

8 benchmarks

6,852 questions × 7 providers × 8 benchmarks

BEST ACROSS THE BENCHMARKS

Exa · 65.3%

first on 4 of 8 benchmarks · 18.9% to 95.7% across them · +5.5 pp over the last provider

8 benchmarks · 6,852 questions · 47,965 graded rows (93 errors, excluded from means) · 2,702 contested questions · best-to-worst spread +5.5 pp

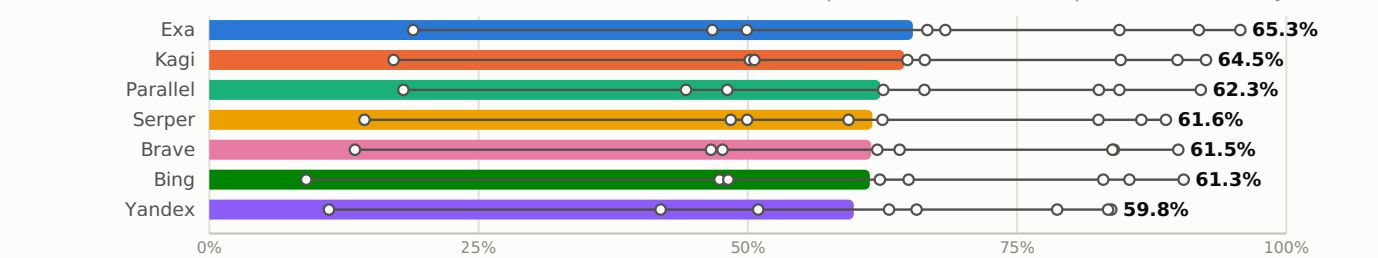
Accuracy by benchmark

One row per benchmark, question count beside it, best provider boxed. The mean row weights every benchmark equally, so a large one cannot outvote a small one.

	Exa	Kagi	Parallel	Serper	Brave	Bing	Yandex
kolas_specific (724)	96	93	83	89	84	90	79
triviroomqa (1,200)	67	66	63	62	64	65	66
msqa (1,064)	47	50	44	48	48	47	42
musique (1,000)	50	51	48	50	47	48	51
frames (824)	68	65	66	59	62	62	63
talcai_searchbench (929)	84	85	84	83	84	83	84
simpleqa_verified (1,000)	92	90	92	87	90	85	83
sealqa_0 (111)	19	17	18	14	14	9	11
mean	65	65	62	62	61	61	60

Ranked

Bar is the mean over benchmarks, each dot one benchmark — a wide scatter is a provider whose result depends on which set you ran



PROVIDER	GRADED	MEAN	RANGE	1ST	AVG RANK	VERDICT MIX	URL RECALL	SEARCH	TOTAL	TOKENS	ERR
Exa	6,839	65.3%	19-96%	4/8	2.1		38%	2.4s	8.6s	9,021	13
Kagi	6,840	64.5%	17-93%	2/8	2.2		41%	2.1s	7.6s	5,812	12
Parallel	6,838	62.3%	18-92%	1/8	3.9		32%	2.2s	8.4s	6,718	14
Serper	6,836	61.6%	14-89%	0/8	4.9		28%	2.2s	8.1s	5,284	16
Brave	6,842	61.5%	14-90%	0/8	4.8		33%	0.7s	26.5s	5,563	10
Bing	6,839	61.3%	9-90%	0/8	5.0		38%	0.6s	25.9s	6,018	13
Yandex	6,837	59.8%	11-84%	1/8	5.0		29%	1.1s	8.0s	6,733	15

correct partial incorrect

 URL recall is the share of a question's reference URLs returned exactly, blank where a provider returns facts rather than pages. Latency and tokens are per question, pooled over every benchmark.

What this says

- Exa has the best mean, first on 4 of 8 benchmarks**
65.3% averaged over the benchmarks it ran, +0.8 pp over Kagi. It is not first everywhere — sealqa_0 is where it slips to 18.9%.
- Bing is the least portable result**
It swings 81.5 pp between benchmarks (9.0% to 90.5%), against 72.6 pp for Yandex. A single benchmark would have told you very different things about these two.
- Indonesian is the hardest language, 49.3 pp behind Czech**
Providers average 41.3% on Indonesian against 90.5% on Czech. Bing holds up best there at 46.9%, 9.9 pp clear of the weakest provider.

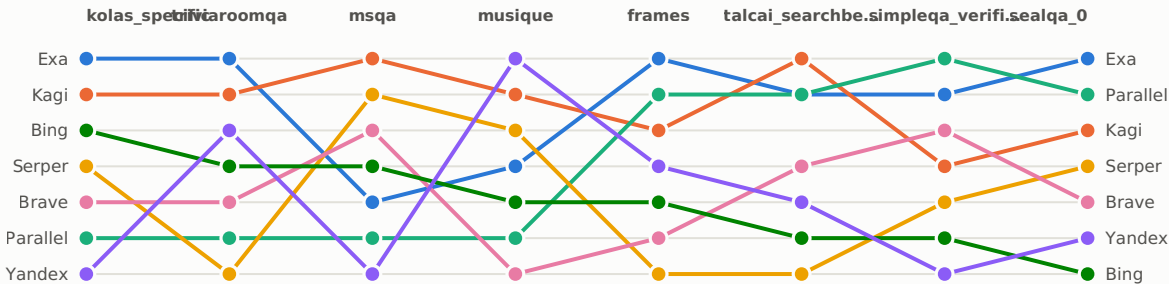
WHERE THEY DIFFER

Does the ranking travel?

Head-to-head figures use the 6,789 questions every provider graded across the 8 benchmarks all of them ran, so no one is credited for a question they never faced.

Rank by benchmark

Position on each benchmark, best at the top. A flat line generalises; a line that crosses belongs to its dataset.



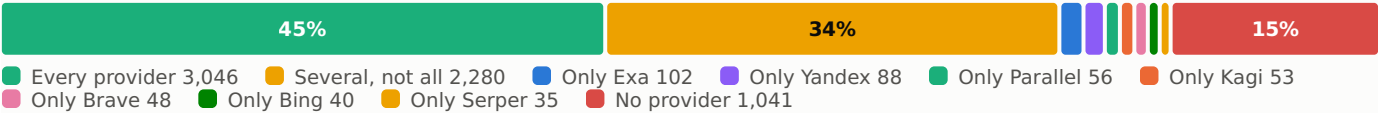
Accuracy by language

One row per language, question count beside it, best provider boxed. Questions carrying no language count as English. Cells pool every benchmark, so a language only some of them cover is not comparable with the rest; under 10 graded questions a cell stays blank.

	Exa	Kagi	Parallel	Serper	Brave	Bing	Yandex	mean of providers
English (4,315)	71	70	69	67	68	67	68	68
French (384)	73	75	70	70	73	75	73	73
Spanish (380)	56	57	52	54	58	56	59	56
German (300)	62	61	57	57	56	59	58	59
Chinese (215)	58	61	52	63	51	50	39	53
Thai (182)	66	70	66	73	68	66	35	63
Korean (161)	81	79	73	77	75	77	71	76
Russian (142)	65	61	58	57	52	59	63	59
Hungarian (111)	95	88	77	83	76	88	79	84
Arabic (99)	96	96	75	94	91	90	89	90
Czech (86)	97	98	85	87	88	94	85	91
Japanese (83)	67	59	48	53	46	53	36	52
Malay (82)	39	50	46	48	49	54	48	48
Indonesian (81)	37	42	41	43	41	47	38	41
Portuguese (80)	42	48	48	48	51	56	41	48
Vietnamese (79)	97	95	84	87	92	91	77	89
Serbian (72)	100	89	79	82	78	90	82	86

Who answers what

Every shared question by how many providers got it right, with the ones won alone split between the providers that own them



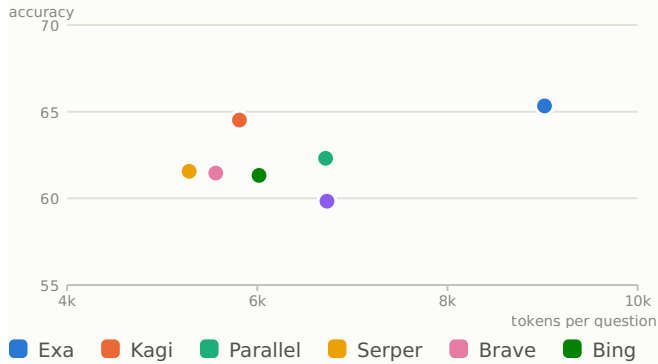
The benchmarks

What each one is worth as a discriminator

BENCHMARK	QUESTIONS	PROVIDERS	MEAN	BEST	LEADER	WHO ANSWERS WHAT
kolas_specific	724	7	88%	96%	Exa	
triviroomqa	1,200	7	65%	67%	Exa	
msqa	1,064	7	47%	50%	Kagi	
musique	1,000	7	49%	51%	Yandex	
frames	824	7	64%	68%	Exa	
talcai_searchbench	929	7	84%	85%	Kagi	
simpleqa_verified	1,000	7	88%	92%	Parallel	
sealqa_0	111	7	15%	19%	Exa	

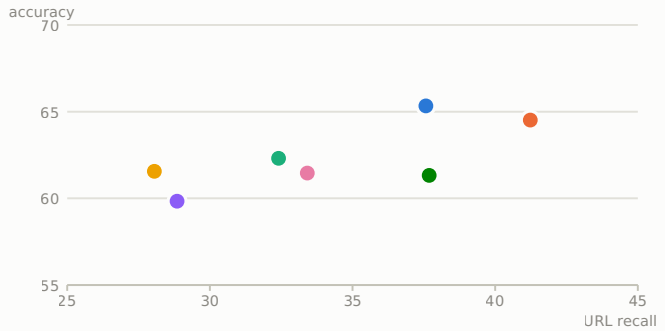
Token efficiency

Mean tokens per question across, accuracy up. Left and high is the same answer for less — the SERP is most of what the model reads.



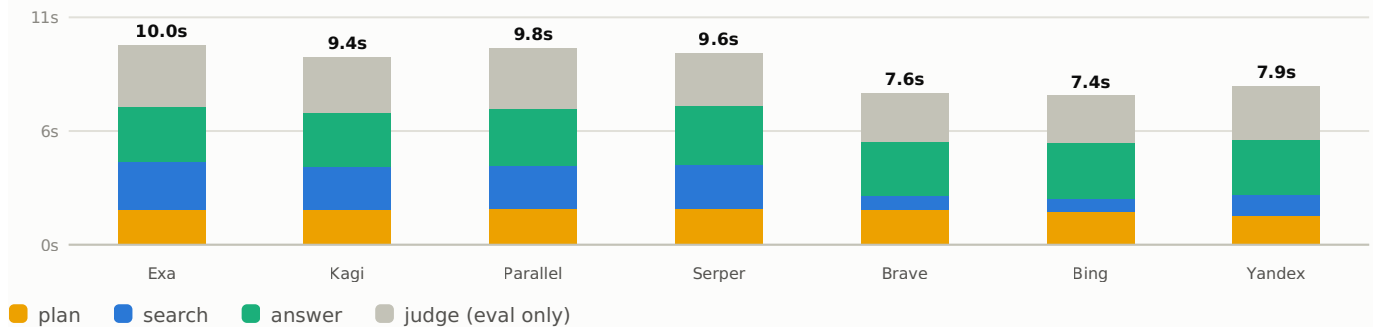
Retrieval against answers

Exact URL recall across, accuracy up, both in percent. Far left and high means answering without the reference page in the results.



Where the time goes

Mean seconds per question, pooled over every benchmark. Only search is the provider's own — planning, answering and judging are the same model for everyone.



Head to head. On the shared set — Exa 70% (102 solo wins, 102 solo misses) · Kagi 69% (53 solo wins, 54 solo misses) · Parallel 67% (56 solo wins, 120 solo misses) · Serper 67% (35 solo wins, 127 solo misses).

A *solo win* is a question only that provider answered; a *solo miss* is one every other provider answered and it did not.

Method. One search call per question, the query planned once per style and shared by every provider asked for it, so all are graded on identical input. An LLM judge scores the answer against the reference; only a full match is correct. Under 50% of a benchmark's questions, a provider is dropped from it.

Page reading off · 260907_1928 260901_1444 260824_1956 260812_2039 260812_2016 260805_2014 260805_1939 260805_1956 · 2026-08-05 → 2026-09-08 · generated 08 Sep 2026, 14:19