Beyond LMArena: 3 Practical Insights for Custom LLM Benchmarks

Prashant Anand

2025.10.07

Product category and attributes

Shopify

https://shopify.engineering/evolution-product-classification

On-call Q&A Slack bot

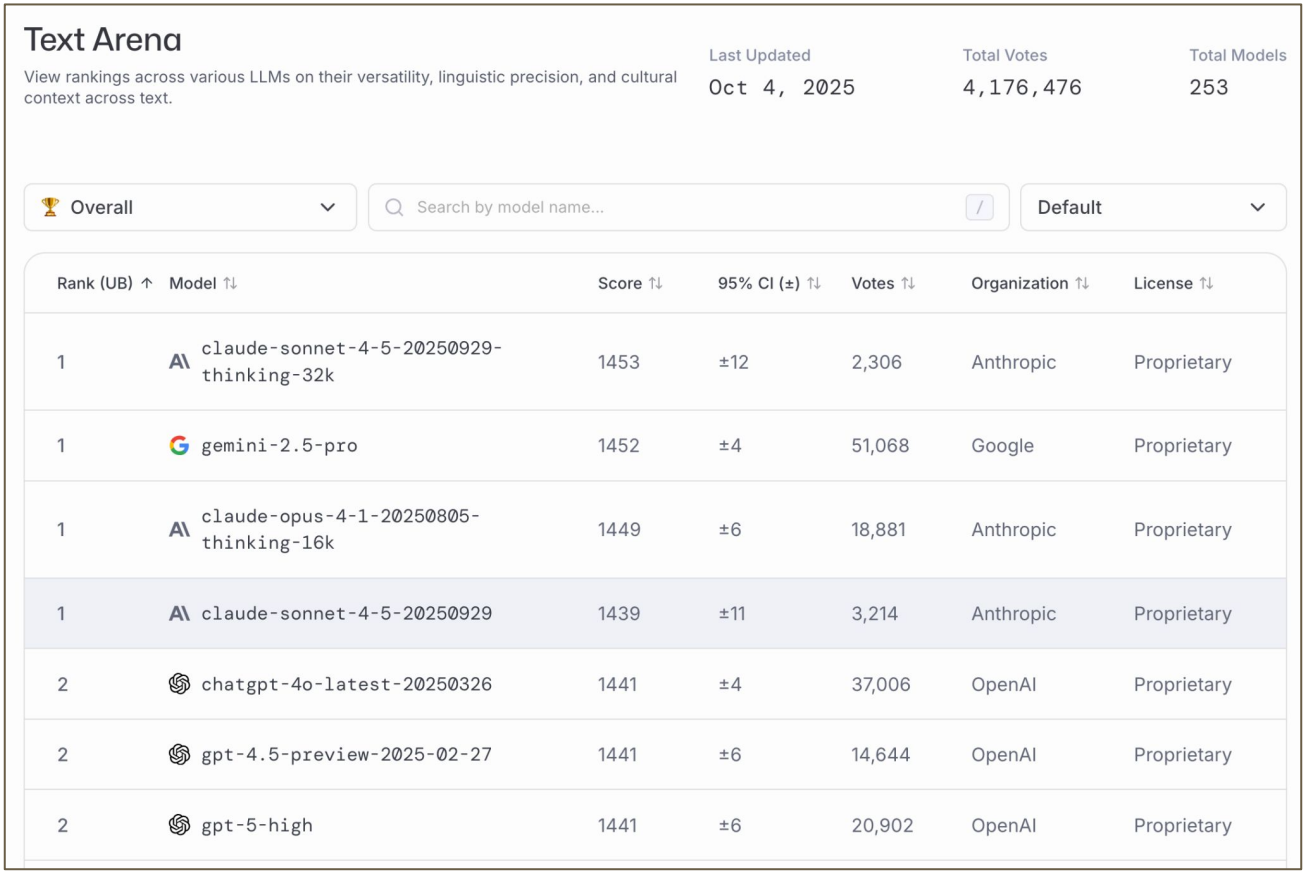
Uber

https://www.uber.com/en-JP/blog/enhanced-agentic-rag/

"prefers spicy Sichuan dishes, avoids dairy"

DoorDash

https://careersatdoordash.com/blog/doordash-profile-generation-llms-understanding-consumers-merchants-and-items/



Text Arena

View rankings across various LLMs on their versatility, linguistic precision, and cultural context across text.

Last Updated: Oct 4, 2025

Total Votes: 4,176,476

Total Models: 253

Overall | Search by model name... / | Default

Rank (UB) ↑	Model	Score	95% CI (±)	Votes	Organization	License
1	claude-sonnet-4-5-20250929-thinking-32k	1453	±12	2,306	Anthropic	Proprietary
1	gemini-2.5-pro	1452	±4	51,068	Google	Proprietary
1	claude-opus-4-1-20250805-thinking-16k	1449	±6	18,881	Anthropic	Proprietary
1	claude-sonnet-4-5-20250929	1439	±11	3,214	Anthropic	Proprietary
2	chatgpt-4o-latest-20250326	1441	±4	37,006	OpenAI	Proprietary
2	gpt-4.5-preview-2025-02-27	1441	±6	14,644	OpenAI	Proprietary
2	gpt-5-high	1441	±6	20,902	OpenAI	Proprietary

https://lmarena.ai/leaderboard/text

Business Constraints

- Performance
- Cost
- Latency

Solves the business problem under the constraints

Over

Scores gold medal in Mathematical Olympiad

Kagi Offline Benchmark

model	%accuracy	Cost($)	time/ task	tokens	TPS	provider
claude-4-opus-thinking	74.3	22.4	13.3	17058	11.0	kagi (ult)
grok-4	73.6	1.0	65.1	3660	0.5	kagi (ult)
claude-4-sonnet-thinking	73.0	5.4	14.1	17872	10.0	kagi (ult)
gpt-5	72.7	7.1	32.8	6282	1.6	kagi (ult)
qwen3-max	72.5	1.9	15.7	148347	22.1	openrouter
o3-pro	72.1	34.2	87.8	12054	1.1	kagi (ult)
gemini-2-5-pro	70.3	1.7	20.9	13581	5.4	kagi (ult)
gpt-5-mini	70.3	4.9	26.7	10113	3.3	kagi
gpt-5-codex	70.3	4.6	20.7	14993	94.7	kagi
deepseek-r1	69.4	9.9	33.6	40707	9.8	kagi (ult)
qwen-3-235b-a22b-thinking	69.4	0.1	27.6	52601	15.8	kagi

https://help.kagi.com/kagi/ai/llm-benchmark.html

Prashant Anand

- ML Engineer at Mercari
- GitHub: @primaprashant
- Love specialty coffee

01 Start Small

- Benchmark with only 10 tasks is better than having none

- Methodological, early signal, and smoke tests

02 Keep It Simple

- Direct LLM API calls over frameworks

03 Pairwise Tests

	gemini-2.5-flash-lite **Correct**	gemini-2.5-flash-lite **Wrong**
gpt-5 **Correct**	60 (a)	**14 (b)**
gpt-5 **Wrong**	**1 (c)**	25 (d)

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

03 Pairwise Tests

	gemini-2.5-flash-lite **Correct**	gemini-2.5-flash-lite **Wrong**
gpt-5 **Correct**	60 (a)	**8 (b)**
gpt-5 **Wrong**	**7 (c)**	25 (d)

03 Pairwise Tests

	gemini-2.5-flash-lite **Correct**	gemini-2.5-flash-lite **Wrong**
gpt-5 **Correct**	60	**18**
gpt-5 **Wrong**	**17**	5

Conclusion

- Having a benchmark with 10 tasks is better than having none.
- Prefer direct LLM API calls over abstract frameworks.
- Use pairwise tests and analyze failure model to come up with portfolio of models.

https://docs.google.com/presentation/d/10c8wukqtkaPiQ5TqeiKAJ64Loq0yCjyYIx0A-xFHM50/edit?usp=sharing