Tavi People Search Benchmark Report

Evaluation on PSB-Recruiting

May 22, 2026

Abstract

This report summarizes Tavi's evaluation on PSB-Recruiting, the 30-query recruiting category of PeopleSearchBench. PeopleSearchBench is introduced in the public paper PeopleSearchBench: A Multi-Dimensional Benchmark for Evaluating AI-Powered People Search Platforms, arXiv:2603.27476. Tavi achieved an overall score of **97.03** across seven evaluator runs, with **97.26** Relevance Precision, **99.90** Effective Coverage, and **93.94** Information Utility. The result measures candidate-search quality: whether Tavi returns relevant people, fills the candidate slate, ranks strong matches highly, and provides useful review information.

Contents

1 Executive Summary

This report evaluates Tavi on **PSB-Recruiting**, the recruiting subset of PeopleSearchBench. The benchmark asks a people-search system to answer realistic recruiting briefs and return ranked candidate lists. Each returned candidate is evaluated against explicit criteria extracted from the query, using web evidence and an LLM judge.

Tavi achieved an **overall score of 97.03** across 30 recruiting queries and seven independent evaluator runs. The result was stable across evaluator seeds, with an overall standard deviation of **0.10**. Tavi also returned nearly a full qualified slate on average, with **14.99 qualified candidates per query** out of a maximum of 15.

97.03	**97.26**	**99.90**	**93.94**
Overall	Relevance	Coverage	Utility

The strongest interpretation is that Tavi's sourcing layer performs very well on structured recruiting search: it finds candidates who match the requested criteria, returns enough qualified candidates to be useful, and provides actionable candidate information. The careful interpretation is that this is an internal benchmark run, not a third-party audited certification.

2 Benchmark Scope

PeopleSearchBench is a public benchmark for evaluating people-search systems. The full benchmark contains 119 queries across four categories:

Category	**Query count**
Recruiting	30
B2B prospecting	32
Deterministic or expert lookup	28
Influencer or KOL search	29
Total	119

This report covers **PSB-Recruiting only**: the 30 recruiting queries. That is the relevant benchmark slice for evaluating candidate sourcing and recruiting search. It is also the subset used in the published recruiting comparisons discussed later in this report.

Each query is a natural-language hiring brief. Tavi is asked to return a ranked list of candidates. The benchmark permits up to 15 candidates per query, and the evaluator focuses heavily on whether the best candidates appear near the top of the list.

3 System Under Test

Tavi was evaluated as a people-search and candidate-sourcing system. For each benchmark brief, Tavi:

1. Interpreted the recruiting intent and constraints in the query.
2. Ran candidate-search strategies to identify people likely to match the brief.
3. Ranked candidates by expected fit.
4. Returned structured candidate information, including profile context and match evidence where available.
5. Exported the results into the PeopleSearchBench submission format.

The evaluation measured search and evidence quality. It did not measure outreach deliverability, candidate response rates, interview conversion, close rates, or end-to-end hiring outcomes.

4 Evaluation Methodology

The evaluation follows a criteria-grounded verification workflow. The benchmark does not just ask, "does this candidate look good?" Instead, it breaks the hiring brief into checkable criteria and evaluates every returned candidate against those criteria.

4.1 Step 1: Criteria Extraction

For each recruiting query, the evaluator extracts explicit requirements from the brief. For example, a query might imply criteria such as:

- The person has a specific role or seniority level.
- The person works in a target industry or company type.
- The person has experience with a required technology, function, or market.
- The person is located in a required geography.

These criteria become the rubric for candidate-level evaluation.

4.2 Step 2: Candidate Verification

Each returned candidate is checked against the criteria. The evaluator uses web search evidence to verify whether the candidate satisfies each requirement. For each criterion, the candidate can be judged as shown in the table below.

The candidate's relevance grade is the average of these criterion-level scores.

Judgment	Meaning	Score
Met	Evidence supports that the candidate satisfies the criterion.	1.0
Partially met	Evidence supports a partial or adjacent match.	0.5
Not met	Evidence does not support the criterion.	0.0

4.3 Step 3: Ranking and Metric Computation

The benchmark then turns candidate judgments into query-level metrics. A system gets more credit when it returns relevant candidates, returns enough of them, ranks the strongest candidates first, and includes information that makes the result actionable.

The same submitted Tavi results were evaluated across seven independent evaluator runs. Repeating the evaluator helps reduce the effect of judge variance and makes the final score more stable.

4.4 Evaluator Stack

The evaluator used:

- The public PeopleSearchBench evaluation workflow and prompts.
- Tavily web search for grounding and evidence collection.
- Gemini 3 Flash Preview as the LLM judge through OpenRouter.
- Seven evaluator runs using the same Tavi candidate submissions.

The evaluator prompt was not edited for this run. The Tavi-side harness was used to run searches and format outputs for the benchmark, but candidate results were not manually edited, benchmark queries were not removed, and the final result includes all seven successful evaluator runs.

5 Metrics Explained

5.1 Relevance Precision

Relevance Precision measures whether the returned people match the query and are ranked well. The core ranking metric is **padded nDCG@10**.

nDCG@10 checks whether the best candidates are near the top of the first 10 results. The top-ranked candidate matters more than the tenth-ranked candidate. A system loses credit if strong candidates are buried lower in the list.

The benchmark uses a padded version, meaning it assumes 10 strong results should be possible. This prevents a system from returning only a few perfect candidates and receiving an artificially high ranking score.

$$\text{nDCG@10} = \frac{\sum_{i=1}^{10} \frac{rel_i}{\log_2(i+1)}}{\sum_{i=1}^{10} \frac{1}{\log_2(i+1)}}$$

Here, rel_i is the evaluator's relevance grade for the candidate in rank position i.

5.2 Effective Coverage

Effective Coverage measures how many qualified candidates the system found. A candidate is treated as qualified if the relevance grade is at least 0.5, meaning the candidate satisfies at least half of the extracted criteria.

This metric rewards systems that can fill the slate, not only systems that find one or two strong matches. In recruiting terms, it asks: did the system return enough plausible people for a recruiter or hiring team to work with?

5.3 Information Utility

Information Utility measures how useful the candidate output is without requiring the user to do all the research again. It considers:

- **Profile completeness**: whether the returned profile contains useful candidate details.
- **Query-specific evidence**: whether the result explains why the candidate matches the brief.
- **Actionability**: whether the user can shortlist, contact, or review the person based on the returned information.

5.4 Overall Score

The overall score is the equal-weight average of Relevance Precision, Effective Coverage, and Information Utility:

$$\text{Overall} = \frac{\text{Relevance} + \text{Coverage} + \text{Utility}}{3}$$

6 Tavi Results

Tavi's final PSB-Recruiting result was:

Metric	Score
Overall	**97.03**
Relevance Precision	97.26
Effective Coverage	99.90
Information Utility	93.94
Task Completion Rate	100.00
Mean Qualified Results	14.99 / 15
Overall Standard Deviation Across Evaluator Runs	0.10

6.1 Per-Run Stability

The seven evaluator runs produced highly consistent scores:

Run	Relevance	Coverage	Utility	Overall
1	97.09	99.78	94.16	97.01
2	97.13	99.78	94.15	97.02
3	97.45	100.00	93.87	97.11
4	96.99	100.00	93.83	96.94
5	97.17	99.78	93.69	96.88
6	97.53	100.00	93.98	97.17
7	97.45	100.00	93.91	97.12

The tight spread suggests the headline score is not the result of one unusually favorable evaluator pass.

7 Comparative Context

The chart below compares Tavi's internal PSB-Recruiting result with publicly available PSB-Recruiting category scores for other systems. These are Recruiting-specific scores, not full 119-query PeopleSearch-Bench overall scores.

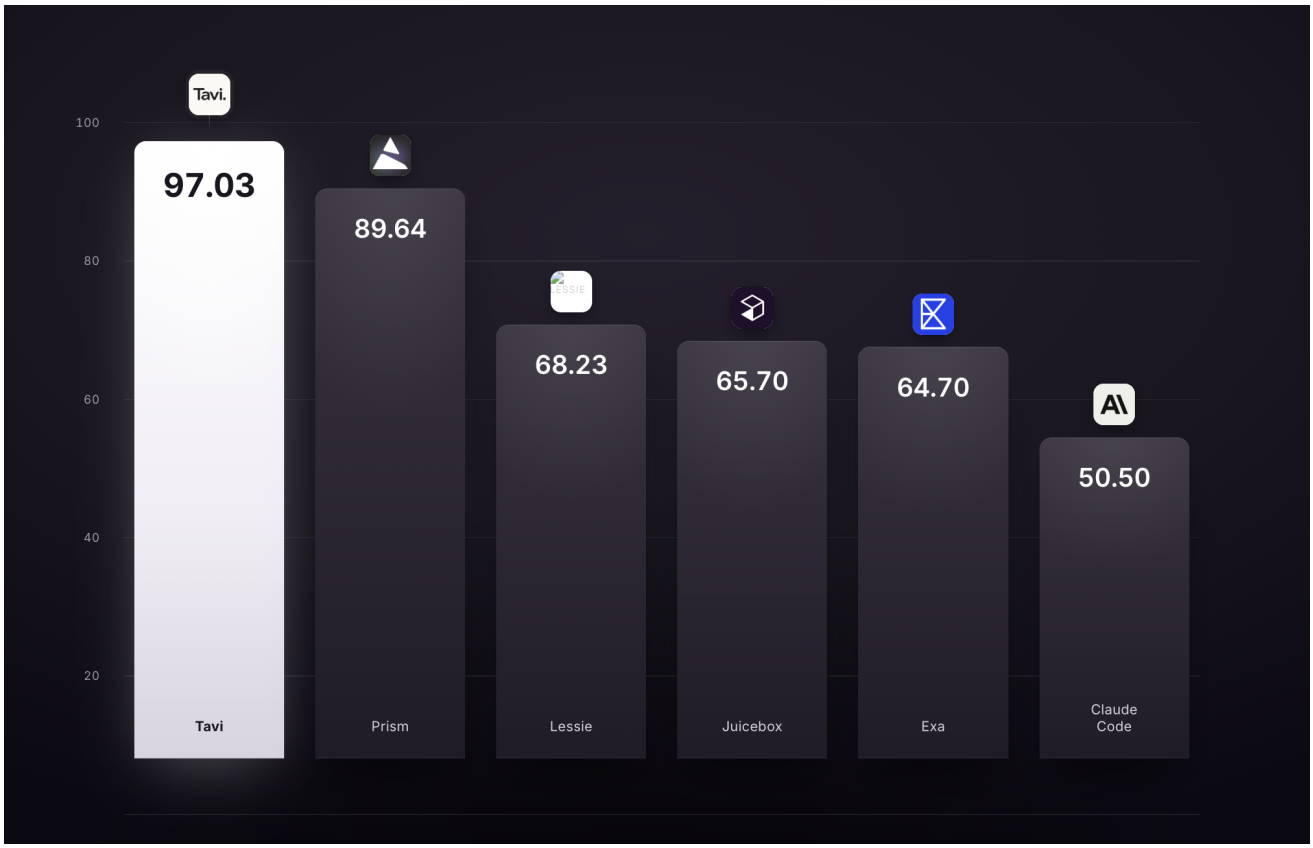

Figure 1: PSB-Recruiting benchmark comparison.

7.1 Scope of the Comparison

The comparison uses PSB-Recruiting category scores. These are category-level scores for the 30 recruiting queries, not full-suite scores across all 119 PeopleSearchBench queries.

For example, Lessie's full 119-query overall score is 65.2, while its Recruiting category score is 68.23. Prism uses the same Recruiting-scope framing because its report is scoped to PSB-Recruiting.

Tavi's internal score is 7.39 points higher than Prism's published PSB-Recruiting score. The competitor scores are published reference scores rather than contemporaneous side-by-side reruns in our environment.

8 Interpretation

The result suggests three things about Tavi's current sourcing layer.

First, Tavi is finding candidates that match the brief. A Relevance Precision score of 97.26 means the returned people generally satisfy the extracted recruiting criteria, and the strongest candidates are ranked near the top.

Second, Tavi is filling the slate. Effective Coverage of 99.90 and 14.99 qualified results per query indicate that Tavi is not only finding isolated strong matches. It is returning close to the full allowed candidate set with qualified profiles.

Third, Tavi's outputs are actionable. Information Utility of 93.94 indicates that the returned information is generally useful for review, shortlisting, and next-step decision-making.

For a recruiting workflow, that combination matters. High precision without coverage creates too few options. High coverage without precision creates review burden. High precision and high coverage without candidate context still leaves the user doing manual research. The benchmark rewards all three.

9 Conclusion

Tavi achieved an overall score of **97.03** on PSB-Recruiting across 30 recruiting queries and seven evaluator runs. The result reflects high relevance, near-complete coverage, and strong information utility: Tavi found candidates that matched the briefs, returned nearly full qualified slates, and included enough context for review and shortlisting.

The clearest readout is that Tavi's people-search layer is performing at a benchmark-leading level on structured recruiting search. In this internal PSB-Recruiting evaluation using the public benchmark workflow, Tavi outperformed the published PSB-Recruiting category scores available for other systems while maintaining a stable score across repeated evaluator runs.