Kello Alpha on PeopleSearchBench-Recruiting

Kello · 17 November 2026

Summary. We ran Kello Alpha, as live in production on 17 November 2026, on the 30 recruiting queries of PeopleSearchBench (PSB), a public people-search benchmark made by LessieAI. We ran the full set three times and scored each run with PSB's own judge: Gemini 3 Flash, checking claims on the web with Tavily. Kello Alpha scored **94.7** out of 100, the average of 94.7, 95.1 and 94.3. The highest score on PSB's own leaderboard for these 30 queries is Lessie's 68.2. Every query returned people in every run, and 1,313 of 1,348 returned people were qualified under PSB's rule (meeting at least half the criteria).

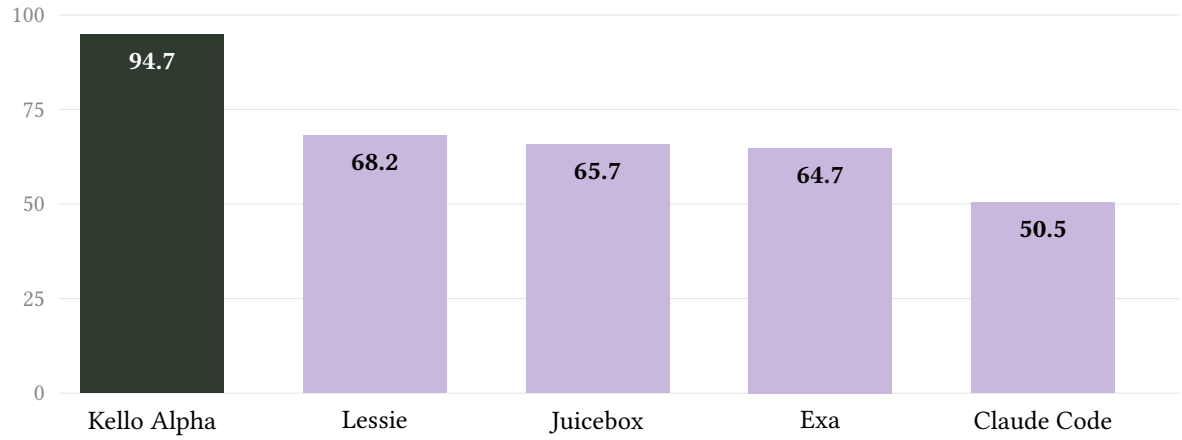


Figure 1: Recruiting score on PSB's 30 recruiting queries. Kello Alpha is the average of our three runs on 17 Nov 2026. The other four are PSB's published scores: the average of the three recruiting numbers in PSB's score files (Lessie, Juicebox, Exa, Claude Code).

1. What we measured

Recruiters describe the person they want in plain words. A people-search tool should return real people who fit. PeopleSearchBench is a public test of this. It has 119 queries in four groups. Kello is a recruiting company, so we ran only the 30 recruiting queries.

The queries are PSB's own. They range from "backend developers in London" to "secretaresses in overheidsinstanties" (Dutch) to "strength and conditioning coaches in portugal". They are in English, Dutch and Portuguese.

2. How PSB scores a result

This follows PSB's methodology (docs/methodology.md). For each query, the system returns up to 15 people. The judge then:

1. Breaks the query into checkable criteria, for example "Person is based in Toronto" or "Person has experience with Go".

2. Checks every returned person against every criterion, using the profile data sent and live web search (Tavily), and marks each one met, partially met, or not met.
3. Gives each person a relevance grade: met = 1, partially met = 0.5, not met = 0, averaged over the criteria. A person matching 2 of 3 criteria scores 0.67.
4. Rates each person's record for information utility.
5. Combines these into three scores.

Score	How PSB computes it
Relevance precision	Padded nDCG@10 over the relevance grades. The ideal list always assumes 10 perfect people exist, so a system that returns fewer than 10 is penalised. Mean across queries × 100.
Effective coverage	A person is qualified if their grade is 0.5 or higher. Per query: qualified ÷ 15, capped at 1. Multiplied by the share of queries with at least one qualified person. × 100.
Information utility	The judge rates each person on profile completeness, evidence that they match the query, and whether the user can act on the result. PSB's code weights these 50 / 15 / 35, averages them per query, then across queries × 100.
Overall	The equal-weight average of the three.

For coverage we divide by 15, the value PSB's own scoring tool uses (tools/compute_scores.py). PSB's README mentions 10; their code uses 15, and their published numbers only reproduce with 15.

3. How we ran it

1. **Ask Kello Alpha.** Each of the 30 queries was sent, word for word, to Kello Alpha in production, 10 at a time. Each query became a search thread under a Kello account.
2. **Take the top 15.** We kept the first 15 people Alpha returned for each query, in Alpha's order.
3. **Write the CSV.** We wrote one row per person in PSB's format: name, title, company, location, LinkedIn URL, and the public profile (work history, education, skills).
4. **Judge.** We ran PSB's own judge on that file, with Tavily web checks on and the differences listed in the table below.
5. **Score.** We used PSB's own scoring functions (benchmark/metrics.py), with the same settings as PSB's scoring tool.
6. **Repeat and average.** We did this three times, from scratch, and report the average.

The people Alpha returned are real people, so, like PSB, we do not publish per-person results, for privacy.

Item	What we used	Same as PSB?
Queries	data/queries/recruiting.jsonl, all 30	Yes
Benchmark code	LessieAI/people-search-bench, commit f1afbb9 (25 Aug 2026)	Yes
Result format	PSB CSV columns, up to 15 people per query	Yes
Judge model	Gemini 3 Flash (gemini-3-flash-preview)	Yes
How the judge is called	Google's API directly. PSB's default goes through OpenRouter.	Same model, different route
Web checks	Tavily on, advanced search, as in PSB's code	Yes
Judge instructions and scoring	PSB's code, unchanged	Yes
Failed web search	If Tavily rejects a search, that person scores 0. PSB's code lets the judge grade them anyway.	Stricter on precision and coverage
Profile data sent	Public profile only: work history, education, skills, location, LinkedIn. No emails or phone numbers.	Stricter; can only lower our score

4. Results

Platform	Relevance precision	Effective coverage	Information utility	Recruiting score
Kello Alpha	**92.3**	**97.3**	**94.5**	**94.7**
Lessie (published)	74.8	75.6	54.3	**68.2**
Juicebox (published)	66.1	75.3	55.8	**65.7**
Exa (published)	66.2	73.8	54.0	**64.7**
Claude Code (published)	59.0	46.7	45.8	**50.5**

Table 1: PSB recruiting, 30 queries. Kello Alpha is the average of our three runs. The other rows are PSB's published numbers (Lessie, Juicebox, Exa, Claude Code); we did not re-run them.

Kello Alpha	Relevance precision	Effective coverage	Information utility	Recruiting score
Run 1	92.3	97.3	94.5	**94.7**
Run 2	92.7	97.6	95.1	**95.1**
Run 3	91.9	96.9	94.0	**94.3**
Average	**92.3**	**97.3**	**94.5**	**94.7**

Table 2: Each run on its own, and the average.

Across all three runs	Value
Queries that returned people	90 / 90
People returned	1,348
People qualified under PSB's rule	1,313
Qualified people per query, on average	14.6
Checks marked met / partly met / not met	4,774 / 228 / 262
People the judge could not grade	21

Table 3: Totals across all three runs. People the judge could not grade, or whose web search Tavily rejected, score 0 for precision and coverage and are left out of the utility average, as in PSB's code.

5. Every query

Table 4: Recruiting score for every query, in each run and on average. Each score is the average of relevance precision, effective coverage and information utility for that query.

ID	Query	Run 1	Run 2	Run 3	Average
rec_0001	Find backend developers in London with experience in microservices architecture	93.4	96.0	94.6	**94.7**
rec_0002	Looking for product managers in Seattle who have experience in developer tools and platforms	99.6	99.8	99.3	**99.6**
rec_0003	Find UX designers in Berlin with 2+ years experience who specialize in mobile app design	99.8	99.4	90.1	**96.5**
rec_0004	Looking for machine learning engineers in Boston who have worked on large language models	92.3	98.8	99.7	**96.9**
rec_0005	Find senior data scientists with 5+ years experience in New York City, working in fintech companies, and holding a PhD degree	96.8	99.3	95.3	**97.2**
rec_0006	i am looking for a new head of fundraising to help me with fundraising for my new tech start up. I am looking for VC people who are based in Singapore to join me. They should at least have 5 years of VC investment team experience and they should be based in singapore. However, they should not be over 40 years of age.	87.4	97.1	92.8	**92.4**
rec_0007	Find senior executives who have successfully sold their previous companies for 100M+ and have a track record of building distributed teams across multiple time zones	94.1	89.6	92.3	**92.0**
rec_0008	Find me talents based in Hangzhou who's role is AI Agent Engineer and works in Alibaba or ByteDance	87.2	97.8	98.4	**94.5**
rec_0009	Find me top AI developers in Bay Area, and graduated from Tsinghua University after 2010	99.4	98.9	97.3	**98.5**
rec_0010	Product managers in Paris who have worked on consumer mobile apps and previously worked at Series A+ companies	99.2	93.1	98.4	**96.9**
rec_0011	Find sales directors in Chicago with 5+ years experience in B2B software sales	98.3	88.8	98.9	**95.3**
rec_0012	Find senior backend engineers in Toronto experienced with Go or Rust, working in startups, 5+ years	89.8	83.4	91.8	**88.3**
rec_0013	Find tax advisors in Europe trusted by startup founders who regularly share practical compliance advice on LinkedIn	95.9	95.7	91.2	**94.3**

ID	Query	Run 1	Run 2	Run 3	Average
rec_0014	Find product designers in Amsterdam who have experience in B2B SaaS design	99.8	99.6	99.7	**99.7**
rec_0015	Looking for growth marketers in Amsterdam who have experience with B2B SaaS products and 3+ years in the field	99.3	99.1	87.0	**95.1**
rec_0016	Find agricultural scientists in Africa working on food security, crop science, or sustainable farming	99.2	96.7	97.6	**97.8**
rec_0017	find top booking agents for electronic music in France, that book artists for clubs such as Raspoutine Paris, Buddah Bar Paris, Casa Amor, Baoli Cannes, Jimmyz, Isabelle sent Tropez, fvtvr Paris, etc	72.3	85.7	78.9	**79.0**
rec_0018	Zoek secretaresses in overheidsinstanties die waarschijnlijk regelmatig moeten notuleren.	99.4	99.2	96.6	**98.4**
rec_0019	Find full-stack developers in Austin with experience in React, Node.js, and Python	99.1	99.5	93.7	**97.4**
rec_0020	i am looking for a new head of fundraising to help me with fundraising for my new tech start up. I am looking for VC people who are based in Singapore to join me. They should at least have 5 years of VC investment team experience and they should be based in singapore..	92.8	87.9	86.3	**89.0**
rec_0021	profissionais liberais ativos no instagram	94.7	97.4	94.8	**95.6**
rec_0022	Find senior executives who have successfully sold their previous companies for 100M+ and have a track record of building distributed teams across multiple time zones	91.8	87.3	75.7	**85.0**
rec_0023	Find product marketing managers in San Francisco with experience in go-to-market strategies	97.3	99.5	99.8	**98.9**
rec_0024	Especialistas de ia	99.6	99.9	99.9	**99.8**
rec_0025	Looking for senior frontend engineers in Boston with experience in performance optimization	85.6	90.5	96.0	**90.7**
rec_0026	Encontra optometristas em Lisboa	93.9	95.4	95.7	**95.0**
rec_0027	Looking for senior frontend developers in Stockholm who specialize in React and TypeScript	97.9	97.6	97.8	**97.8**
rec_0028	encontra os melhores analistas de jogo e scouting no futebol mundial	88.2	87.4	93.2	**89.6**
rec_0029	Find iOS developers in San Francisco with 3+ years experience in Swift and SwiftUI	99.8	94.3	97.1	**97.1**
rec_0030	Bes strength and conditioning coaches in portugal	97.5	99.1	98.7	**98.4**