EPHOR SLINGSHOT

Model Comparison Leaderboard Report

Generated: 12/4/2025

Summary Statistics

Metric	Value
Total Tests	1029
DeepSeek Win Rate	46.7%
DeepSeek Top-2 Rate	80.8%
DeepSeek Parity Rate	50.3%

Consensus Rankings

Model	Wins	Top-2	Avg Rank	Points
DeepSeek V3.2	518	831	1.73	4379
Gemini 3 Pro	354	773	1.95	4172
GPT-5	135	539	2.44	3667
Claude Opus 4.5	157	506	2.48	3618

Suite Breakdown

Suite	Tests	Win%	Top-2%	Parity%
Math & Reasoning	299	45.8%	78.3%	50.2%
Coding	278	54.7%	89.2%	57.2%
Creative Writing	231	26.4%	65.8%	30.3%
General Knowledge	121	52.1%	89.3%	56.2%
agentic-tasks	100	68.0%	89.0%	71.0%

Speed Rankings

Model	Avg Latency	Tests
Gemini 3 Pro	5.59s	1029
Claude Opus 4.5	9.29s	1029
GPT-5	11.25s	1027
DeepSeek V3.2	14.24s	1025

Cost Rankings

Model	Avg Cost	Tests
DeepSeek V3.2	$0.000370	1025
GPT-5	$0.005835	1027
Gemini 3 Pro	$0.009018	1029
Claude Opus 4.5	$0.011015	1029

Factual Accuracy Rankings (Objective Questions)

Model	Accuracy	Correct	Partial	Incorrect	Value Wins
Claude Opus 4.5	126.7%	739 correct	211 partial	11 incorrect	94
DeepSeek V3.2	115.4%	738 correct	124 partial	9 incorrect	379
GPT-5	103.9%	718 correct	61 partial	32 incorrect	90
Gemini 3 Pro	100.0%	627 correct	123 partial	123 incorrect	168

Cost Per Correct Answer

Model	Cost/Correct	
DeepSeek V3.2	$0.0004	Best Value
GPT-5	$0.0053	
Gemini 3 Pro	$0.0084	
Claude Opus 4.5	$0.0109	