AI Resume Shortlister – Non-Functional Requirements (NFR)

Version: 1.0
Owner: Recruitment Solutions Team
Date: 23 Aug 2025

---

1) Purpose & Scope

This document specifies the non-functional requirements (NFRs) for the AI Resume Shortlister product, covering performance, scalability, availability, resiliency, security, privacy, compliance, observability, operability, maintainability, portability, cost, accessibility, and responsible-AI constraints. All requirements include measures, targets, and verification methods where applicable.

---

2) References

- Functional Requirements Document (FRD) v1.0
- Solution Architecture (SAD) (forthcoming)
- Data Governance & PII Handling Plan (forthcoming)
- Threat Model & Security Plan (forthcoming)
- Test Strategy & IR Evaluation Plan (forthcoming)

---

3) Global System Constraints & Assumptions

- **Primary Region:** AWS us-east-1 (Prod) with us-east-2 as DR.
- **Stack:** S3, Lambda, ECS on EC2, FastAPI, Python, FAISS/Chroma, DynamoDB or RDS (PostgreSQL), React/Next.js, CloudWatch, Cognito, WAF, KMS.
- **Traffic Profile:** US business hours heavy, bursty around new requisitions.
- **Data Sensitivity:** Contains PII; resumes may include sensitive fields; treat as confidential.
- **Workloads:** Ingestion (batch/event-driven), Online search (JD → Top-20), Reporting (HTML/PDF), Admin (weights/taxonomy).

---

4) Performance & Capacity

4.1 API Latency Targets (SLI/Targets)

Endpoint	SLI Definition	P50	P95	P99	Notes
`POST /search` (JD→Top-20)	End-to-end server time	≤ 1200 ms	≤ 3000 ms	≤ 6000 ms	Index ≤ 5M resumes; hybrid retrieval + re-rank
`POST /jd/parse`	Server time	≤ 400 ms	≤ 1200 ms	≤ 2500 ms	Includes skill extraction
`POST /report` (single)	Server time	≤ 800 ms	≤ 2000 ms	≤ 4000 ms	HTML; PDF can be async batch
`POST /report` (Top-20 batch)	Job wall-clock	—	≤ 3 min	≤ 6 min	Generates ZIP/PDF bundle
`GET /candidate/{id}`	Server time	≤ 200 ms	≤ 600 ms	≤ 1200 ms	Reads cache/DB + S3 head

4.2 Throughput & Concurrency

- **Sustained QPS:** 20 req/s across APIs in normal load; **Burst:** 100 req/s for 10 min.
- **Concurrent Recruiters:** 50 sustained, 200 peak.
- **Ingestion:** ≥ 30,000 resumes/day; spikes to 100,000/day supported with backpressure; **Ingestion SLA:** 95% ingested, parsed, embedded, and indexed within **30 min** of upload under normal load, **2 h** under peak.

4.3 Embeddings & Indexing Capacity

- **Embeddings Throughput:** ≥ 120 documents/min/node (avg resume 2–4 pages) on c6i.xlarge or equivalent; horizontally scalable.
- **Vector Index Size:** Support 5M resumes v1; roadmap 20M.
- **Shard Strategy:** ≤ 2M vectors/shard; target RAM ≤ 75% per node; spill to IVF/Flat hybrid as needed.
- **Index Build/Rebuild:** Initial 5M build ≤ 24 h; partial nightly maintenance windows ≤ 2 h.
- **Snapshot:** Nightly index snapshot to S3; restore ≤ 2 h.

4.4 UI Performance

- **Initial load (SPA):** TTI ≤ 3 s on 4G/standard laptop; core web vitals CLS ≤ 0.1, LCP ≤ 2.5 s.
- **Table interactions:** Sort/filter respond ≤ 150 ms (client-side).

4.5 Verification

- Synthetic and real-user monitoring; load tests (k6/Locust) to validate P50/P95/P99; profiling to ensure GC/IO not bottlenecks; capacity tests quarterly.

---

5) Scalability

- **Horizontal scaling** on ECS services (parser, embedder, retriever, re-ranker).
- **Auto Scaling Policies:** CPU > 60% for 5 min **or** P95 latency > target → scale out; idle < 20% for 30 min → scale in.
- **Storage Scaling:** S3 virtually unlimited; DynamoDB on-demand with autoscaling; RDS with read replicas and storage autoscaling (if chosen).
- **Queueing/Backpressure:** SQS for ingestion pipelines with DLQs; max-inflight caps to protect downstream.

---

6) Availability, Resiliency & DR

- **Service Availability:** 99.9% monthly for public APIs; scheduled maintenance outside 9am–7pm ET.
- **Multi-AZ:** All stateful services deployed multi-AZ; ECS services spread across AZs.
- **Graceful Degradation:** If generator is down, still return Top-20 with basic explanations; if report generator down, allow CSV export.
- **Circuit Breakers & Retries:** Exponential backoff, idempotent endpoints; bulkheads between components.
- **RTO/RPO:** RTO $\leq$ 4 h; RPO $\leq$ 24 h; index snapshots nightly; DB PITR enabled.
- **DR:** Warm standby in us-east-2; infra as code to recreate; cross-region S3 replication of curated data and snapshots.

---

7) Reliability & Data Integrity

- **Idempotency** for ingestion and indexing; dedupe by content hash + fuzzy signatures.
- **Exactly-once semantics** for index updates; versioned embeddings and taxonomy.
- **Consistency:** Read-after-write for metadata; eventual for search results within 5 min.
- **Checksums** stored and validated on S3 objects; periodic consistency audits.

---

8) Security

8.1 Identity & Access

- **AuthN:** Amazon Cognito (OIDC/JWT).
- **AuthZ:** Role-based access control (RBAC) with fine-grained permissions (search, download, admin).
- **Principle of Least Privilege:** IAM scoped to buckets/prefixes, ECS task roles; no long-lived keys.

8.2 Data Protection

- **Encryption at Rest:** S3 SSE-KMS; EBS encrypted; DynamoDB/RDS encryption; KMS customer-managed keys.
- **Encryption in Transit:** TLS 1.2+; HSTS; perfect forward secrecy ciphers.
- **Secrets:** AWS Secrets Manager or SSM Parameter Store; rotation every 90 days.

8.3 Network & Edge Security

- Private subnets for ECS/EC2; NAT for egress; VPC endpoints for S3, DynamoDB.
- WAF on ALB/API Gateway with managed rules + custom bots/regex.
- Security groups deny-all by default; ingress only from ALB.

8.4 Secure SDLC & AppSec

- SAST/DAST/SCA in CI; dependency pinning; container image signing (Sigstore/COSIGN).
- OWASP ASVS & Top 10; file upload validation; PDF/DOCX sanitization.
- **Vulnerability Mgmt:** Critical patch < 7 days; High < 14 days.

8.5 Audit & Logging

- CloudTrail enabled; access logs immutable in S3 with Object Lock (Governance).
- Admin actions (weights/taxonomy) logged with user, timestamp, before/after.

---

9) Privacy & Compliance

- **PII Minimization:** Store only necessary fields; mask PII in UI by default for sharing.
- **Candidate Consent:** Documented basis for processing; opt-out workflows.
- **DSAR:** Fulfill data-subject access/delete within **30 days**; hard delete within **60 days** including derived embeddings.
- **Retention:** Default **2 years** of inactivity; configurable per client.
- **Regulatory:** CCPA/CPRA, GDPR (if applicable), and **EEOC** fairness guidance; U.S. employment practices.
- **Cross-Border:** Keep US candidates in US regions; disclose if otherwise.

---

10) Responsible AI & IR Quality

10.1 Matching Quality Metrics (IR)

- **Precision@20 $\geq$ 0.65**, **Recall@200 $\geq$ 0.75**, **nDCG@20 $\geq$ 0.80** on labeled JD↔resume gold set.
- **Time-to-first-useful-result:** $\leq$ 2 s P95.
- **Explanation Sufficiency Score:** $\geq$ 0.85 (human-rated) — evidence snippets clearly justify top matches.

10.2 Fraud Detection Metrics

- **Precision (High-risk flag):** $\geq$ 0.80; **Recall:** $\geq$ 0.60; **FPR:** $\leq$ 5%.
- All fraud outcomes are **assistive** (re-ranking penalties), not automatic rejections.

10.3 Bias, Fairness & Guardrails

- No use of protected attributes (race, gender, age, religion, etc.) in scoring or features.

- **Bias Audits:** Quarterly disparate-impact checks on model outputs; keep **impact ratio ≥ 0.8** across protected groups where data is available.
- **Explainability:** Provide per-feature contributions and matched evidence.
- **Grounding:** All generated summaries must cite resume provenance; no hallucinated claims.

10.4 Verification

- Gold set curation; blinded human judging; A/B testing with error-budget policies; offline eval + online metrics.

---

11) Observability

- **Logging:** JSON structured logs; request IDs propagated (traceparent).
- **Metrics:** API latency/throughput, error rates, queue depth, index load factor, cache hit rate, embedding TPS, snapshot duration.
- **Tracing:** OpenTelemetry traces across UI→API→services.
- **Dashboards:** Service health, ingestion pipeline, search latency, IR metrics.
- **Alerts:** P95 search latency > 3 s (5 min), error rate > 1% (5 min), queue age > 10 min, node memory > 80%, disk > 80%.

---

12) Operability & SRE

- **SLOs:**
- Availability 99.9% monthly.
- P95 search latency ≤ 3 s.
- Ingestion 95% ≤ 30 min.
- **Error Budgets:** Trigger release freeze if SLOs breached in last 30 days.
- **Incident Mgmt:** Severity definitions, on-call 24×7; MTTA ≤ 10 min, MTTR ≤ 60 min (Sev-1).
- **Runbooks:** Failover, cache warm, index restore, partial degradation procedures.

---

13) Maintainability & Supportability

- **Code Quality:** Linting, type-checking (mypy), unit test coverage ≥ 80%.
- **Docs:** ADRs for key decisions; API docs (OpenAPI) auto-published.
- **Config Mgmt:** All config in env/SSM; no config in code; feature flags supported.
- **Backwards Compatibility:** API v1 is stable for ≥ 12 months; deprecation policy with 90-day notice.

---

14) Deployability & Release Engineering

- **CI/CD:** GitHub Actions → ECR → ECS blue/green with canary 10% for 30 min.
- **Rollbacks:** One-click rollback within 5 min; database migrations are backward-compatible.

- **DORA Targets:** Deployment frequency ≥ weekly; change failure rate ≤ 15%; MTTR ≤ 1 h; lead time ≤ 1 day.

---

15) Compatibility, Accessibility & UX Quality

- **Browser Support:** Latest 2 versions of Chrome, Edge, Firefox; Safari latest.
- **Accessibility:** WCAG 2.1 AA; keyboard navigation, focus states, ARIA labels, color-contrast checks.
- **Localization:** US English; full Unicode support for names; right-to-left layout not required v1.

---

16) Interoperability & Integration

- **File Types:** PDF, DOCX (v1); TXT optional.
- **APIs:** REST (JSON) with pagination; rate limits 60 rpm/user; burst 120 rpm.
- **Webhooks:** Ingestion completion callbacks; report-ready notifications.
- **Future Connectors:** ATS (Greenhouse, Lever), Slack/Email share links.

---

17) Data Quality & Governance

- **Parsing Accuracy:** ≥ 95% field-level accuracy on name/title/dates; ≥ 90% on skills extraction (top-k).
- **Normalization:** Skills mapped to controlled vocabulary; synonyms list maintained; drift monitoring.
- **Data Lineage:** Provenance for each evidence snippet; maintain version tags for resumes and embeddings.
- **Retention & Purge:** As per §9; verifiable deletion across S3, DB, embeddings, and caches.

---

18) Cost & FinOps

- **Tagging:** All resources tagged with `Product`, `Env`, `Owner`, `CostCenter`.
- **Budgets & Alarms:** Monthly budget per env; 80/100/120% alerts.
- **Unit Economics:** Target ≤ **$0.05** per resume ingested; ≤ **$0.01** per JD query at 5M scale.
- **Scaling Policies:** Rightsize instances quarterly; Spot for batch where feasible; S3 lifecycle tiers for aged artifacts.

---

19) Legal, Ethical & Policy

- **EEOC Compliance:** No disparate treatment; exclude protected attributes; audit logs for decisions.
- **Explainability Policy:** Provide reasons for ranking; show missing skills and evidence snippets.
- **Content Policy:** Block profane/unsafe generated text; sanitize uploads; malware scanning of attachments.

---

20) Backup, Recovery & Archival

- **Backups:** Nightly DB snapshots; S3 versioning; index snapshots nightly.
- **Verification:** Weekly restore test in staging.
- **Archival:** Move stale logs/data to S3 Glacier after 90 days; retain audit logs ≥ 1 year (configurable).

---

21) Measurement & Acceptance

For each SLI/SLO, define: **metric name**, **dashboard location**, **alert threshold**, **test method**, **owner**.
- Example: *Search Latency P95* → `api.search.p95_ms`, Dashboard: `Search`, Alert: >3000ms for 5 min, Test: load test k6 profile `jd_mix.json`, Owner: SRE Lead.

---

22) Roadmap for NFR Maturity

- **v1.0 (Go-Live):** Core performance, availability, security, privacy, observability baselines.
- **v1.1:** Bias audit automation; cost anomaly detection; golden-set expansion; chaos engineering drills.
- **v1.2:** Cross-region active/active; autoscaling on custom IR signals; continuous canary load.

---

23) Appendices

- **A. SLO Registry (initial):**
- *SLO-001:* Availability 99.9% monthly.
- *SLO-002:* Search P95 ≤ 3 s.
- *SLO-003:* Ingestion 95% ≤ 30 min.
- *SLO-004:* Fraud FPR ≤ 5%.
- *SLO-005:* DSAR close ≤ 30 days.
- **B. Glossary:** SLI/SLO, DR, RTO/RPO, DSAR, nDCG, PII, ADR, DORA.