

Web search API benchmark results: Keenable vs Exa, Parallel, Tavily, Brave and 18 more

By Andrey Styskin, Matthias Petri and Ilya Gusev, Keenable. Updated Oct 22, 2026, 14:25 UTC after every benchmark run.

As of Oct 22, 2026, NEEDLE compares 23 web search APIs on the same queries in five verticals. Keenable leads overall with 73% of the ceiling (Exa 68%, Perplexity 67%, Brave LLM Context and Perplexity Fast 66%), and leads Scholar (83%), AgenticRare (50%; Exa 48%, intervals overlap) and Legal (82%; Exa 79%, intervals overlap). Brave LLM Context leads News (64%; DeepSeek 63%, intervals overlap), where Keenable is fifth (59%); Exa leads Finance (96%; Exa Instant 94%, intervals overlap), where Keenable ties for third (93%). Keenable has the lowest median latency, 142 ms, at \$1.00 per 1,000 queries.

Leaders by vertical

Best score in each vertical and Keenable's place, last 7 days

Vertical	Leader	Score	Keenable
Overall	Keenable	73%	1st
Scholar	Keenable	83%	1st
News	brave-llmcontext	64%	5th, 59%
Finance	exa	96%	3rd, 93% (tied with perplexity and parallel)
AgenticRare	Keenable	50%	1st (exa 48%, intervals overlap)
Legal	Keenable	82%	1st (exa 79%, intervals overlap)

Results for all 23 engines

NEEDLE web search API benchmark, last 7 days to Oct 22, 2026, 14:25 UTC: share of the ceiling by vertical, price, latency and result uniqueness for 23 engines. The best value in each column is in bold.

engine	overall	overall 95% CI	Scholar	News	Finance	AgenticRare	Legal	\$ / 1k queries	p50 latency	p95 latency	unique URLs
Keenable	73%	[72%, 75%]	83%	59%	93%	50%	82%	\$1.00	142 ms	342 ms	9%
exa	68%	[67%, 70%]	62%	56%	96%	48%	79%	\$7.00	2.9 s	6.2 s	7%
perplexity	67%	[65%, 68%]	69%	63%	93%	45%	62%	\$5.00	292 ms	1.6 s	8%
brave-llmcontext	66%	[65%, 67%]	60%	64%	91%	46%	68%	\$5.00	1.2 s	2.4 s	2%
perplexity-fast	66%	[64%, 67%]	68%	63%	92%	45%	59%	\$1.00	165 ms	291 ms	8%
parallel	62%	[61%, 63%]	65%	45%	93%	34%	72%	\$5.00	3.4 s	5.1 s	3%
exa-instant	60%	[59%, 61%]	46%	46%	94%	43%	71%	\$7.00	499 ms	833 ms	6%
linkup	60%	[59%, 62%]	54%	57%	88%	40%	62%	\$5.00	1.3 s	1.9 s	6%
brave	60%	[58%, 61%]	56%	56%	85%	35%	67%	\$5.00	510 ms	736 ms	1%
you	58%	[57%, 60%]	54%	56%	84%	34%	64%	\$5.00	515 ms	693 ms	1%
deepseek	57%	[55%, 59%]	28%	63%	85%	37%	71%	\$0.75	567 ms	686 ms	12%
context.dev	57%	[55%, 58%]	58%	47%	86%	32%	61%	\$0.50	1.1 s	4.6 s	5%
parallel-turbo	54%	[52%, 55%]	43%	42%	91%	32%	60%	\$1.00	351 ms	613 ms	7%
kagi	53%	[52%, 55%]	62%	32%	81%	28%	62%	\$12.00	1.3 s	2.1 s	1%
firecrawl	52%	[50%, 53%]	55%	28%	88%	25%	63%	\$1.20	766 ms	4.6 s	0%
claude-search	51%	[48%, 55%]	51%	45%	76%	28%	56%	\$10.00	35 s	51 s	3%
octen	51%	[49%, 53%]	33%	50%	82%	31%	58%	\$1.00	206 ms	416 ms	6%
google/serper	49%	[47%, 51%]	54%	26%	82%	24%	59%	\$0.30	1.5 s	3.4 s	1%
bing/searchapi	49%	[47%, 51%]	50%	31%	80%	18%	65%	\$1.00	2.5 s	9.1 s	1%
chatgpt-search	48%	[42%, 53%]	29%	32%	75%	33%	69%	\$10.00	40 s	54 s	6%
jina	45%	[43%, 47%]	52%	21%	81%	16%	55%	\$3.40	17 s	25 s	0%
tavily	44%	[42%, 45%]	34%	30%	89%	27%	38%	\$5.00	2.0 s	3.7 s	2%
tinyfish	40%	[38%, 42%]	14%	26%	83%	21%	56%	\$0.00 (free plan)	1.9 s	3.7 s	1%

Scores are shares of the ceiling, a synthetic engine that pools every engine's results per query, so 100% is the best result list the whole field found together. Prices are the lowest public self-serve price per 1,000 queries, collected 2026-08-21.

Runs behind the 7-day means: Scholar 10 (latest Oct 22, 14:25 UTC), News 158 (latest Oct 22, 14:18 UTC), Finance 7 (latest Oct 21, 04:54 UTC), AgenticRare 6 (latest Oct 22, 09:33 UTC), Legal 8 (latest Oct 21, 21:36 UTC).

Download the data

[results.csv](#) [results.json](#) [results.md](#) [results.pdf](#)

The files are rewritten after every benchmark run.

Methodology

- Every engine gets the same query text in each run, one request at a time, so latency is comparable across engines.
- Queries are made fresh from public sources: News every hour, the other verticals every day. There is no fixed set to overfit.
- An LLM judge that does not see the engine name scores News and AgenticRare. Scholar and Legal are known-item tasks: a hit is the known paper, court opinion or regulation. Finance checks whether the results contain the known answer from public registries and SEC filings, with an LLM check of results that the text match does not accept.
- We score what the engine returned: ranking, titles and snippets, with the same 2,000-character snippet budget for every engine. We do not fetch pages and we do not re-rank.
- Failed searches are left out of the mean and reported as errors. They are not scored as zero.
- Keenable operates NEEDLE, and the judge has not yet had a systematic human-agreement audit. The queries, prompts and raw results are public, so anyone can score any result again.

[Full methodology and appendices](#)

Columns

overall

Mean share of the ceiling across the verticals where the engine has a score, last 7 days, with a 95% confidence interval. A dash marks a vertical with no score.

Scholar, News, Finance, AgenticRare, Legal

Share of the ceiling in each vertical, last 7 days.

\$ / 1k queries

Lowest public self-serve price per 1,000 queries, without enterprise quotes; collected 2026-08-21. The plan behind each price is in the [price table](#).

p50 latency, p95 latency

Search latency of successful calls, last 7 days.

unique URLs

Share of the engine's URLs that no other engine family returned for the same query and that are relevant, last 7 days. Engines of one family, such as Exa and Exa Instant, do not count against each other.

How to cite

Styskin, A., Petri, M., & Gusev, I. (2026). NEEDLE: A live open-source search benchmark for AI agents (results as of October 2026). Keenable.
<https://keenableai.github.io/needle/results/> , <https://doi.org/10.5281/zenodo.22095155>

```
@misc{needle2026,  
  author = {Styskin, Andrey and Petri, Matthias and Gusev, Ilya},  
  title  = {{NEEDLE}: A Live Open-Source Search Benchmark for {AI} Agents},  
  year   = {2026},  
  url    = {https://keenableai.github.io/needle/results/},  
  doi    = {10.5281/zenodo.22095155},  
  note   = {Live benchmark; results as of October 2026}  
}
```

[Interactive dashboard](#), [source code on GitHub](#), [raw results on Hugging Face](#), [DOI](#).