Context Arena

Theme: System

OpenAI-MRCR | Fiction.LiveBench

Needles: 2 (Easy)

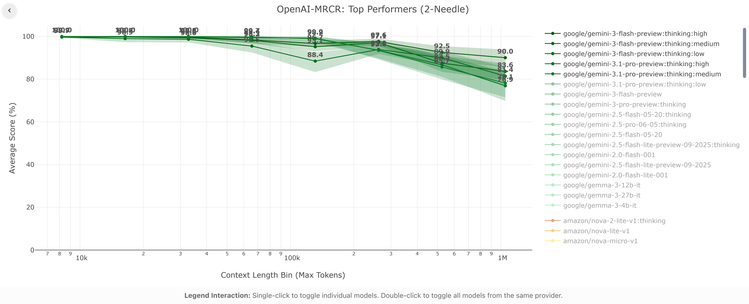
Chart Options

Line

Use Random Colors

Show Chart Labels

Show CI 95%



Legend Interaction: Single-click to toggle individual models. Double-click to toggle all models from the same provider.

Leaderboard | Controls | FAQ

Explore technical milestones and the evolution of long-context retrieval performance.

Note: Deprecated models and older revisions are hidden by default. To view all models, go to the Controls tab and adjust the filters.

Model	Max Ctx	128k (%)	1M (%)	AUC @128k (%)	AUC @1M (%)	Model AUC (%)	Tok Eff.	Input Cost ($)	Output
gemini-3-flash-preview:high	1,048,576	96.7% #1	90.0%	98.3% #1	93.8%	93.8%	1.05x	US$77.6185	US$
gemini-3-flash-preview:medium	1,048,576	95.1% #2	83.6%	97.8% #2	90.0%	90.0%	1.05x	US$77.6185	US$
gemini-3-flash-preview:low	1,048,576	88.4%	81.4%	94.7% #3	89.0%	89.0%	1.05x	US$77.6185	US$
gemini-3.1-pro-preview:high	1,048,576	98.9%	76.9% #2	99.6%	87.8%	87.8%	1.05x	US$620.9480	US$1
gemini-3.1-pro-preview:medium	1,048,576 #2	99.2%	78.1% #3	99.5%	87.7%	87.7%	1.05x	US$620.9480	US$1
gemini-3.1-pro-preview:low	1,048,576	97.4%	79.3%	99.0%	87.4%	87.4%	1.05x	US$620.9480	US$
gemini-3-flash-preview	1,048,576	80.3% #3	82.5%	88.8%	85.5%	85.5%	1.05x	US$77.6185	US
gemini-3-pro-preview OLD	1,048,576 #3	99.0%	72.2%	99.5%	85.5%	85.5%	1.05x	US$620.9480	US$1
gemini-2.5-flash-05-20 OLD	1,048,576	93.5%	68.1%	91.5%	78.3%	78.3%	1.05x	US$46.5711	WARN US
gemini-2.5-pro-06-05 OLD	1,048,576	84.4%	63.9%	89.6%	77.5%	77.5%	1.05x	US$388.0925	WARN US$
gemini-2.5-flash-05-20 OLD	1,048,576	82.9%	64.9%	81.7%	70.2%	70.2%	1.05x	US$46.5711	WARN US
nova-2-lite-v1	1,000,000	92.6%	59.4%	92.9%	63.5%	63.5% #2	0.99x	US$44.1462	US
gpt-4.1 OLD	1,047,576	55.5%	42.1%	61.6%	53.2%	53.2%	1.02x	US$300.6806	US$
gpt-4.1-mini OLD	1,047,576	47.2%	32.5%	49.4%	43.6%	43.6%	1.02x	US$60.1361	US
grok-4.1-fast OLD	2,000,000	49.0%	32.0%	57.4%	40.3%	40.3%	1.01x	US$59.5186	US
llama-4-maverick OLD	1,048,576	50.3%	30.1%	52.7%	39.9%	39.9%	1.03x	US$22.7873	US
grok-4.1-fast OLD	2,000,000	54.5%	22.6%	78.3%	36.8%	36.8%	1.01x	US$59.6695	US
gemini-2.5-flash-lite-preview-09-2025 OLD	1,048,576	39.6%	23.8%	47.9%	34.3%	34.3%	1.05x	US$15.5237	US
gemini-2.0-flash-001 OLD	1,048,576	60.1%	13.3%	56.0%	32.1%	32.1%	1.04x	US$15.3839	US
gemini-2.5-flash-lite-preview-09-2025 OLD	1,048,576	34.0%	23.8%	40.5%	31.7%	31.7%	1.05x	US$15.5237 #3	US
minimax-01 OLD	1,000,192	71.9%	N/A	71.6%	31.4%	63.2%	1.07x	US$31.0056	US
minimax-m1 OLD	1,000,000	68.5%	N/A	75.4%	30.4%	61.3%	1.06x	US$199.3104	US$
gpt-4.1-nano OLD	1,047,576	36.4%	13.0%	42.6%	24.6%	24.6%	1.02x	US$15.0340	US
gpt-5.2:xhigh	400,000 #1	99.5%	N/A	99.8% #1	24.3%	99.5% #1	1.02x	US$264.3818	US$
gemini-2.0-flash-lite-001 OLD	1,048,576	38.6%	17.0%	38.4%	24.2%	24.2%	1.04x	US$11.5379	US
gpt-5.2	400,000	95.8%	N/A	98.4%	23.7% #2	97.3%	1.02x	US$264.3818	US$
gpt-5 OLD	400,000	95.0%	N/A	96.7%	22.9% #3	93.8%	1.02x	US$188.8442	US$
gpt-5.2	400,000	85.2%	N/A	93.8%	21.2%	86.8%	1.02x	US$264.3818	US$
gpt-5-mini OLD	400,000	87.1%	N/A	92.6%	20.6%	84.4%	1.02x	US$37.7688	US$
llama-4-scout OLD	1,310,720	33.1%	13.4%	38.1%	20.5%	20.5%	1.03x	US$37.9911	US
kimi-k2.5	262,144	81.9%	N/A	83.9%	20.0%	81.9%	1.03x	US$34.9740	US$
grok-4-07-09 OLD	256,000	52.8%	N/A	79.6%	15.6%	63.9%	1.00x	US$831.6127	US$
claude-sonnet-4 OLD	1,000,000	52.8%	N/A	71.2%	15.1%	62.0%	1.15x	US$1,016.8112	US$
claude-sonnet-4 OLD	1,000,000	62.3%	N/A	68.5%	15.1%	62.0%	1.15x	US$1,016.2476	US$
o4-mini OLD	200,000	57.4%	N/A	76.0%	15.0%	61.3%	1.02x	US$166.1829	US$
o3 OLD	200,000	55.4%	N/A	72.7%	14.1%	57.9%	1.02x	US$302.1507	US$
seed-1.6	262,144	47.3%	N/A	69.9%	13.2%	54.2%	1.07x	US$78.4265	US$
kimi-k2 OLD	262,144	52.1%	N/A	59.1%	13.0%	53.3%	1.03x	US$76.0289	US$
claude-3-haiku OLD	200,000	58.6%	N/A	52.9%	12.3%	50.5%	1.15x	US$42.3437	US
claude-3.7-sonnet OLD	200,000	51.1%	N/A	55.9%	12.1%	49.6%	1.15x	US$508.1181	US$
claude-3.7-sonnet OLD	200,000	52.9%	N/A	55.5%	11.7%	47.8%	1.15x	US$508.1299	US$
claude-3.5-sonnet OLD	200,000	46.1%	N/A	51.2%	10.9%	44.8%	1.15x	US$1,016.2415	US$
qwen3-next-80b-a3b OLD	262,144	46.2%	N/A	49.2%	10.3%	42.3%	1.04x	US$13.8113	US
claude-3.5-haiku OLD	200,000	46.2%	N/A	50.0%	10.3%	42.0%	1.15x	US$133.1825	US
claude-opus-4.5	200,000	74.0%	N/A	86.5%	10.2%	86.5%	1.15x	US$847.4530	US$1
qwen3-next-80b-a3b OLD	262,144	41.5%	N/A	48.7%	9.7%	39.6%	1.04x	US$23.0133	US$
gpt-5-nano OLD	400,000	38.3%	N/A	44.2%	9.6%	39.4%	1.02x	US$7.5538	US
claude-sonnet-4.5 OLD	1,000,000	55.0%	N/A	76.6%	9.0%	76.6%	1.15x	US$1,016.9436	US$1
ministral-14b-2512	262,144	37.9%	N/A	42.1%	9.0%	36.9%	1.08x	US$31.7339	US
seed-1.6-flash	262,144	41.9%	N/A	47.0%	8.8%	36.0%	1.06x	US$15.6704	US
o3-mini OLD	200,000	34.7%	N/A	43.8%	8.7%	35.8%	1.02x	US$165.3700	US$
claude-opus-4:curated UNRANKED OLD	200,000	63.2%	N/A	73.1%	8.6%	73.1%	1.15x	US$1,688.0723	US$1
grok-3-beta OLD	131,072	60.0%	N/A	73.0%	8.6%	73.0% #1	0.99x	US$436.6788	US$
glm-4.7	200,000	22.1%	N/A	59.3%	8.6%	35.1%	1.02x	US$90.4283	US$
deepseek-v3.2	131,072	54.4%	N/A	70.6%	8.3%	70.6%	1.02x	US$25.0903	US
ministral-8b-2512	262,144	35.0%	N/A	37.5%	7.6%	30.9%	1.08x	US$23.8004	US
claude-opus-4 OLD	200,000	25.9%	N/A	63.2%	7.5%	63.2%	1.15x	US$2,465.6002	US$1
deepseek-r1-0528 OLD	163,840	44.5%	N/A	61.0%	7.2%	61.0%	1.02x	US$47.7723	US
claude-haiku-4.5 OLD	200,000	39.8%	N/A	59.4%	7.0%	59.4%	1.15x	US$169.4906	US$
deepseek-v3.2	131,072	43.2%	N/A	58.9%	7.0%	58.9%	1.02x	US$33.0686	US
deepseek-v3.2-exp OLD	163,840	45.4%	N/A	55.7%	6.6%	55.7%	1.02x	US$30.9996	US
grok-3-mini-beta OLD	131,072	37.0%	N/A	54.8%	6.5%	54.8%	1.00x	US$44.2934	US
qwen-turbo OLD	131,072	14.2%	4.2%	24.5%	6.4%	6.4%	1.04x	US$7.6699	US
grok-3-mini-beta:high OLD	131,072	34.8%	N/A	50.3%	5.9%	50.3% #3	1.00x	US$44.4349	US
minimax-m2.1	204,800	42.8%	N/A	49.2%	5.8%	49.2%	1.01x	US$44.8923	US
mistral-small-3.1-24b-instruct OLD	128,000	48.7%	N/A	47.7%	5.6%	47.7%	1.08x	US$43.7673	US
qwen3-14b OLD	131,072	34.5%	N/A	44.9%	5.3%	44.9%	1.04x	US$20.7283	US
mistral-medium-3 OLD	131,072	33.2%	N/A	43.0%	5.1%	43.0%	1.08x	US$59.9857	US
qwq-32b OLD	131,072	14.3%	N/A	41.7%	4.9%	41.7%	1.04x	US$29.1730	US
deepseek-r1-0120 OLD	64,000	18.0%	N/A	40.6%	4.8%	40.6%	1.03x	US$101.8653	US$
gemma-3-12b-it OLD	131,072	35.0%	N/A	39.9%	4.7%	39.9%	1.07x	US$4.7571	US
minimax-m2 OLD	204,800	29.6%	N/A	39.3%	4.6%	39.3%	1.05x	US$39.4336	US$
qwen3-30b-a3b OLD	131,072	29.7%	N/A	39.1%	4.6%	39.1%	1.03x	US$11.8091	US
llama-3.3-70b-instruct OLD	131,072	17.1%	N/A	37.9%	4.5%	37.9%	1.04x	US$15.5790	US
qwen3-32b OLD	131,072	29.3%	N/A	36.5%	4.3%	36.5%	1.04x	US$9.4689	US
deepseek-chat-v3-0324 OLD	163,840	33.8%	N/A	34.9%	4.1%	34.9%	1.01x	US$27.6889	US
gpt-oss-120b OLD	131,072	20.7%	N/A	30.3%	3.6%	30.3%	1.03x	US$4.6471	US
qwen3-235b-a22b OLD	131,072	14.9%	N/A	29.6%	3.5%	29.6%	1.04x	US$41.4563	US
command-r7b-12-2024 OLD	128,000	17.3%	N/A	28.7%	3.4%	28.7%	1.07x	#3 US$4.5529	US
gemma-3-27b-it OLD	131,072	25.1%	N/A	28.2%	3.3%	28.2%	1.04x	US$9.6677	US
qwen3-8b OLD	131,072	21.3%	N/A	27.4%	3.2%	27.4%	1.04x	US$10.6603	US
gpt-oss-20b OLD	131,072	15.3%	N/A	25.7%	3.0%	25.7%	1.03x	#1 US$3.5777	US
mistral-large-2411 OLD	131,072	0.0%	N/A	24.1%	2.8%	24.1%	1.20x	INC US$304.2297	INC US
nova-lite-v1 OLD	300,000	6.0%	N/A	17.9%	2.7%	11.1%	1.40x	US$13.5991	US
qwen3-4b:free OLD	40,960	11.7%	N/A	18.0%	2.1%	18.0%	1.04x FREE	US$0.0000 FREE	US
gemma-3-4b-it OLD	131,072	8.1%	N/A	14.5%	1.7%	14.5%	1.07x	US$4.8146 #1	US
ministral-3b-2512	131,072	6.3%	N/A	12.8%	1.5%	12.8%	1.08x	US$12.5113	US
nova-micro-v1 OLD	128,000	4.0%	N/A	9.7%	1.1%	9.7%	1.29x	#2 US$4.5526	#2 US
nemotron-3-nano-30b-a3b	262,144	0.0%	N/A	8.3%	1.0%	4.0%	1.10x	US$8.1178	US
router UNRANKED OLD	131,072	2.2%	N/A	5.8%	0.7%	5.8%	1.22x	US$120.4235	US
mistral-nemo OLD	131,072	0.0%	N/A	1.5%	0.2%	1.5%	1.08x	INC US$2.6120	INC US

Table Interaction: Click headers to sort. ↑_ (Model header): Reload data. (Model cell): View model performance chart. (Bin header): View Cost/Score plot. (Score cell): View test details.

Notes & Definitions:

1. N/A indicates no results for this model/bin, potentially due to context window limits.
2. AUC (Area Under Curve) normalized to 100% reflects overall performance across bins, weighted by bin width. Model AUC is normalized only for context bins within the model's max context length.
3. Badge definitions:
   - #1 #2 #3 : Top 3 models for that metric (higher score / lower cost is better).
   - UNRANKED : Unranked due to known issues.
   - WARN : Cost inaccuracy reported.
   - INC : Incomplete cost data (potentially underestimated cost, excluded from cost rank).
   - FREE : Free model or cost data unavailable (excluded from cost rank).
   - OLD : Old model (released >1yr ago).