 230 | 0 |
| 3 | Claude Opus 4.8 | **92.8** | 92.4–93.1 | $0.980 | 61 | 0 |
| 4 | DeepSeek V4 Pro | **92.5** | 92.1–93.0 | $0.250 | 66 | 0 |
| 5 | Qwen 3.7 Max | **92.2** | 91.7–92.6 | $0.440 | 61 | 0 |
| 6 | Claude Sonnet 4.6 | **91.5** | 90.8–92.2 | $0.400 | 38 | **2** |
| 7 | Grok 4.3 | **91.4** | 90.9–91.9 | $0.080 | 159 | 0 |
| 8 | GPT-5.5 | **91.1** | 90.2–91.9 | $1.090 | 42 | 1 |
| 9 | DeepSeek V4 Flash | **91.0** | 89.9–91.7 | $0.009 | 62 | **2** |
| 10 | Gemini 3.5 Flash | **90.8** | 90.1–91.5 | $0.630 | 148 | 1 |
| 11 | Gemini 3.1 Pro Preview | **90.8** | 90.3–91.2 | $0.780 | 89 | 0 |

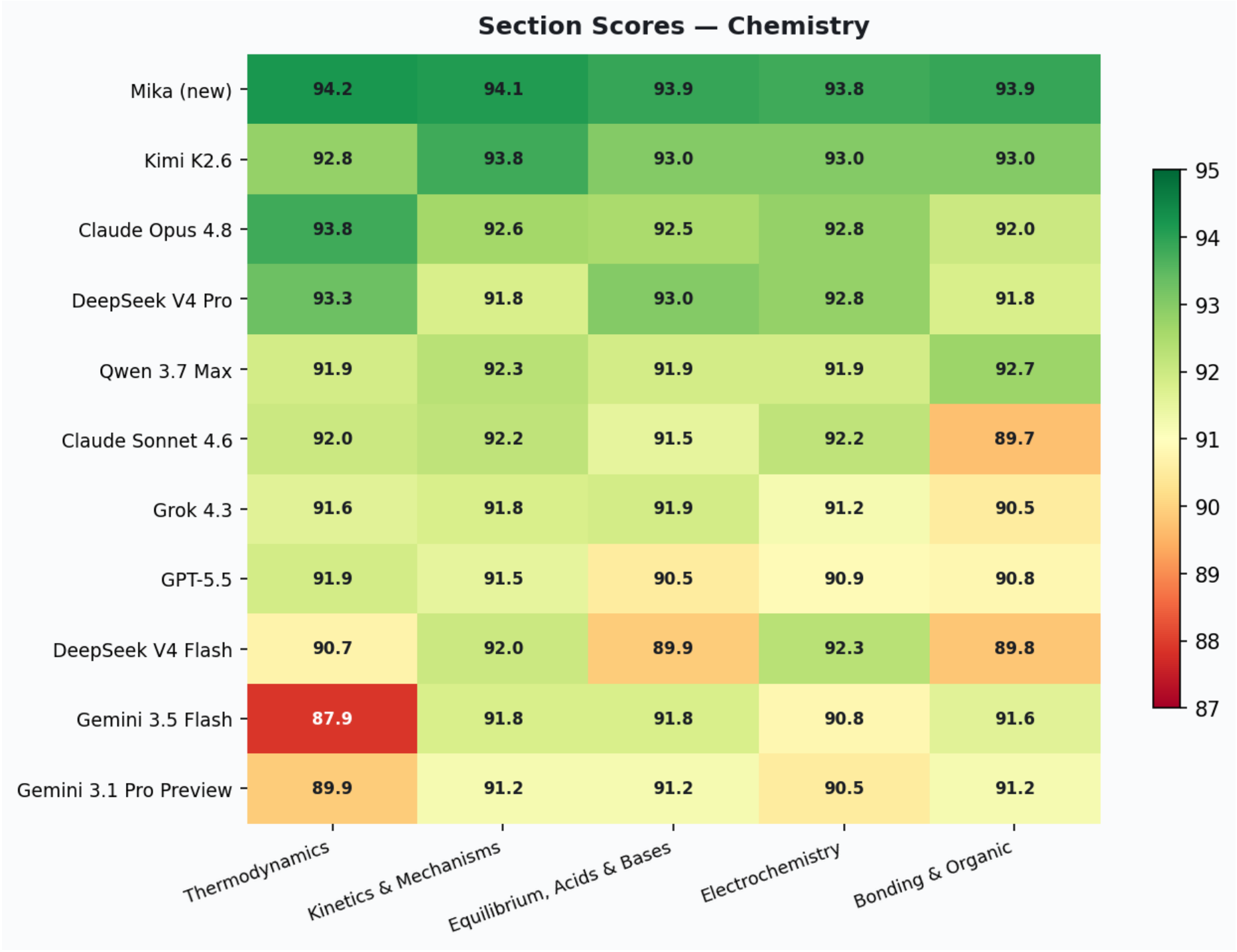
## 1. Overall Scores

Chemistry scores are higher and tighter than biology — the range across models is only 3.2 points (90.8–94.0) vs 5.2 in biology. GPT-5.5 shows its best subject performance, closing the gap on mid-tier models. DeepSeek V4 Flash remains remarkable value at $0.009 despite two flags.



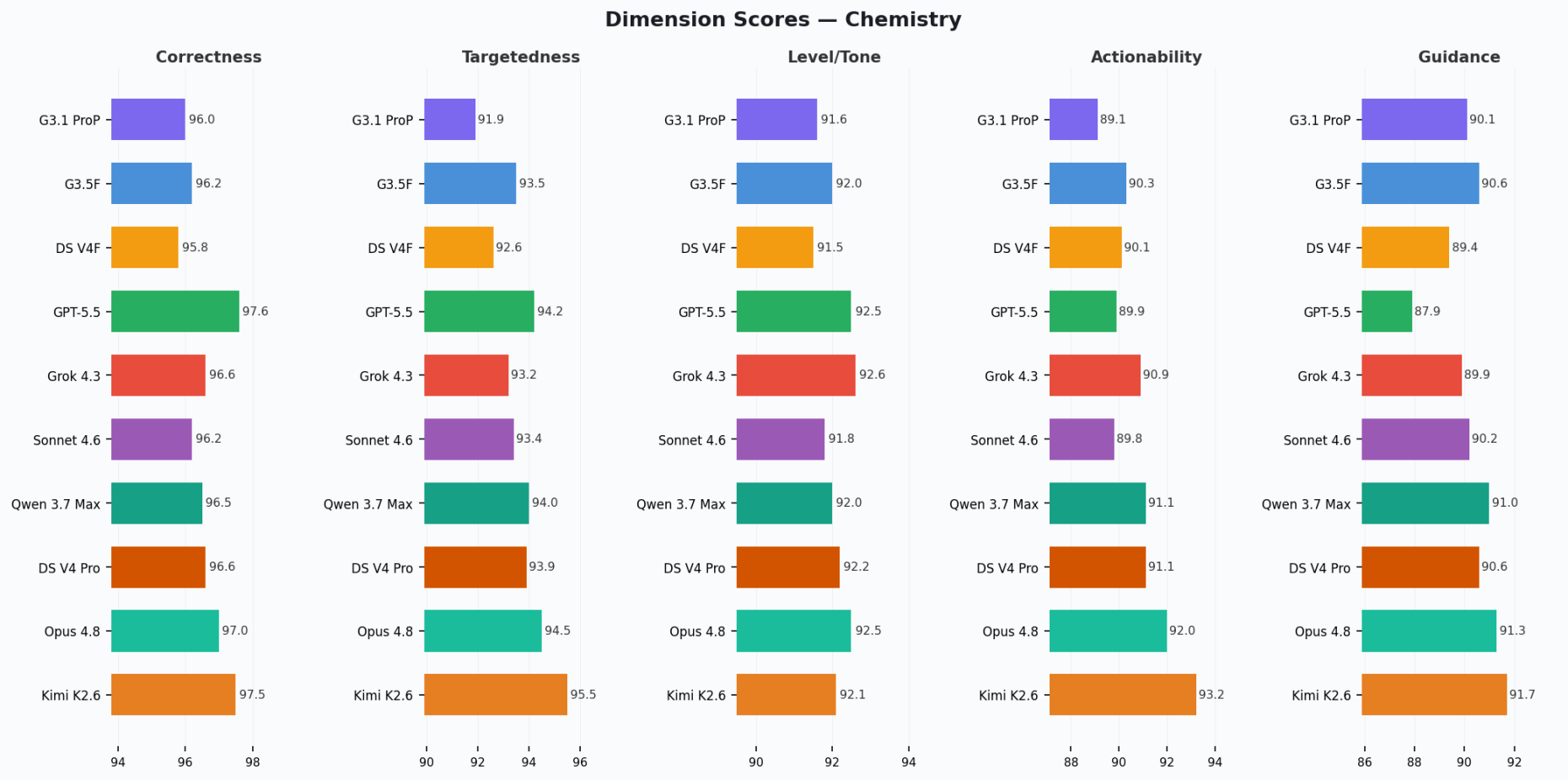
## 2. Section Scores

Thermodynamics & Energetics is the weakest section for Gemini Flash (87.9) — driven by CHEM-1003. Every other model handles that item cleanly, suggesting it is a Flash-specific gap rather than a difficult item. Bonding & Organic is the weakest section for the most models.


Section Scores — Chemistry
Mika (new)
Kimi K2.6
Claude Opus 4.8
DeepSeek V4 Pro
Qwen 3.7 Max
Claude Sonnet 4.6
Grok 4.3
GPT-5.5
DeepSeek V4 Flash
Gemini 3.5 Flash
Gemini 3.1 Pro Preview
94.2 94.1 93.9 93.8 93.9
92.8 93.8 93.0 93.0 93.0
93.8 92.6 92.5 92.8 92.0
93.3 91.8 93.0 92.8 91.8
91.9 92.3 91.9 91.9 92.7
92.0 92.2 91.5 92.2 89.7
91.6 91.8 91.9 91.2 90.5
91.9 91.5 90.5 90.9 90.8
90.7 92.0 89.9 92.3 89.8
87.9 91.8 91.8 90.8 91.6
89.9 91.2 91.2 90.5 91.2
Thermodynamics
Kinetics & Mechanisms
Equilibrium, Acids & Bases
Electrochemistry
Bonding & Organic
95
94
93
92
91
90
89
88
87

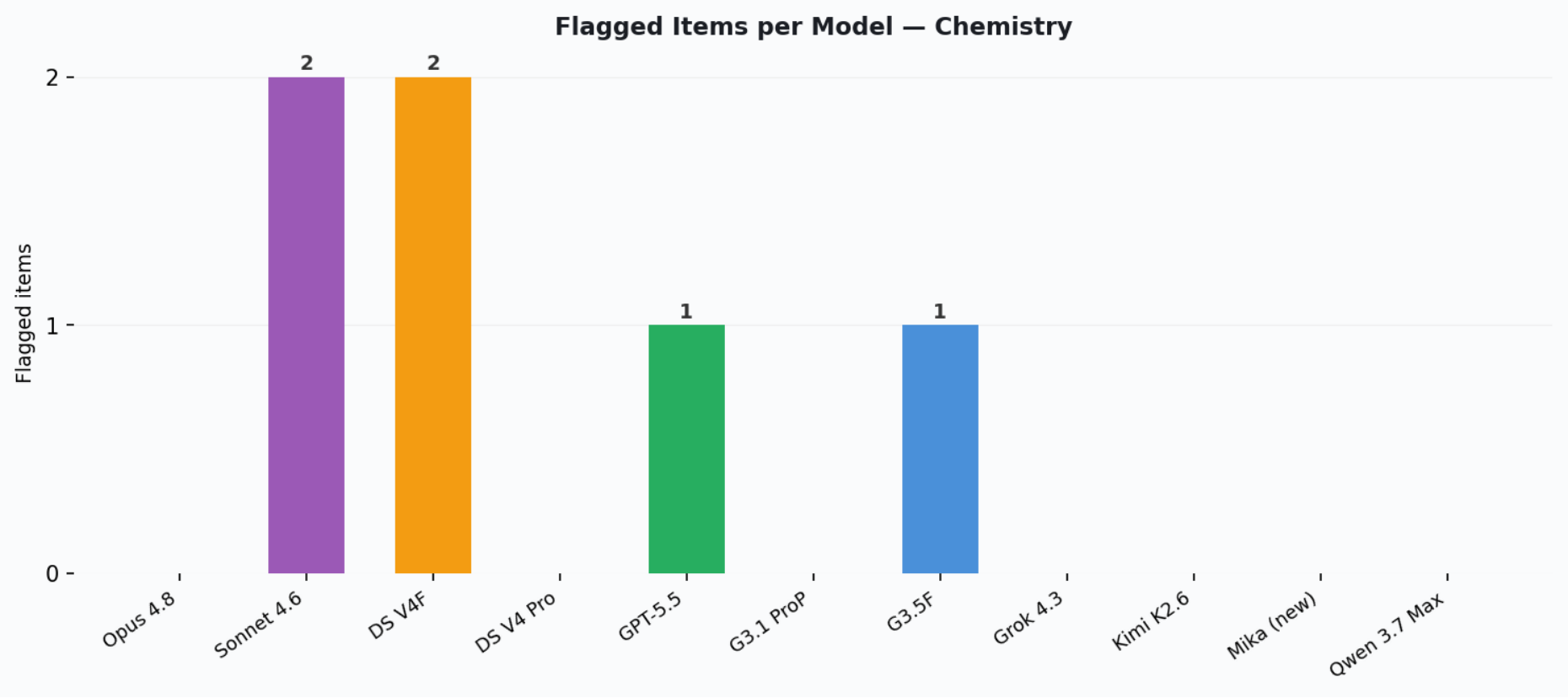
## 3. Dimension Scores

Chemistry shows the highest Correctness scores of any subject — nearly every model sits above 96. Guidance remains the universal weak point. GPT-5.5's Guidance score (87.9) is the only dimension below 88 in the entire chemistry batch.



## 4. Flagged Items

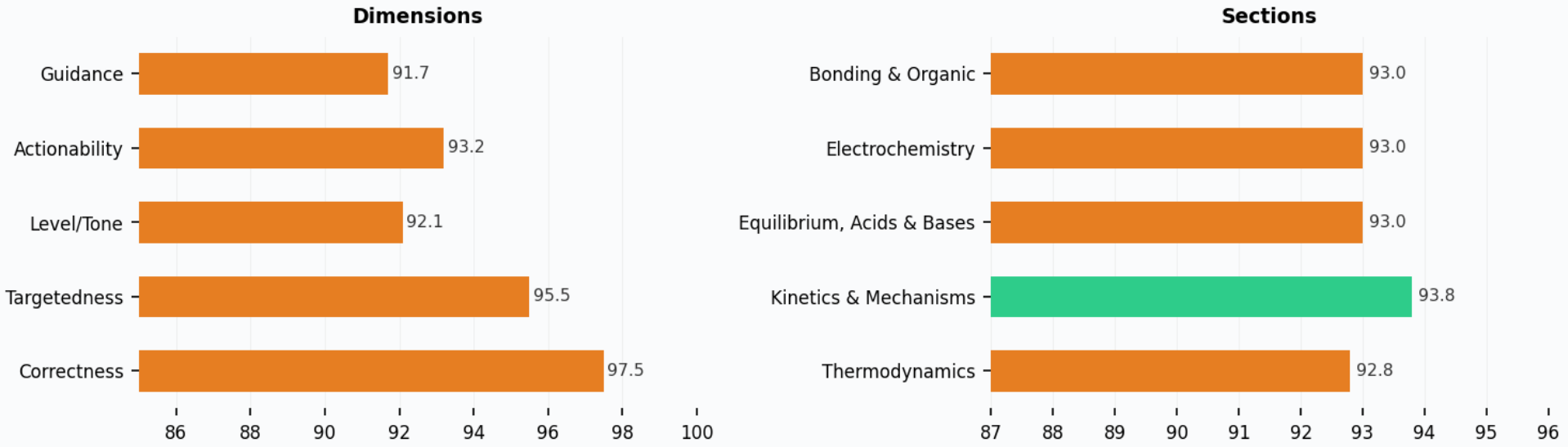
Only 6 flags total across 11 models — the cleanest subject in the benchmark. CHEM-1017 (resonance hybrid misconception) flags in DS Flash and Sonnet 4.6 for the same actionability gap. Mika, Kimi, Opus, Pro Preview, Grok, DeepSeek Pro, and Qwen all have zero flags.



## 5. Individual Model Cards

### Kimi K2.6

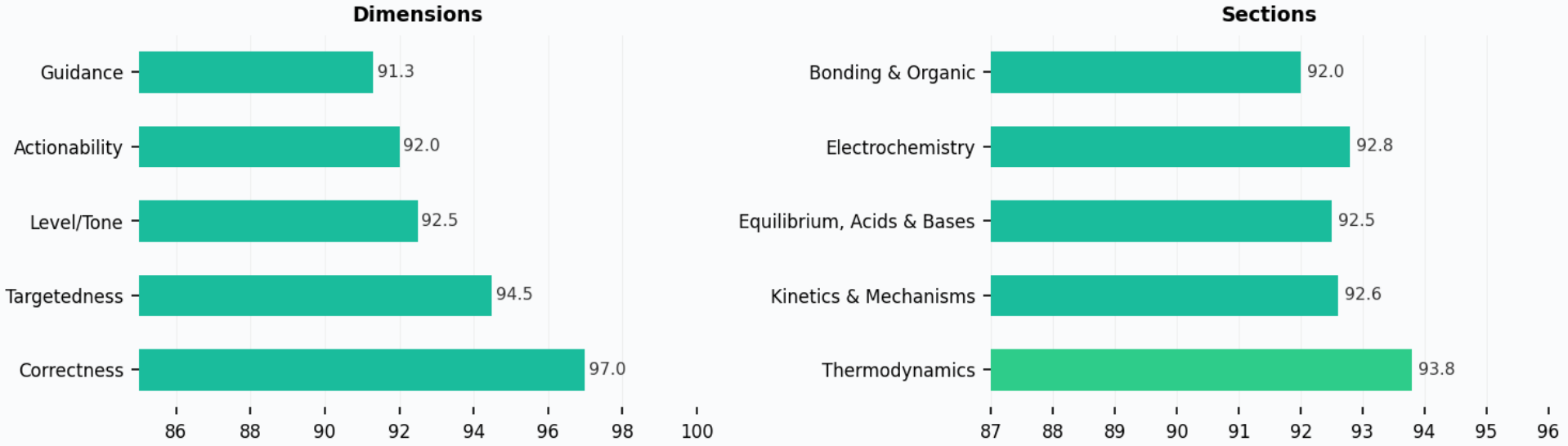
Score: **93.1** CI 92.7–93.6 ✓ **None**




Cost: $0.660 Speed: 230 tok/s Flagged: 0

## Claude Opus 4.8

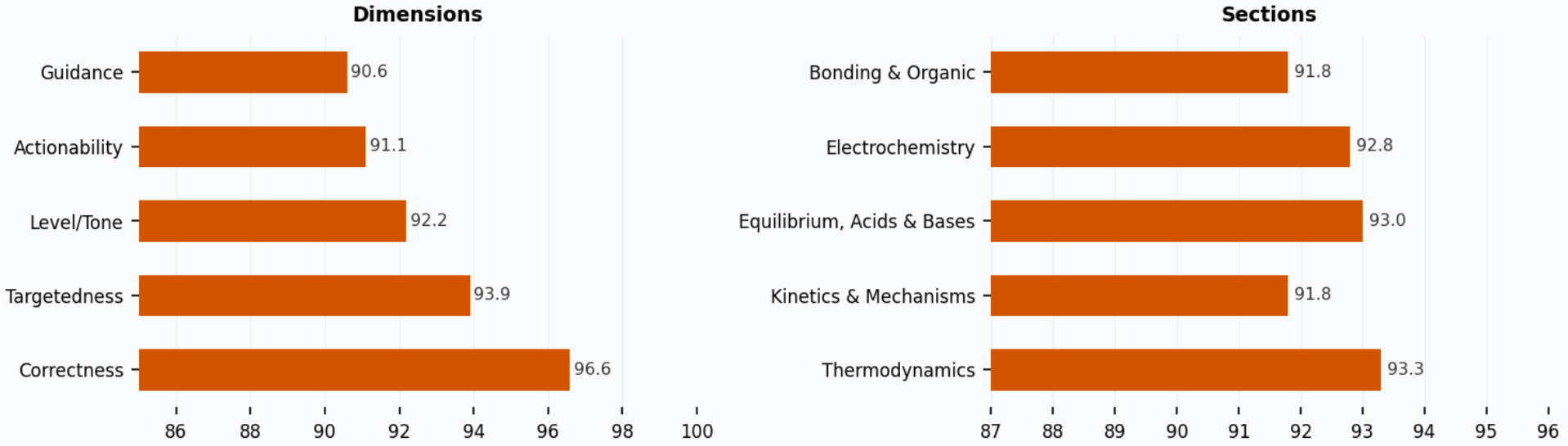
Score: **92.8** CI 92.4–93.1 ✓ **None**




Cost: $0.980 Speed: 61 tok/s Flagged: 0

## DeepSeek V4 Pro

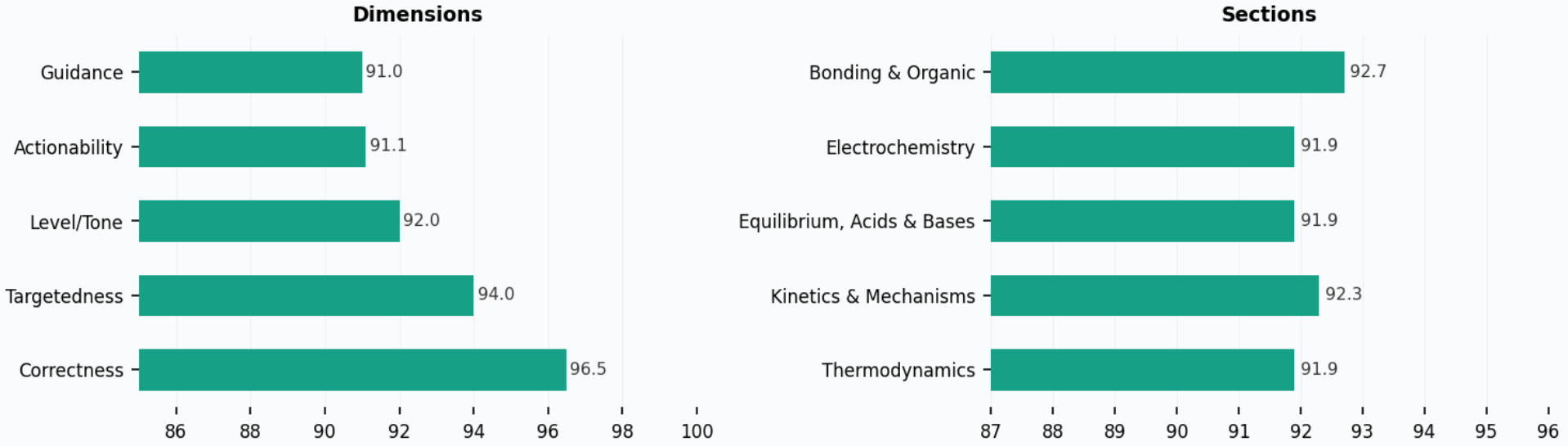
Score: **92.5** CI 92.1–93.0 ✓ **None**




Cost: $0.250 Speed: 66 tok/s Flagged: 0

## Qwen 3.7 Max

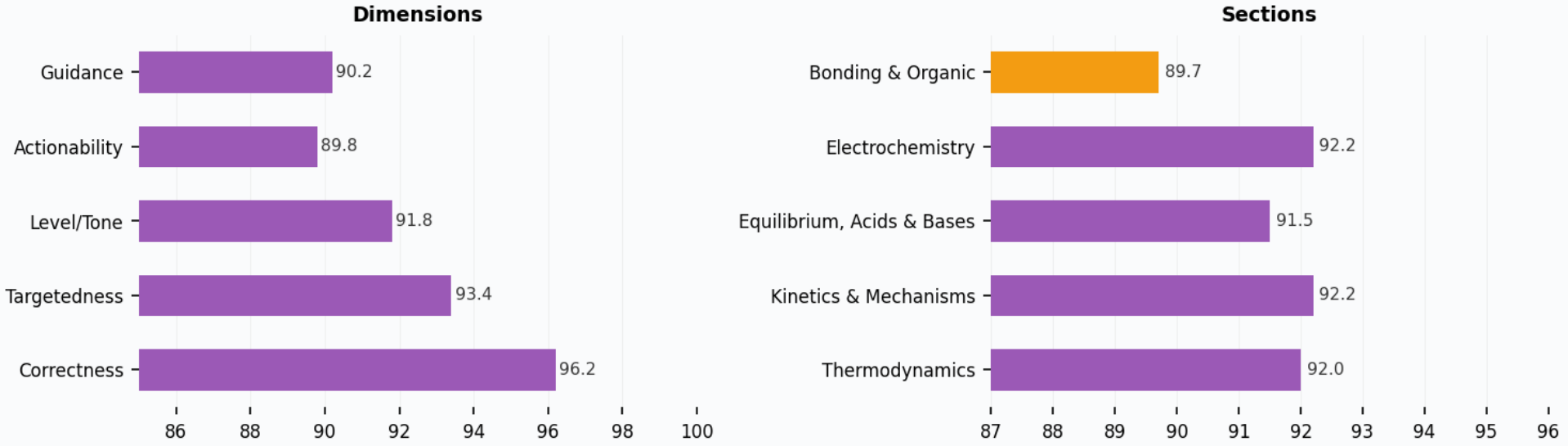
Score: **92.2** CI 91.7–92.6 ✓ **None**




Cost: $0.440 Speed: 61 tok/s Flagged: 0

## Claude Sonnet 4.6

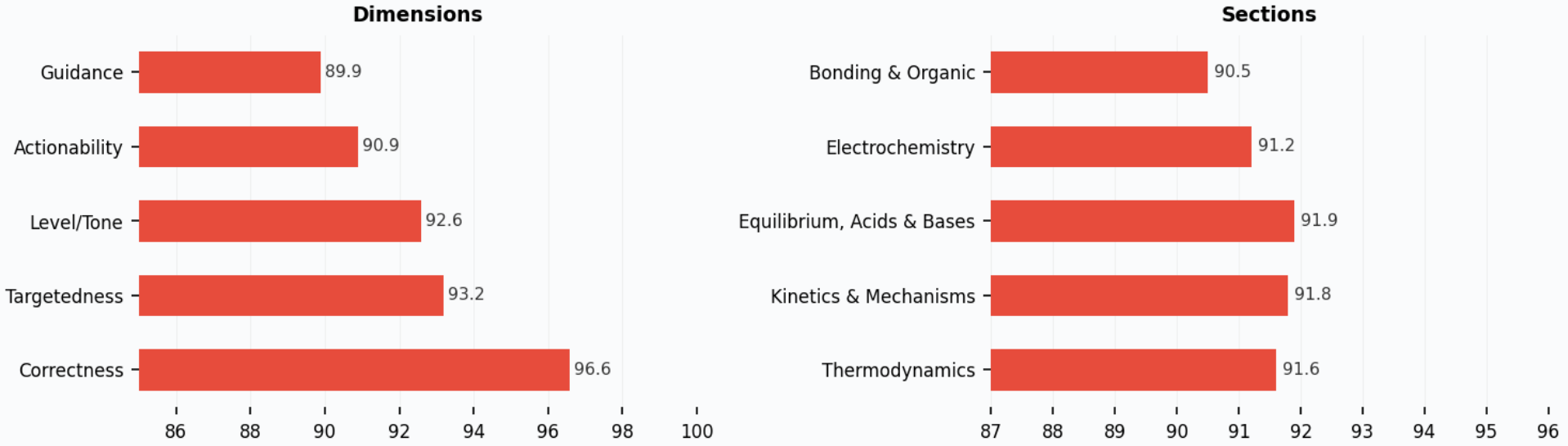
Score: **91.5** CI 90.8–92.2 ⚠ **2 items**




Cost: $0.400 Speed: 38 tok/s Flagged: 2

## Grok 4.3

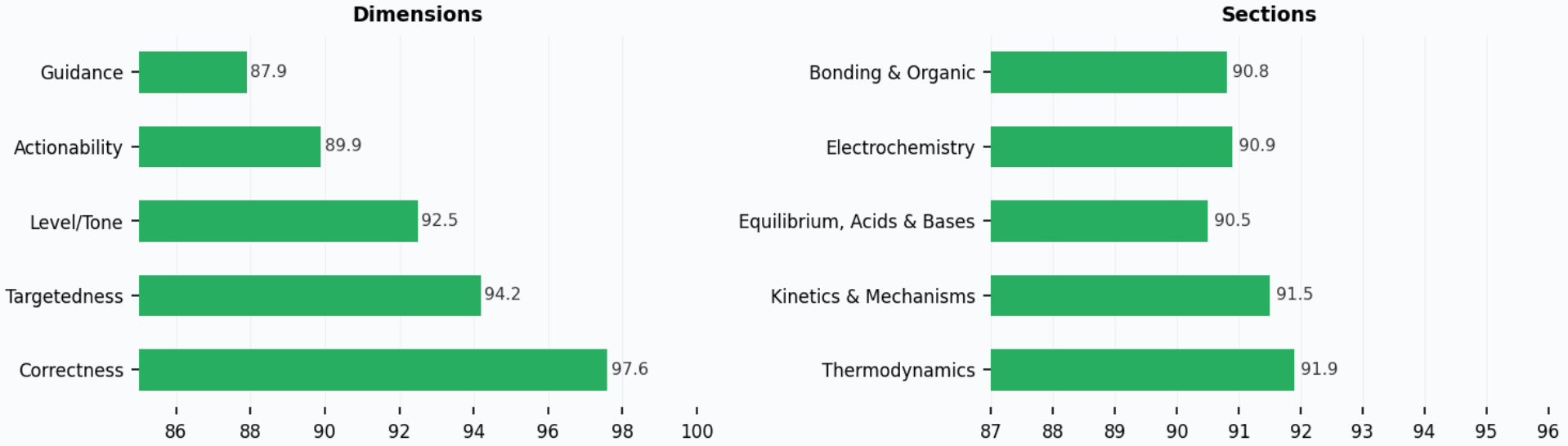
Score: **91.4** CI 90.9–91.9 ✓ **None**




Cost: $0.080 Speed: 159 tok/s Flagged: 0

## GPT-5.5

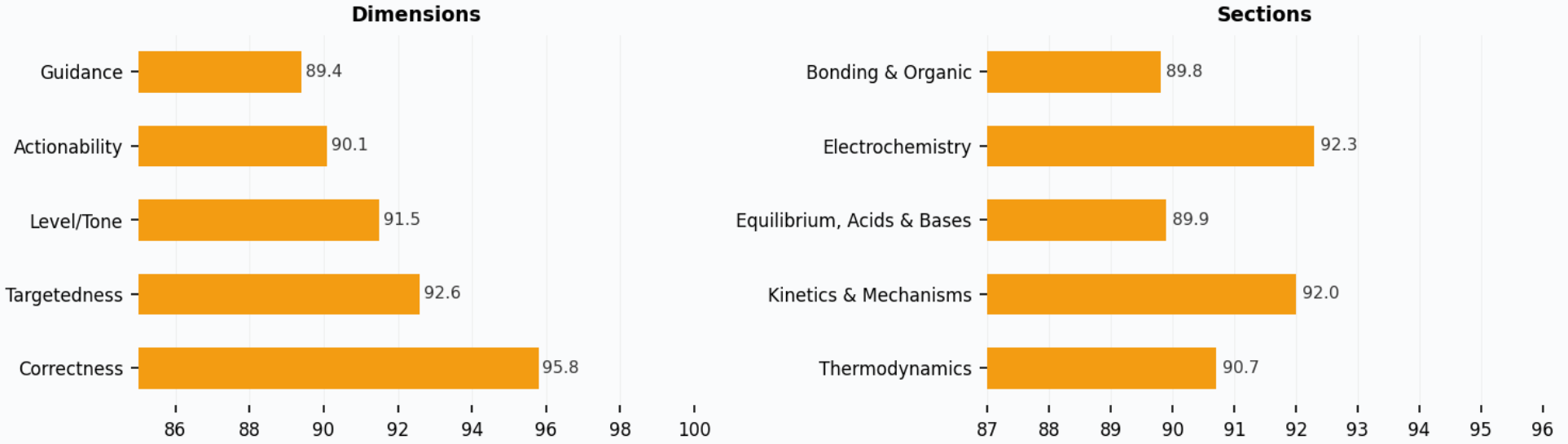
Score: **91.1** CI 90.2–91.9 ⚠ **1 item**




Cost: $1.090 Speed: 42 tok/s Flagged: 1

## DeepSeek V4 Flash

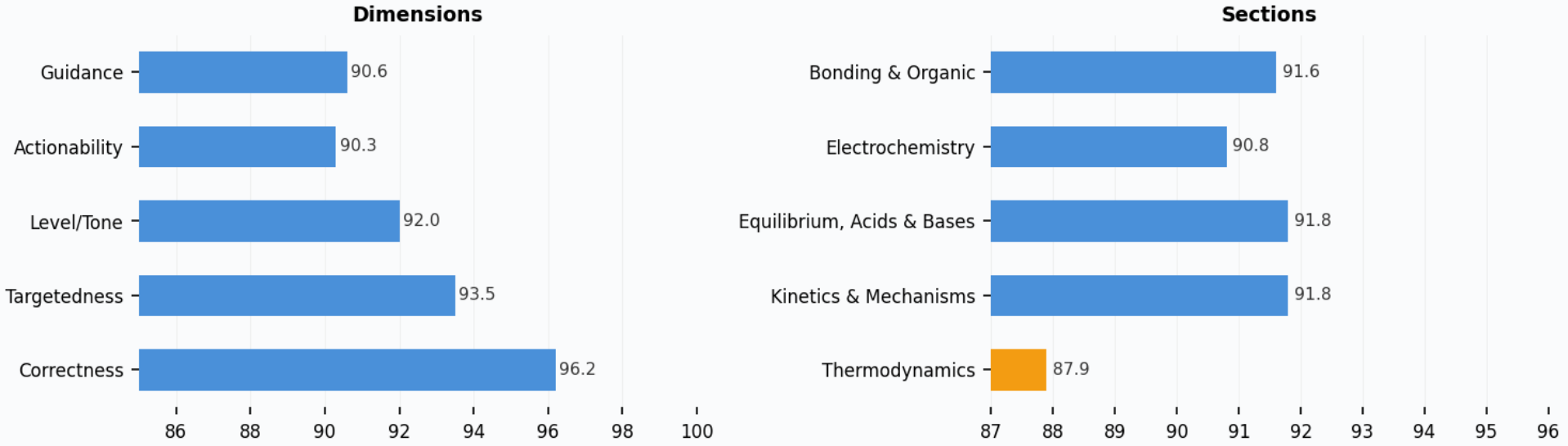
Score: **91.0** CI 89.9–91.7 ⚠ **2 items**




Cost: $0.009 Speed: 62 tok/s Flagged: 2

## Gemini 3.5 Flash

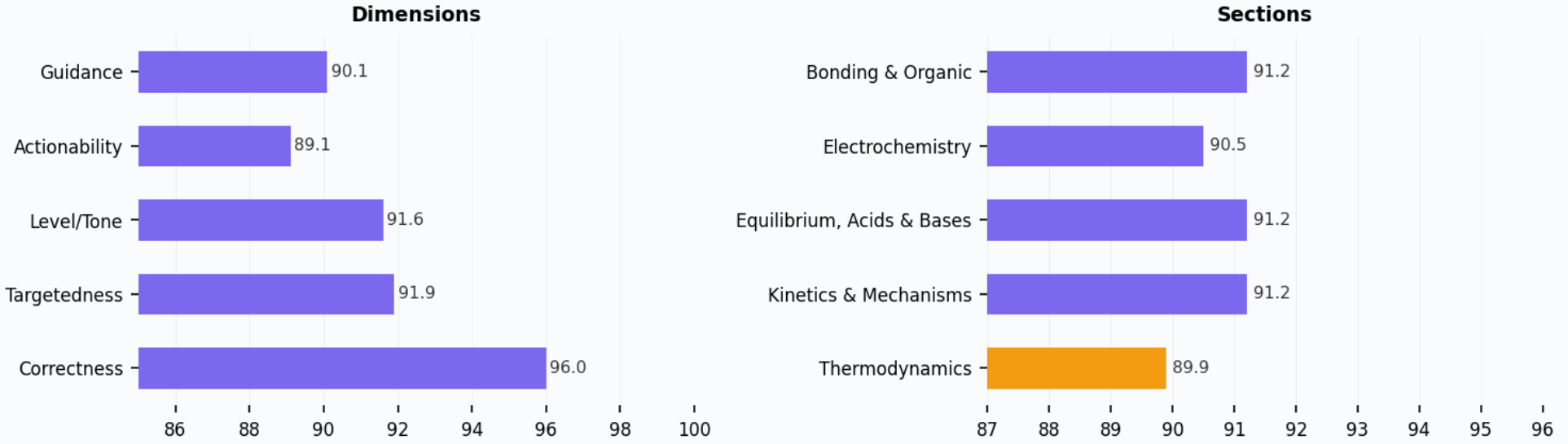
Score: **90.8** CI 90.1–91.5 ⚠ **1 item**




Cost: $0.630 Speed: 148 tok/s Flagged: 1

## Gemini 3.1 Pro Preview

Score: **90.8** CI 90.3–91.2 ✓ **None**




Cost: $0.780 Speed: 89 tok/s Flagged: 0

All scores are LLM-judged by openai/o4-mini using judge-v0.3-100pt. Reasoning OFF only — no reasoning-on condition run for Chemistry. Results should be validated against human ratings before publication. Mika cost and speed are proprietary and not disclosed. Report generated: 2026-06-09 · RedPenBench v1 · Chemistry.