Candidate Matching & Recommendation Technology Fact Sheet

May 20, 2025

Table of Contents

1. Overview

Fastr.ai provides an AI-powered candidate matching and recommendations platform purpose-built for enterprise-scale talent acquisition. The system serves recruiters, hiring managers, and talent acquisition teams by identifying the strongest candidates for any given role across databases of tens of millions of candidates, delivering top results in seconds.

At its core, Fastr.ai employs a proprietary Transformer-based dense retrieval architecture that encodes full resume documents and job descriptions into high-dimensional semantic representations. This approach understands candidates holistically, capturing the meaning, context, and transferability of their experience, rather than relying on keyword overlap or categorical matching. A software engineer with extensive distributed systems experience, for example, is correctly identified as a strong match for a cloud infrastructure role even when their resume uses none of the terminology found in the job description.

The platform supports three core matching modalities: candidate-to-job recommendations, job-to-candidate search, and candidate-to-candidate similarity. It operates at two levels of scale: fine-grained recommendation on curated shortlists within an ATS, and broad discovery search across external data sources containing more than 40 million candidate profiles.

Matching results are presented as actionable HOT / WARM / COLD qualification badges, calibrated through a rigorous human-in-the-loop annotation process with professional recruiters. These badges are complemented by structured information filters (skills, certifications, location, experience, languages, and more) and an AI-generated candidate summary that contextualizes a candidate's qualifications against the specific job description, enabling recruiters to make faster, more informed decisions.

Fastr.ai is designed as a decision-support tool. The system surfaces and prioritizes candidates; humans always make the final hiring decision.

2. Model Description

2.1 Inputs

The matching model ingests the following primary inputs for each candidate-job pair:

- Full resume text — the complete, unstructured document content of the candidate's resume, regardless of length or format.
- Full job description text — the complete job requisition including title, responsibilities, qualifications, and any additional context provided by the hiring organization.
- Candidate's most recent job title — extracted from the resume as a discrete signal.
- Job requisition title — the title of the open position.

These four inputs produce four independent embedding vectors, which are ensembled to compute the final similarity score. The architecture intentionally operates on full raw text rather than pre-parsed categories, enabling the model to capture nuanced relationships, cross-domain skill transferability, and contextual relevance that structured-field approaches miss. A candidate's skills are understood in the context of their job titles, industries, and career trajectory, not as isolated keywords.

During development, additional vector sources were evaluated, including separate embeddings for skills sections and experience sections. These did not improve evaluation metrics over full-text encoding, confirming that the dense retrieval architecture extracts richer signal from whole documents than from decomposed fragments.

In addition to semantic matching, the platform provides structured smart filters parsed from candidate resumes, including:

- Hard skills, soft skills, and technical competencies (context-aware, using a proprietary BERT-based disambiguation model)
- Certifications and professional licenses
- Languages
- Years of experience and experience sections with dates
- Education details
- Location and personal information

Structured filters from job descriptions can also be automatically parsed using Fastr.ai's in-house LLM-based extraction technology and pre-applied for recruiter convenience, reducing manual filter configuration.

2.2 Outputs

For each candidate-job pair, the system produces:

- Qualification badge (HOT / WARM / COLD) — a calibrated assessment of candidate-to-job fit. Badge thresholds are derived from the continuous similarity score and are calibrated during training through Fastr.ai's debiased human annotation process, ensuring that badge assignments reflect professional recruiter judgment, not arbitrary statistical cutoffs.
- Continuous similarity score — the underlying dense retrieval similarity, computed as the mean of the full-text similarity and job-title similarity vectors. This score enables fine-grained ranking within each badge tier, so recruiters can prioritize within their HOT candidates, for example.
- AI Summary — an LLM-generated natural-language summary of the candidate's qualifications in the context of the specific job description, highlighting relevant strengths and potential gaps. This helps recruiters assess candidates faster without reading the full resume.
- Structured attribute data — parsed skills, certifications, languages, experience, and other fields available as smart filters for further refinement of the ranked candidate list.

The combination of a calibrated badge, continuous score, contextual summary, and structured filters gives recruiters both a quick-glance signal and the depth to drill into why a candidate was surfaced, at a level of granularity that coarse-grade systems cannot provide.

2.3 How It Works

Dense Retrieval Architecture

Fastr.ai's matching engine is built on a self-authored Transformer-based dense retriever with custom architectural components. The model encodes arbitrarily long documents, whether a one-page summary or a ten-page CV, into a single dense vector suitable for high-speed similarity search. This is achieved through proprietary top layers that compress full document representations regardless of input length, a capability that standard Transformer models with fixed context windows do not natively support.

The architecture is designed to be modular: the backbone can be replaced with any state-of-the-art dense retriever model as the field advances, while the custom top layers and training infrastructure remain intact. This ensures the system benefits from ongoing advances in foundation model research without requiring a full architectural redesign.

Vector Ensembling

The matching architecture is designed around a flexible vector ensembling framework. Because the Transformer-based Encoder operates on raw text, it can be applied independently to any segment of a document — full resume text, job titles, skills sections, specific experience entries, education, or any other resume component. Each segment produces its

own embedding vector, and these vector similarities are ensembled with configurable weights to produce the final matching score. This flexibility is a configuration choice which would require minor modifications, not an architectural rebuild, and would enable criterion-level explainability while retaining the power of full-database dense retrieval.

The current default configuration ensembles four embedding vectors: full resume text, full job description text, candidate job title, and requisition job title. This configuration was selected based on both evaluation metrics and direct recruiter feedback: matching on full text alone produced highly contextual and accurate results, but recruiters preferred to see candidates with closely matching job titles ranked highest. The four-vector default balances deep semantic understanding with practical usability, and can be extended with additional vector sources (e.g., skills sections, experience sections) to support use cases that require more granular scoring breakdowns, though such approach would be more expensive to maintain, and according to all tests does not produce objectively better results.

Compact, Efficient Embeddings

Through a custom contrastive loss function and training scheme tailored for the matching task, Fastr.ai possess the ability to compress embedding vectors to as few as 64 dimensions while retaining 99.9% of original quality. This compression makes the entire search operation extremely efficient in both memory and computation. Searching the top 100 candidates from a database of over 40 million profiles completes in approximately four seconds, enabling real-time interactive search at scales that batch-processing architectures cannot match.

Structured Information Parsing

Alongside the dense retrieval pipeline, Fastr.ai extracts structured attributes from resumes using a multi-layer parsing approach: a proprietary Named Entity Recognition (NER) model, pattern-matching rules, and a BERT-based disambiguation model that makes skill detection context-aware. This means the system distinguishes between “Java” the programming language and “Java” the geographic reference, or identifies that “leadership” in a management context differs from “leadership” in a student club. Parsed attributes are surfaced as smart filters for recruiters to narrow results after semantic matching.

Multi-Directional Matching

Because the system represents both candidates and jobs as dense vectors in a shared semantic space, it naturally supports three matching directions: best candidates for a given job, best jobs for a given candidate, and most similar candidates to a reference candidate. The latter generalizes from the primary training objective and performs well without dedicated training, demonstrating the robustness of the learned representations.

2.4 How It Was Trained

Patent-Pending Continuous Improvement Framework (Data Engine)

Fastr.ai's models are trained and continuously improved through a proprietary human-in-the-loop Data Engine (patent pending). This framework coordinates professional recruiters who review and refine model recommendations on an ongoing basis, producing the highest-quality labeled training data with full control over what the model learns for each job type.

The Data Engine incorporates:

- Active learning strategies that direct annotator effort toward the most informative examples, maximizing training data quality per annotation hour.
- A debiased annotation process that ensures each candidate is evaluated in a structured, objective manner against the job description, minimizing annotator preference and subjective bias.
- Independent statistical bias audits run after each model is trained, conducted on a quarterly schedule and available on demand, providing quantitative assurance that the model's learned representations do not encode demographic disparities.
- Continuous deployment — models are not periodically retrained on a schedule; they are continuously improved as new annotation data flows in from the Data Engine.

Training Methodology

The model is trained using a custom contrastive loss function with a training and pre-training scheme designed specifically for the candidate-job matching task. This margin-based approach learns to separate strong matches from weak ones with high precision, producing embedding spaces where similarity scores are meaningful and stable.

A Critical Finding: Curated Data Over Activity Logs

> **Key Insight:** Automatically gathered user activity data (candidate pipeline progressions, recruiter accept/reject decisions) proved consistently too scarce at the individual-job level and too noisy to teach the model what makes a strong candidate for a given role. Even for clients with hundreds of thousands of activity records, per-job positive and negative candidate instances were insufficient, and models trained on this signal produced extremely unstable scores. Fastr.ai discovered that even small, highly curated candidate samples per job type, annotated through the Data Engine, give the model a clear understanding of job requirements and produce stable, high-quality scores. These curated models then generalize correctly to analogous jobs they have never seen, demonstrating robust zero-shot performance.

This finding underscores why Fastr.ai invests in purpose-built annotation infrastructure rather than relying solely on implicit behavioral signals. The quality of training data matters far more than its volume, and expert human curation produces fundamentally more reliable models than passive data collection.

2.5 Model QA, Testing & Monitoring

Evaluation Metrics

The following metrics are tracked across all model iterations:

Metric	Score
F-Score (Badge Assignment)	0.753
MNDCG@10 (Ranking Quality)	0.898
Precision@10	0.95

Precision@10 consistently exceeds 0.95, meaning that virtually every candidate in the top 10 results is a genuinely qualified match for the role. MNDCG@10 of 0.898 confirms that the ordering within those top results closely aligns with expert recruiter judgment.

Continuous Monitoring via Data Engine

Because models are continuously trained and evaluated through the Data Engine, Fastr.ai does not rely on periodic monitoring for data drift. The ongoing annotation and retraining cycle means that model performance is validated against fresh expert judgment on an ongoing basis. Every model update is tested against held-out evaluation sets before deployment, and metric trends are tracked over time to detect any degradation.

3. Performance & Scale

Fastr.ai is engineered from the ground up for enterprise-scale operation. The platform is used in production across databases containing tens of millions of candidate profiles, with performance characteristics that support real-time, interactive recruiter workflows.

Capability	Specification
Database Scale	>40 million candidate profiles
Search Latency (Top 100 from 40M)	~4 seconds
Embedding Dimensionality	64 dimensions (compressed)
Quality Retention at Compression	99.9% of full-dimension quality
Matching Modalities	Candidate→Job, Job→Candidate, Candidate→Candidate

The combination of 64-dimensional compressed embeddings and optimized vector search infrastructure enables these latency numbers at scale. Recruiters experience near-instant results when searching across the full database, removing the delay traditionally associated with large-scale candidate discovery. This is not batch processing executed overnight; it is real-time, interactive search that integrates directly into recruiter workflows.

4. Limitations

Fastr.ai is transparent about the boundaries of its technology:

- Input quality dependency. As with any AI system, the quality of matching results depends on the completeness and accuracy of input data. Incomplete or poorly formatted resumes may produce less precise matches. The system mitigates this through its ability to extract signal from full raw text rather than requiring structured fields, but garbage-in/garbage-out constraints remain.
- Novel job types. Models trained through the Data Engine generalize well to analogous job types they have not been explicitly trained on, sometimes exhibiting even high zero-shot performance. However, highly novel or emerging role types that have no analog in the training data may require a Data Engine annotation cycle to achieve optimal performance.
- Language coverage. Model performance is strongest on the languages and regions well-represented in training data. Performance on underrepresented languages may be lower.

5. Responsible AI

5.1 Principles & Governance

Fastr.ai is committed to developing AI that is fair, transparent, and accountable. Recruitment technology directly impacts people's livelihoods and career opportunities, and Fastr.ai treats this responsibility as foundational to its product design, not as a compliance add-on.

Key governance principles include: mitigating bias at the annotation level and rigorously auditing trained models through streamlined, automated bias testing that runs quarterly and on demand, maintaining human oversight at every stage of the recommendation process, investing in continuous improvement rather than point-in-time audits, and operating with transparency about what the system can and cannot do.

5.2 Bias Evaluation

Debiased Annotation Process

Fastr.ai's training data is generated through a multi-step annotation process specifically designed to minimize subjective and demographic bias. Rather than simply asking annotators to rate candidates as "good" or "bad," the process implements a structured evaluation framework that ensures each candidate is scored against the job description on the same objective criteria, with annotator personal preference reduced to a minimum. This process is central to the Data Engine and is applied to every annotation cycle.

Automated Bias Testing

Fastr.ai runs automated bias tests quarterly, with the ability to run them on demand at any frequency across the full production system, testing against all customer data to identify any statistically significant differences in badge distribution, exposure rates, and fairness ratios of model recommendations. The testing process is fully streamlined and automated, enabling fast computation even for large sample sizes. These tests are conducted at scale, covering more than 12 million job–candidate pairs per evaluation cycle, providing robust statistical power to detect even small disparities.

Multi-Stage Bias Mitigation

Bias is addressed at multiple stages: during annotation design (debiased process) and after model training through independent automated bias audits that run quarterly and on demand. This approach ensures that bias is both prevented at the data-curation level and rigorously measured in production, providing the ability to identify and correct sources of bias as well as their downstream effects.

5.3 Explainability & Transparency

The system currently provides explainability through three mechanisms: the HOT/WARM/COLD badge (indicating overall fit), the continuous similarity score (enabling within-tier ranking), and the AI-generated candidate summary (a natural-language explanation of the candidate's qualifications in the context of the job). Recruiters can also view parsed structured attributes (skills, certifications, experience) to understand the factual basis of a candidate's profile.

Fastr.ai is actively developing enhanced explainability features, including a reranking layer which is in final stages of development, that will provide criterion-level scoring breakdowns, showing recruiters how much each qualification dimension contributed to the overall match.

Fastr.ai is transparent about model evaluation metrics (published in Section 2.5) and discloses its approach to training, bias testing, and limitations openly in this document.

5.4 Human Oversight

Fastr.ai is architected as a decision-support system, not a decision-making system. The platform ranks and surfaces candidates; recruiters and hiring managers make all selection and progression decisions. No candidate is automatically advanced or rejected based on the model's output.

Human oversight is also embedded in the model improvement process itself: through the Data Engine, professional recruiters continuously review, validate, and correct model recommendations. Unlike systems where human decisions are only used as implicit training signals, Fastr.ai's human-in-the-loop approach ensures that expert recruiter judgment directly and explicitly shapes what the model learns.

For questions or additional technical details, contact your Fastr.ai account representative.