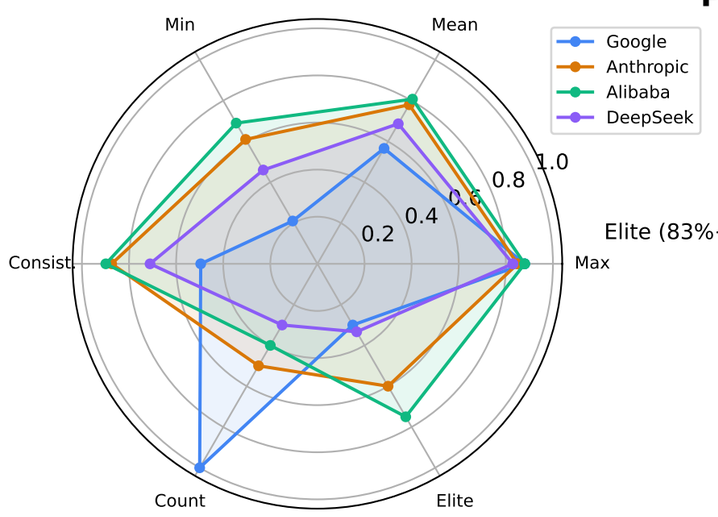
Figure 17: CORE-Bench Novel Insights Dashboard
Comprehensive Analysis of 22 LLMs Across 18 Reasoning Tasks

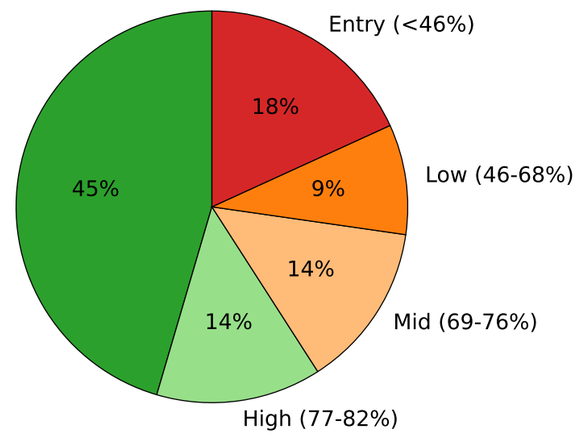
Key Statistics

```
Total Models: 22
Total Tasks: 18

Performance Range:
  • Maximum: 88%
  • Minimum: 21%
  • Mean: 67.5%
  • Median: 77.5%
  • Std Dev: 23.1%

Gini Coefficient: 0.169
(Performance Inequality)
```

Family Profiles



Performance Tier Distribution



Top 5 Models

🥇 Gemini 3 Flash Preview
88% (Google)

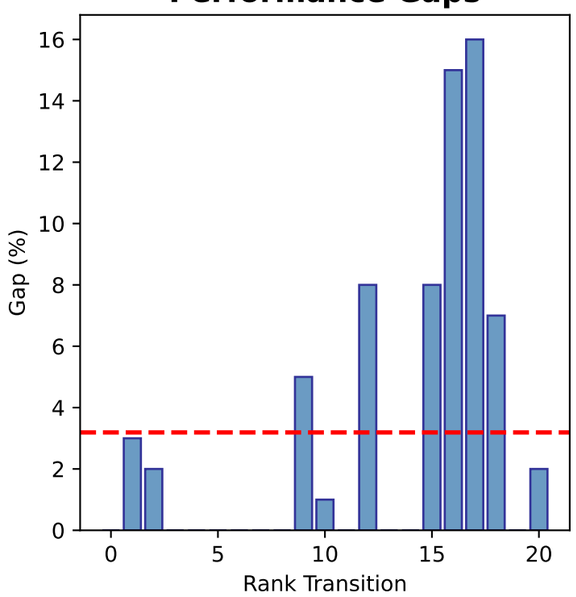
🥈 Qwen 3 Next 80B Thinking
88% (Alibaba)

🥉 Claude Opus 4.1
85% (Anthropic)

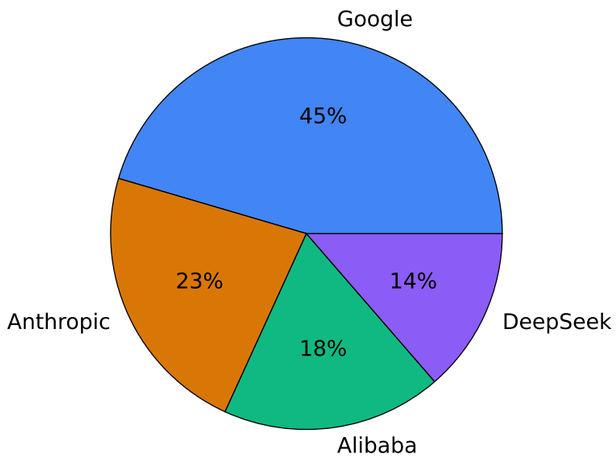
4️⃣ Claude Haiku 4.5
83% (Anthropic)

5️⃣ Claude Sonnet 4.5
83% (Anthropic)

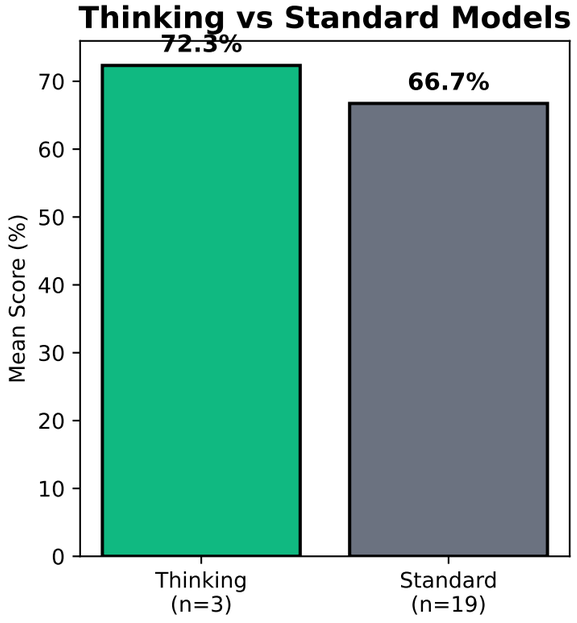
Performance Gaps



Model Distribution by Family



Thinking vs Standard Models



Novel Findings

1. Pareto Analysis: Top 8 models contribute 80% of total perf.

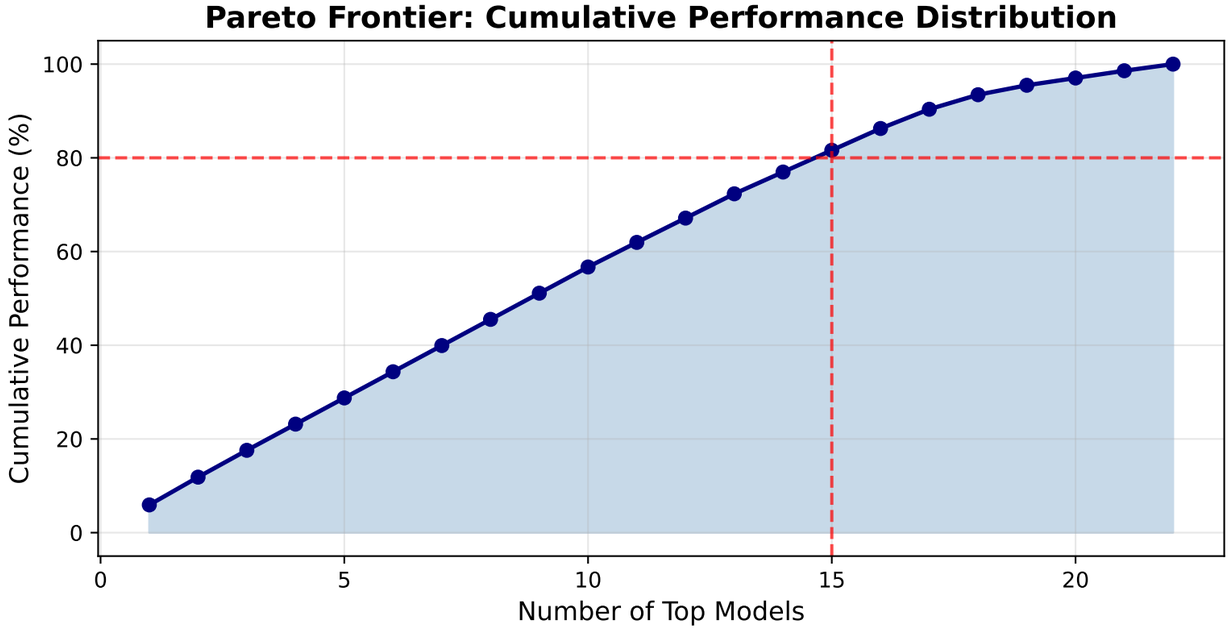
2. Natural Clusters: 2 clusters identified via silhouette

3. Market Concentration: Elite tier highly concentrated

4. Scaling Law: ~7.9% gain per doubling of size

5. Rank Volatility: Mid-tier positions most uncertain

Pareto Frontier: Cumulative Performance Distribution



Complete Leaderboard (Abbreviated)

Rank	Model	Score	Family
1	Gemini 3 Flash Preview	88%	Google
2	Qwen 3 Next 80B Thinki	88%	Alibaba
3	Claude Opus 4.1	85%	Anthropic
4	Claude Haiku 4.5	83%	Anthropic
5	Claude Sonnet 4.5	83%	Anthropic
6	Deepseek V3.1	83%	DeepSeek
7	Gemini 2.5 Flash	83%	Google
8	Gemini 3 Pro Preview	83%	Google
9	Qwen 3 Coder 480B	83%	Alibaba
10	Qwen 3 Next 80B Instru	83%	Alibaba
11	Claude Opus 4.5	78%	Anthropic
...	...	...	...
21	Gemma 3 4B	23%	Google
22	Gemma 3 1B	21%	Google