AI Solution Architecture Addendum

Workforce Optimization & Intelligent Talent Matching Platform

This addendum provides additional technical specifications for the multi-model AI architecture, RAG pipeline for skills intelligence, matching algorithms, and continuous learning infrastructure outlined in the main PRD.

Multi-Model LLM Architecture

The platform employs a specialized multi-model approach where different LLMs handle distinct workforce intelligence tasks based on their strengths:

Component	Model	Rationale
Skill Extraction (Résumés)	GPT-4 Turbo	Superior structured extraction; high accuracy on entity recognition from unstructured documents
Semantic Matching Engine	Claude 3.5 Sonnet	Excellent at nuanced comparison; handles complex multi-factor matching logic with reasoning traces
Requirements Parsing	GPT-4 Turbo	Strong at extracting structured data from RFPs, SOWs, and contract documents
Predictive Analytics	XGBoost + LightGBM	Specialized models for time-series forecasting of staffing demand patterns
Embeddings	text-embedding-3-large	High-quality 3072-dimension vectors for skill/experience semantic matching
Continuous Learning	Claude Opus	Synthesizes feedback patterns; identifies improvement opportunities across matching outcomes

Orchestration Framework

LangGraph manages complex multi-step workflows with human-in-the-loop checkpoints for sensitive staffing decisions:

- **Sequential Workflow:** Employee profile ingestion → Skill extraction → Requirement parsing → Multi-factor matching → Recommendation generation
- **Parallel Execution:** Independent skill extraction and requirement parsing branches merge at matching stage
- **State Management:** Redis-backed checkpointing enables long-running bulk staffing workflows with recovery
- **Human Gates:** Mandatory approval nodes for matches below 70% confidence or for critical roles

RAG Pipeline for Skills Intelligence

The RAG architecture grounds all AI matching decisions in enterprise knowledge, ensuring recommendations are traceable to actual employee data and historical outcomes.

Vector Database Configuration

Component	Specification
Vector Store	Pinecone Enterprise (SOC 2 Type II compliant) with pods sized for 10K+ employee profiles
Embedding Model	text-embedding-3-large (3072 dimensions) for maximum semantic precision
Chunking Strategy	256 tokens for structured HRIS data; 512 tokens for unstructured résumés/documents with 15% overlap
Metadata Schema	employee_id, skill_category, proficiency_level, recency_date, source_system, clearance_level
Retrieval Method	Hybrid search (dense vectors + BM25 keyword) with Cohere Rerank for precision in top-10 results

Knowledge Corpus Structure

The RAG system maintains separate vector collections optimized for different data types:

- **Employee Skills Collection:** 10K+ employee profiles with extracted skills, certifications, experience years, and proficiency levels
- **Project History Collection:** 5 years of assignment records with outcome labels (success/struggle) for model training
- **Requirements Collection:** 2,000+ role descriptions and job requirements normalized to enterprise taxonomy
- **Performance Collection:** 15,000+ performance review summaries with structured manager ratings and competency assessments

Retrieval Pipeline Flow

1. **Requirement Embedding:** Project manager inputs role requirements → System generates dense vector representation
2. **Initial Retrieval:** Retrieve top-50 candidates via hybrid search (semantic + keyword)
3. **Hard Constraint Filtering:** Filter by clearance level, location, availability date, required certifications
4. **Reranking:** Apply Cohere Rerank model considering performance history, skill recency, project fit
5. **Explainable Output:** Present top-10 candidates with match scores and factor-by-factor breakdown

Matching Algorithm Architecture

The multi-dimensional matching engine computes fit scores across weighted factors, with weights dynamically adjusted based on role type and historical success patterns.

Scoring Model Design

Factor	Base Weight	Computation Method
Skill Overlap	30%	Cosine similarity between requirement embedding and employee skill profile embedding
Experience Level	20%	Normalized years in relevant roles with decay function for older experience
Performance History	25%	Weighted average of manager ratings with recency bias (recent reviews weighted 2x)
Availability Alignment	15%	Date proximity score: 100% if available, decaying by days until available
Certification Match	10%	Binary match for required certs; partial credit for related certifications

Dynamic Weight Optimization

Weights are adjusted based on role type and ML-learned patterns from historical success:

- **Senior Roles:** Experience weight increased to 35%; Performance weight increased to 30%
- **Compliance-Critical Roles:** Certification weight increased to 25% (hard requirement becomes filter, not factor)
- **Surge Staffing:** Availability weight increased to 30% to prioritize immediately available candidates
- **ML Refinement:** Gradient boosting model trained on historical match outcomes optimizes weights monthly

Continuous Learning Pipeline

The platform improves over time through feedback loops connecting staffing outcomes back to model training.

Feedback Collection

- **Manager Acceptance:** Track accept/reject decisions on AI recommendations with optional rejection reason
- **Project Outcomes:** 30/60/90 day check-ins collecting manager rating (1-5) on assignment success
- **Employee Feedback:** Post-assignment survey on role fit, skill utilization, and satisfaction
- **Implicit Signals:** Early terminations, extensions, performance improvement plans as outcome indicators

Model Retraining Cadence

Component	Frequency	Trigger
Weight Optimization	Weekly	New feedback data exceeds 50 records

Component	Frequency	Trigger
Skill Taxonomy	Quarterly	Governance committee review + AI-detected new skills
Demand Forecasting	Monthly	New pipeline data from project management systems
Full Model Retrain	Quarterly	Acceptance rate drops >10% or new labeled data exceeds 500 records

Bias Detection & Responsible AI

Given the sensitive nature of workforce decisions, the platform implements comprehensive bias monitoring and mitigation strategies.

Fairness Constraints

- **Protected Characteristic Exclusion:** Age, gender, race, religion, disability status explicitly removed from all model inputs
- **Proxy Detection:** Regular audits for features that correlate with protected characteristics (e.g., graduation year as age proxy)
- **Disparate Impact Analysis:** 4/5ths rule monitoring: flag if any group's selection rate is <80% of the highest group
- **External Audit:** Annual third-party fairness audit by independent consultant with EEOC expertise

Explainability Requirements

Every recommendation includes transparent reasoning for regulatory compliance and user trust:

- **Match Score Breakdown:** Factor-by-factor contribution to overall score displayed in UI
- **"Why Not" Explanations:** Employees can request explanation for why they weren't selected for a role
- **Model Cards:** Versioned documentation for each AI component including training data, limitations, and fairness metrics

Infrastructure & Compliance

Requirement	Specification
Cloud Provider	AWS GovCloud (FedRAMP High) for government contract compliance
Encryption	FIPS 140-2 compliant; AES-256 at rest, TLS 1.3 in transit
Matching Latency	< 30 seconds for top-10 candidates; < 10 minutes for 500+ role bulk staffing
Audit Retention	7 years for all staffing decisions; immutable logging for EEOC/OFCCP compliance

Requirement	Specification
Availability	99.9% uptime SLA; multi-region DR with US-only data residency

[End of AI Solution Architecture Addendum]