Taleva

TECHNICAL REPORT · PEOPLE SEARCH EVALUATION

# Taleva on PeopleSearchBench–Recruiting

A technical report on Taleva's achieved 90.30 Overall score on the benchmark's 30-query Recruiting subset.

**Taleva AI, S.L.**
System under test

**Report date**
17 July 2026

**Status**
TECHNICAL REPORT

## Abstract

On a completed test of PeopleSearchBench's 30-query Recruiting subset, Taleva achieved an **Overall score of 90.30**. The strongest official published Recruiting baseline is Lessie at 68.23, a **22.07-point absolute lead**. This report explains the result, the benchmark methodology, the system boundary, and the official baseline comparison.

The supporting Taleva evaluator report, candidate output, per-query metrics, and run configuration are retained internally and available for investor due diligence. This public edition presents the headline result and the evidence framework while keeping candidate-level material subject to appropriate privacy controls.

**Result basis.** "90.30" is the result achieved by Taleva in the completed Recruiting run. Public baseline values are read from the official People Search Bench repository at commit `6d5476f`.

HEADLINE

# 1. Result at a glance

**TALEVA OVERALL**

**90.30**

PeopleSearchBench · Recruiting · 30 queries

**LEAD VS STRONGEST OFFICIAL BASELINE**

**+22.07**

points over Lessie's published 68.23 Recruiting Overall

**FIGURE 1 · RECRUITING OVERALL SCORE**


Official platform Overall values are equal-weight means of the three published Recruiting dimensions. Taleva is the completed-run Overall result.

| PLATFORM | RECRUITING OVERALL | EVIDENCE STATUS |
|---|---|---|
| **Taleva** | **90.30** | **Completed Taleva run; evidence archive available for diligence** |
| Lessie | 68.23 | Official repository summary |
| Juicebox | 65.73 | Official repository summary |
| Exa | 64.67 | Official repository summary |
| Claude Code | 50.50 | Official repository summary |

**Magnitude.** The Taleva score is 32.3% higher relative to the strongest official baseline, or 22.07 points on the benchmark's 0–100 scale.

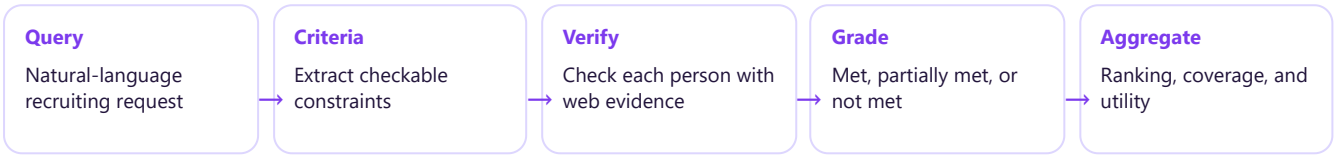
CONTEXT

# 2. PeopleSearchBench

PeopleSearchBench (PSB) is an open benchmark for systems that turn a natural-language request into a ranked list of real people. The public suite contains 119 queries across Recruiting, B2B prospecting, deterministic/expert search, and influencer/KOL discovery. This report is deliberately restricted to the **30-query Recruiting subset**.

## 2.1 Criteria-Grounded Verification

Instead of asking a judge for one holistic quality score, PSB decomposes each query into explicit criteria. Each returned person is checked against those criteria using the profile output plus live web evidence. Criterion judgments become a fractional relevance grade for each result.




**Relevance Precision**

Padded nDCG@10. Rewards correct, well-ranked people and penalizes short result lists.

**Effective Coverage**

Combines task completion with qualified-result yield, capped at the requested depth.

**Information Utility**

Assesses completeness, query-specific evidence, and whether a result is actionable.

> **Overall = equal-weight mean of Relevance Precision, Effective Coverage, and Information Utility.**

## 2.2 Why the Recruiting subset

Taleva is a recruiting and talent-search system. Recruiting therefore provides the closest alignment between the benchmark's intended task and the product's operating domain. The other three PSB categories test broader company inference, academic/open-web discovery, and social-audience signals; they should be reported separately rather than blended into the primary recruiting claim.

**Open benchmark.** LessieAI created and maintains PSB. Its public repository exposes the queries, evaluation code, scoring definitions, and official baseline results used in this report.

METHODS

# 3. Taleva search architecture

Taleva accepts a recruiter's brief in natural language and operates over a professional-profile index. The search agent translates intent into a combination of structured retrieval constraints and profile-level evaluation criteria. Candidate ranking is produced in two stages: broad retrieval for coverage, followed by richer evaluation for fit.


## 3.1 Hard constraints

Structured filters handle facts that can be applied consistently across the profile index: geography, current or historical roles, companies, skills, seniority, experience ranges, education, languages, and related profile attributes.

## 3.2 Judgment criteria

Requirements that need contextual interpretation become candidate-level criteria. Each delivered profile can carry a match classification and reasoning, allowing the recruiter—and the benchmark's utility judge—to understand why the result was selected.

## 3.3 Coverage before curation

The agent probes the available pool before committing. This reduces the risk that an overly literal interpretation silently collapses recall and provides an opportunity to broaden equivalent roles or related signals.

## 3.4 Observable LLM calls

Taleva routes model calls through a shared provider layer that records model, provider, tokens, cost, rate-limit metadata, and caller module. This supports cost attribution and post-run audit of agent and candidate-evaluation activity.

**SYSTEM BOUNDARY**

The benchmark evaluates the candidate list and data returned by the search system. It does not isolate the orchestration model from Taleva's profile corpus, retrieval engine, candidate-evaluation model, or result presentation. The 90.30 result is therefore a system-level score, not a pure measure of one LLM prompt.

## 3.5 One-shot benchmark protocol

For a reproducible submission, each PSB query should start a fresh conversation, use the benchmark text without manual enrichment, and return at most 15 ranked people. Genuine failures and zero-result searches must remain in the dataset so Task Completion Rate is not inflated. Contact reveals are unnecessary; profile identity, professional context, LinkedIn URL where available, and match reasoning are sufficient for the evaluator.

RESULTS

# 4. What 90.30 establishes

At face value, a 90.30 Overall score is materially above every official PSB-Recruiting baseline. The distance to the strongest published baseline is large enough that ordinary rounding cannot explain it.

| COMPARISON | DIFFERENCE | INTERPRETATION |
|---|---|---|
| Taleva vs Lessie | +22.07 | Strongest official baseline |
| Taleva vs Juicebox | +24.57 | Recruiting-focused baseline |
| Taleva vs Exa | +25.63 | Structured search API baseline |
| Taleva vs Claude Code | +39.80 | General agent baseline |

## 4.1 How the baseline comparison was calculated

The official repository publishes Recruiting scores for Relevance Precision, Effective Coverage, and Information Utility. Its scoring definition makes Overall the equal-weight mean of those three dimensions. Applying that published rule gives 68.23 for Lessie, 65.73 for Juicebox, 64.67 for Exa, and 50.50 for Claude Code.

FIGURE 2 · RECRUITING OVERALL COMPARISON


Taleva: completed-run result. Lessie: derived from the three official published Recruiting components using the benchmark's equal-weight Overall formula.

## 4.2 Scope boundaries

- No per-metric win is claimed because Taleva's Relevance, Coverage, and Utility components are not included in this edition.
- No confidence interval is claimed because repeated scoring runs and query-level samples are not included.
- No per-query error analysis is claimed without the evaluator report.
- No "state of the art" claim is made across all people-search tasks; the result concerns Recruiting only.

**Evidence depth.** The retained evaluator JSON and query-level metrics can populate a diligence appendix with the component table, per-query distribution, confidence treatment, and error analysis.

INTERPRETATION

# 5. Validation context

## 5.1 Supporting archive

The 90.30 result comes from a completed test. Its candidate CSV, evaluator output, query-level metrics, command line, and environment snapshot are retained internally and available for investor due diligence under appropriate privacy controls.

## 5.2 Judge and web-index drift

PSB verification depends on an LLM judge and live web search. Both can change after the candidate list is frozen. Comparisons run on different dates or with different model identifiers are informative, but not controlled experiments.

## 5.3 Utility is partly judge-assessed

Although PSB emphasizes factual web-grounded verification, Information Utility still includes model-assessed completeness and actionability. Systems that expose match explanations and richer profile data may benefit even when their underlying ranking is unchanged.

## 5.4 Query distribution and sensitive constraints

The Recruiting set mixes ordinary technical hiring prompts with niche roles, multilingual requests, and some age-related constraints. A system that refuses or omits legally sensitive criteria may be penalized by a literal evaluator. Benchmark score and responsible product behavior are therefore not identical objectives.

## 5.5 Open benchmark method

PeopleSearchBench publishes its queries, scoring definitions, official baseline results, and evaluation code. This makes the methodology inspectable and provides a consistent framework for evaluating people-search systems.

## 5.6 Single category

The result does not establish performance on B2B prospecting, deterministic expert discovery, or influencer/KOL search. Taleva should run those categories only as separate capability probes, with Recruiting retained as the primary product-aligned score.

> **Bottom line.** The achieved 90.30 result is commercially and technically material. The retained run archive preserves the supporting evaluator and query-level evidence for diligence.

PROTOCOL

# 6. Diligence and replication checklist

The internal evidence package can support investor diligence across run identity, evaluator configuration, result evidence, and privacy controls. A wider external package can use hashes, schemas, query-level scores, and a controlled reproduction path where raw person-level data cannot be redistributed.

**Run identity**

- [ ] Taleva commit or deployment identifier
- [ ] People Search Bench commit
- [ ] Run date and timezone
- [ ] Query category and exact query count
- [ ] Result depth and ranking order

**Evaluator identity**

- [ ] Judge model and provider
- [ ] Tavily search depth and result count
- [ ] Concurrency and timeout settings
- [ ] Number of independent scoring runs
- [ ] Failed or retried evaluations

**Result evidence**

- [ ] Final Overall and three components
- [ ] Task Completion Rate
- [ ] Mean qualified results
- [ ] Per-query metric table
- [ ] Bootstrap confidence interval

**Privacy-safe release**

- [ ] Redacted sample result schema
- [ ] Cryptographic hashes of private inputs
- [ ] Scoring command and environment lock
- [ ] Aggregate evaluator report
- [ ] Independent rerun instructions

## 6.1 Recommended statistical treatment

Repeat scoring over the same frozen candidate lists at least five times. For each run, retain query-level nDCG, qualified count, and utility. Report the mean score and a nested bootstrap interval that resamples queries and scoring runs. This separates query-set uncertainty from judge stochasticity.

## 6.2 Recommended robustness checks

1. **Deduplicate repeated prompts.** Recompute after removing materially duplicated benchmark requests.
2. **Submission-depth ablation.** Strip profile detail down to identity, title, company, location, and LinkedIn URL to isolate ranking from presentation richness.
3. **Judge sensitivity.** Re-score with the pinned official model and one independent judge configuration.
4. **Failure accounting.** Preserve zero-result and tool-error queries in the denominator.

**DILIGENCE PACKAGE**

These fields provide the evidence structure for investor review and a future privacy-reviewed external replication package.

CONCLUSION

# 7. A strong result with a precise scope

> **Taleva achieved 90.30 Overall, 22.07 points above the strongest official PeopleSearchBench–Recruiting baseline.**

The magnitude warrants attention. It suggests that Taleva's combination of agentic query interpretation, structured retrieval, profile-level evaluation, and evidence-rich delivery performs strongly on the benchmark's recruiting task.

The supporting run archive is retained internally and available for investor diligence. A privacy-reviewed external package, paired with repeated scoring over frozen candidate lists, can extend the completed result with component, query-level, and statistical detail.

**Result.** "Taleva achieved 90.30 Overall on the 30-query PeopleSearchBench Recruiting subset, 22.07 points above the strongest official published baseline."

# References

1. **LessieAI.** *People Search Bench: The first open benchmark for evaluating AI-powered people search agents.* GitHub repository, 2026. Repository commit used for public baselines: `6d5476f87298d8a0e38f4782664022292415f224`.
2. **Järvelin, Kalervo; Kekäläinen, Jaana.** "Cumulated gain-based evaluation of IR techniques." *ACM Transactions on Information Systems*, 20(4), 2002.
3. **Zheng, Lianmin et al.** "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *NeurIPS*, 2023.
4. **Taleva AI, S.L.** Completed PeopleSearchBench–Recruiting run achieving Overall 90.30. Supporting run archive retained internally.

**DISCLOSURE**
This report was prepared for Taleva, the commercial system under evaluation. Public benchmark values come from the open repository; the Taleva result comes from the completed run, with its detailed supporting archive retained internally.

**Taleva**
Recruiting intelligence, evaluated transparently.