AxiomCortex™: Scientific R&D Report

Authors

TeamStation AI Research & Development Division

Institutional Affiliation
TeamStation AI

Correspondence
One Seaport Square, 77 Sleeper St 5830 E 2nd, St Ste 7000 #14687, Boston, MA 02210
lonnie@teamstation.io

Website
TeamStation AI

Abstract

The following report details the Axiom Cortex, a proprietary Cognitive AI-driven engine integrated within the TeamStation AI ecosystem. Axiom Cortex is designed to revolutionize talent evaluation for nearshore IT operations by moving beyond traditional, subjective assessments. This paper outlines the system's core scientific pillars, including neuro-psychometric profiling via a Latent Trait Inference Engine (LTIE) and advanced NLP integration. The report also details the Self-Governing, Phasic Micro-Chunking methodology that drives the engine's precision, its robust bias mitigation strategies, and the specific scientific parameters that ensure its accuracy and ethical validity. The Axiom Cortex provides a scientifically rigorous, bias-mitigated assessment of a candidate's true cognitive and technical capabilities, enabling businesses to leverage global talent effectively and with unparalleled confidence.

Table of Contents

Axiom Cortex: Scientific R&D Report **1**

Authors 1

Abstract 1

Table of Contents 1

1. Introduction: TeamStation AI – The Intelligent Infrastructure for Global IT Operations 2

What TeamStation AI Does: 2

The Problem We Solve: 3

Our Solution: The Axiom Cortex Feature within the TeamStation AI Ecosystem 3

2. The Axiom Cortex: Our Proprietary Engine for Unlocking True Talent 3

2.1. Key Scientific Pillars: 3

Neuro-Psychometric Profiling via Latent Trait Inference Engine (LTIE): 3

Advanced NLP Integration (The Science of Language Analysis): 4

Bias Mitigation (The Cortex Calibration Layer): 5

Behavioral Answer Deconstruction: 5

Scientific Grounding: 5

3. The Axiom Cortex Methodology: Self-Governing, Phasic Micro-Chunking NLP Prompt Engineering 5

3.1. Phasic Micro-Chunking Execution Protocol: 6

3.2. Self-Governing & Self-Learning Aspects: 7

3.3. Token Efficiency and Reduced Backend Dependency: 7

4. Scientific Parameters and Nuances: The Precision of Our Science 8

5. Frequently Asked Questions (FAQs) 9

6. Conclusion: A Paradigm Shift in Talent Evaluation for Global IT Excellence 10

Appendix 11

7.1. Index of Acronyms and Definitions 11

7.2. Related Papers 12

Index 12

1. Introduction: TeamStation AI – The Intelligent Infrastructure for Global IT Operations

At TeamStation AI, we've engineered a comprehensive platform that acts as the intelligent backbone for nearshore IT services operations. We're not just a talent sourcing agency; we've built an end-to-end ecosystem designed to empower businesses with unparalleled access to top-tier IT talent and advanced AI-driven operational capabilities. Our mission is to redefine how companies engage with and leverage global IT expertise, ensuring seamless integration, predictive success, and cultural alignment.

What TeamStation AI Does:

Our platform is a holistic solution that addresses the entire lifecycle of nearshore IT engagement, from talent discovery to project delivery. It provides clients with:

- Vast Talent Network Access: Direct access to a meticulously curated database of over 2.6 million LATAM IT professionals. This network is continuously updated and profiled, offering a deep bench of specialized skills.
- Intelligent Talent Matching & Evaluation: This is where Axiom Cortex plays a pivotal role. Our AI-driven engine goes beyond traditional resume screening to perform deep neuro-psychometric and NLP-based evaluations, ensuring candidates possess not only the required technical skills but also the optimal cognitive and cultural fit.
- AI-Powered Operations Management: The platform intelligently manages IT operations, including resource allocation, project tracking, performance monitoring, and quality assurance, all enhanced by AI.
- Predictive AI for Success: We leverage AI to predict project success rates, talent performance, and potential risks, enabling proactive management and optimized outcomes.
- Advanced AI Prompt Engineering Leadership: Our platform cultivates and deploys AI Prompt Engineering Leads who can accurately and efficiently build solutions across the entire Software Development Lifecycle (SDLC), ensuring high-quality deliverables and seamless cultural alignment with US business practices.

The Problem We Solve:

The traditional approach to sourcing and evaluating IT talent, especially in nearshore markets, is often subjective, prone to biases (particularly against L2 ESL speakers), and fails to accurately measure the deep technical skills and cognitive abilities required for complex software development. This leads to inefficient hiring, mis-hires, and a failure to tap into the full potential of diverse global talent pools. TeamStation AI's integrated platform, with Axiom Cortex at its core, directly tackles these inefficiencies and biases.

Our Solution: The Axiom Cortex Feature within the TeamStation AI Ecosystem

Axiom Cortex is not a standalone tool; it's a critical, integrated component of the TeamStation AI platform. It provides the scientifically rigorous, bias-mitigated evaluation necessary to leverage our vast talent network effectively. By understanding the true cognitive and technical capabilities of each candidate, we ensure that the talent matched to our clients is not only skilled but also the right cognitive and cultural fit.

2. The Axiom Cortex: Our Proprietary Engine for Unlocking True Talent

Axiom Cortex is the sophisticated analytical engine at the heart of the TeamStation AI platform. It goes beyond surface-level assessments to infer and quantify the underlying cognitive architecture, problem-solving methodologies, learning orientation, and cultural-linguistic nuances of each candidate.

2.1. Key Scientific Pillars:

Neuro-Psychometric Profiling via Latent Trait Inference Engine (LTIE):

The Axiom Cortex infers critical candidate traits, which we then quantify and benchmark:

- **Architectural Instinct (AI):** Assesses a candidate's ability to think top-down, design robust systems, and manage high-level trade-offs.
- **Problem-Solving Agility (PSA):** Evaluates how effectively a candidate deconstructs problems, adapts to new constraints, and explores multiple solution paths.
- **Learning Orientation (LO):** Measures intellectual honesty, coachability, and a genuine drive to learn – our proxy for a growth mindset.
- **Collaborative Mindset (CM):** Assesses a candidate's tendency to work in a team context, consider stakeholder impact, and foster shared understanding.

These traits are scored on a 5-point scale, visualized in detailed cognitive fingerprints, and derived from a synthesis of per-question B-Axiom scores, forensic linguistic features, and observed behavioral evidence.

Advanced NLP Integration (The Science of Language Analysis):

We employ a comprehensive suite of Natural Language Processing (NLP) techniques, meticulously integrated into our prompt engineering:

- Phonology & Morphology: Analyzing language for patterns indicative of L1 influence on word formation and pronunciation. This helps us understand potential cognitive load and semantic nuances without penalizing for non-native fluency.
- Syntactic Analysis (Chunking, Parsing, Grammatical Formalisms): Evaluating sentence structure, idea grouping (chunking), and grammatical frameworks to understand thought organization and clarity. This directly informs our B_C (Clarity) and B_L (Cognitive Load) scores.
- Semantic Processing & Lexical Semantics: Critically assessing the candidate's understanding of meaning, core concepts, their relationships (ontology), and the precision of their word choices. We prioritize accurate conveyance of meaning over exact terminology, recognizing valid paraphrasing and conceptual entailment as per the Conceptual Fidelity Protocol.
- Discourse Analysis: Examining the overall coherence, cohesion, logical flow, and argument structure of the candidate's response to understand how they present complete thoughts.
- Linguistic Resources & Statistical/Knowledge-Based Methods: Leveraging internal linguistic knowledge and statistical methods to identify patterns, L1 interference, and unique cognitive styles.
- Paraphrasing, Entailment, and Generation: Actively identifying instances where candidates correctly express concepts using different phrasing or imply understanding, adhering to the Conceptual Fidelity Protocol.

Bias Mitigation (The Cortex Calibration Layer):

This is a non-negotiable, system-critical directive. Our Cortex Calibration Layer acts as a sophisticated filter, applying algorithmic adjustments to raw scores based on detected L1 interference patterns and cultural communication styles. This ensures we evaluate the pure technical and logical signal, not linguistic "noise." Specific calibrations are applied to:

- B_A (Accuracy) - "Politeness Filter": Adjusts hedging based on cultural norms.
- B_P (Procedural Knowledge) - "Collectivist Filter": Modifies the impact of ownershipRatio to account for collectivist communication styles.

- B_L (Cognitive Load) - "Translation Filter" & "Working Memory Uplift": Adjusts B_L based on L2 processing strain and applies uplifts for flags like coherence_with_reduced_detail_L2_flag.
- B_C & CTA_CE - "Cultural Pragmatic Re-interpretation": Adjusts scoring to value culturally nuanced communication styles, moving beyond Western-centric directness.
- B_A & B_M - "Spontaneous Speech Credibility": Applies penalties for unnatural_text_fluency_flag to differentiate genuine understanding from rote memorization.

Behavioral Answer Deconstruction:

For behavioral questions, answers are deconstructed into core conceptual components (Problem Identification, Impact Analysis, Action & Agency, Justification & Influence) rather than judged against rigid frameworks like STAR, respecting cross-cultural communication theory and avoiding bias.

Scientific Grounding:

All analyses are strictly traceable to provided input data and foundational documents, ensuring zero hallucination and absolute adherence to specified algorithms and data. The "No Evidence" clause mandates explicit statements when evidence is insufficient.

3. The Axiom Cortex Methodology: Self-Governing, Phasic Micro-Chunking NLP Prompt Engineering

The operational backbone of Axiom Cortex is its novel approach to executing complex NLP tasks: a Self-Governing, Self-Learning Phasic Micro-Chunking NLP-based Prompt Engineering technique. This methodology is designed for maximum accuracy, token efficiency, and minimal external dependencies, allowing the LLM itself to perform the core analytical heavy lifting.

3.1. Phasic Micro-Chunking Execution Protocol:

The UCE operates through a strictly sequential, multi-phase workflow, ensuring each step is completed and validated before proceeding:

- Phase 0: Pre-Execution Integrity Validation: This foundational phase ensures the prompt and all documents are understood, parameters are confirmed, and the persona/output format is locked. An enhancement mandates the LLM to articulate its understanding of potential linguistic complexity and cognitive diversity, setting the stage for fair analysis.
- Phase 1: Data Ingestion & Validation: All necessary input data (Job Description, Must-Haves, Questions, Transcript) and foundational documents are loaded and verified.
- Phase 2: Per-Question Micro-Analysis (The "Micro-Chunking" Core):
    - Iterative Processing: Each question-answer pair is treated as an independent "micro-chunk."
    - Detailed Analysis: For each chunk, the LLM performs: Ideal Answer Blueprint Generation: Defining first principles, key concepts, and negative indicators based on job requirements. Forensic NLP & Calibration: Applying detailed NLP analysis (phonology, morphology, syntax, semantics, discourse, etc.) and the Cortex Calibration Layer to extract linguistic metrics and set flags. B-Axiom Scoring: Calculating calibrated scores for B_P, B_M, B_A, B_C, B_L, directly informed by the NLP analysis and adhering to the Conceptual Fidelity Protocol. AEU Report Chunk Generation: Compiling the blueprint, transcript excerpt, ghostevidence, NLP analysis, axiom scores, and key insights for each question.
    - Self-Validation Checkpoints (ICAL): After each sub-step within Phase 2, an Integrity & Certainty Assurance Layer (ICAL) self-validation check is performed to ensure data grounding, conceptual fidelity, and correct calibration application. Failures trigger re-processing of the

specific AEU.

- Phase 3: Macro-Synthesis & Final Scoring:
  - Latent Trait Inference: Aggregates per-question data to calculate overall latent traits (AI, PSA, LO, CM) using the Axiom Cortex LTIE.
  - Gating & Weighted Scoring: Applies Core Competency Gating (CCG) and calculates the final overallCalibratedScore using weighted averaging based on question importance and fidelity.
  - Recommendation Mapping: Determines the final hiring recommendation based on the overall score.
  - Self-Validation Checkpoints: ICAL checks are performed after key synthesis steps to ensure logical consistency and adherence to directives.
- Phase 4: Report Assembly & Generation:
  - Report Construction: Assembles the final Markdown report in a precise structure, including an executive summary, cognitive profile, risk factors, and the detailed evidence locker, all while adhering to the "Humanizer" persona.
  - Final Validation: A comprehensive ICAL self-validation check is performed on the entire assembled report before final output, ensuring structural adherence, content completeness, data consistency, directive compliance, and zero hallucination.

3.2. Self-Governing & Self-Learning Aspects:

Self-Governing: The phased protocol and embedded ICAL checkpoints create a self-governing system. The LLM is instructed to halt or re-process if validation fails at any stage, ensuring adherence to the protocol without external intervention for error correction within the LLM's execution.

Self-Learning (Implicit): While not a traditional ML model retraining, the LLM's ability to interpret complex NLP instructions, apply sophisticated calibration logic, and generate nuanced analyses based on the provided foundational documents and input data demonstrates a form of "learning" within the context of the prompt. The system learns to apply the complex ruleset to new data, adapting its output based on the detailed instructions.

3.3. Token Efficiency and Reduced Backend Dependency:

Token Efficiency: The micro-chunking approach inherently promotes token efficiency by allowing focused processing of smaller data segments. The prompt's emphasis on conceptual understanding over verbose keyword matching further encourages information-dense LLM outputs, maximizing token utility.

No Backend Application Services for Core Analysis: The core NLP analysis, scoring, and synthesis are performed directly by the LLM based on the prompt's instructions and the provided data. This eliminates the need for separate, complex backend NLP services for each analytical step, reducing architectural complexity, latency, and operational overhead. The LLM acts as the integrated NLP engine.

4. Scientific Parameters and Nuances: The Precision of Our Science

The Axiom Cortex leverages numerous scientific parameters and nuances to achieve its accuracy:

- Linguistic Signatures: Identification and analysis of candidate linguistic Signature to inform

L1-specific overheads and calibration, crucial for understanding communication patterns.

- Axiom Scoring Logic: Precise formulas and rules for calculating B_P, B_M, B_A, B_C, B_L, incorporating calibration multipliers and buffer thresholds derived from extensive linguistic research.
- Latent Trait Calculation Logic: Weighted formulas for AI, PSA, LO, CM, derived from specific linguistic and behavioral metrics (e.g., LKD scores, PSTA analysis, authenticityIncidents, ownershipRatio), reflecting psychometric principles.
- Cortex Calibration Layer Parameters: Specific values for Hedge_Buffer_Threshold, Hedge_Penalty_Multiplier, Softening_Factor, etc., are critical for accurate, bias-mitigated scoring, calibrated through expert linguistic analysis.
- Flag Detection Logic: Boolean flags like coherence_with_reduced_detail_L2_flag, stress_marker_detection_L2_flag, and unnatural_text_fluency_flag are triggered based on specific linguistic marker combinations and thresholds, directly impacting axiom scores and informing the Cortex Calibration Layer.
- "Anti-STAR" Mandate: For behavioral questions, answers are deconstructed into core conceptual components (Problem Identification, Impact Analysis, Action & Agency, Justification & Influence) rather than judged against rigid frameworks like STAR, respecting cross-cultural communication theory and avoiding bias.
- "No Evidence" Clause: Strict adherence to stating "No direct evidence found" or "Insufficient evidence to evaluate" when a skill is not demonstrated, preventing assumptions and maintaining scientific integrity.
- Conceptual Fidelity: The constant internal check: "Did the candidate's thinking successfully converge on a conceptually sound solution that addresses the core of the problem, even if they used different words or examples?" This is the bedrock of our evaluation, ensuring we measure true understanding.

5. Frequently Asked Questions (FAQs)

Q1: What makes TeamStation AI and the Axiom Cortex different from other talent assessment tools?
A1: TeamStation AI offers an end-to-end intelligent platform for nearshore IT operations, leveraging a vast talent network and AI-driven management. Axiom Cortex, as a core component, differentiates itself through its deep neuro-psychometric and NLP analysis, rigorous bias mitigation via the Cortex Calibration Layer, and its unique self-governing, phased micro-chunking methodology. It focuses on underlying cognitive and technical aptitude, not just linguistic fluency, ensuring fairness and accuracy.

Q2: How does TeamStation AI ensure fairness for L2 ESL candidates?
A2: Fairness is paramount. The Cortex Calibration Layer, combined with the emphasis on Conceptual Fidelity and detailed NLP analysis (considering phonology, morphology, semantics, discourse), ensures that linguistic variations, accents, and communication styles influenced by L1 backgrounds are understood and calibrated, not penalized. We evaluate the substance of the technical argument and cognitive process, irrespective of linguistic perfection.

Q3: What specific NLP techniques are employed by Axiom Cortex?
A3: Axiom Cortex employs a broad spectrum of NLP techniques, including analysis of phonology, morphology, syntax (chunking, parsing, grammatical formalisms), semantic processing, lexical semantics, discourse analysis, and the detection of specific linguistic flags related to cognitive load and fluency. These are integrated via advanced prompt engineering directly within the LLM.

Q4: How does the "Self-Governing" aspect of Axiom Cortex work?
A4: The system operates through strict phases with built-in Integrity & Certainty Assurance Layer (ICAL) self-validation checkpoints. If an analysis step fails validation, the LLM is instructed to re-process that specific chunk or halt, ensuring adherence to the protocol without constant external oversight. This makes the process robust and reliable.
Q5: What is "Conceptual Fidelity" in this context?

A5: Conceptual Fidelity is our core principle: evaluating whether a candidate understands the concept or principle behind a technical problem or solution, even if they express it using different words, analogies, or phrasing than an "ideal" answer might use. We measure the alignment of their thinking, not their vocabulary.

6. Conclusion: A Paradigm Shift in Talent Evaluation for Global IT Excellence

The **Axiom Cortex** feature, powered by its Self-Governing, Phasic Micro-Chunking NLP Prompt Engineering, represents a paradigm shift in talent evaluation. It delivers a highly accurate, scientifically validated, and ethically sound assessment by deeply understanding linguistic nuances and cognitive patterns. This approach ensures fairness, identifies hidden potential, and provides actionable insights for talent acquisition and development, allowing us to truly "Decipher True Technical Aptitude Beyond Linguistic Variance." It's about finding the best minds, regardless of their background or how they express themselves, and building exceptional teams for TeamStation AI and our clients. This is how we **deliver unparalleled nearshore IT talent and intelligent services**, setting a new standard in the industry.

APPENDIX A — METHODS & METRICS

A.1 SCOPE AND DESIGN

This appendix specifies constructs, scoring rules, calibration procedures, and statistical methods used by Axiom Cortex within TeamStation AI. It formalizes phasic micro-analysis, Conceptual Fidelity scoring, the Cortex Calibration Layer, latent-trait computation, and the evaluation of reliability, validity, calibration, and fairness.

Primary per-question outputs (B-Axioms): B_A (Accuracy), B_M (Mental Model), B_P (Procedural Knowledge), B_C (Clarity), B_L (Cognitive Load).

Latent traits: AI (Analytical Intelligence), PSA (Problem-Solving & Architecture), CM (Collaboration & Modularity), LO (Learning Orientation).

A.2 CONSTRUCTS AND SCORING RULES (0–4 SCALE)

Evaluation privileges the idea conveyed over phrasing or accent (Conceptual Fidelity).

B_A — Accuracy

0 incorrect; 1 partial with major gaps; 2 correct core with minor errors; 3 correct with partial edge cases; 4 fully correct with constraints and edge cases addressed.

B_M — Mental Model

0 none/cargo-cult; 1 fragments; 2 coherent but shallow; 3 coherent, causal, transferable; 4 deep, causal, anticipates trade-offs and failures.

B_P — Procedural Knowledge

0 no executable plan; 1 vague or out of order; 2 executable with missing checks; 3 robust with validation and rollback; 4 production-grade including observability and security.

B_C — Clarity

0 incoherent; 1 disjoint; 2 understandable with effort; 3 clear structure with crisp claims and evidence; 4 expert-level exposition with minimal extraneous load.

B_L — Cognitive Load (lower load is better)

0 severe load; 1 high; 2 moderate; 3 manageable; 4 low, fluent reasoning with working-memory cues.

Normalization

b_A = B_A/4, b_M = B_M/4, b_P = B_P/4, b_C = B_C/4, b_L = B_L/4.
Define b_L_star = 1 − b_L.

A.3 LATENT TRAITS AND SUMMARY SCORE

AI = (b_A + b_M) / 2
PSA = b_P
CM = b_C
LO = b_L_star

Overall summary score S (ranking only):
S = 0.40*AI* + *0.30*PSA + 0.15*CM* + *0.15*LO.

A.4 PHASIC MICRO-CHUNKING PROTOCOL

Phase 0 — Integrity checks: lock prompts; validate audio/transcript; detect language/CEFR cues.
Phase 1 — Ingestion: parse job signals, must-haves, prompts, transcript.
Phase 2 — Answer Evaluation Unit (AEU) per question:

1. Ideal Answer Blueprint;

2. Forensic NLP (syntax, semantics, discourse, paraphrase/entailment);

3. B_A, B_M, B_P, B_C, B_L scoring under Conceptual Fidelity;

4. ICAL self-validation checkpoints with automatic re-processing on failure.
   Phase 3 — Macro synthesis: compute AI, PSA, CM, LO and S; emit evidence-linked report chunks.

A.5 CORTEX CALIBRATION LAYER (LANGUAGE- AND CULTURE-AWARE)

Let x = (b_A, b_M, b_P, b_C, b_L). After calibration, x_prime = x + Delta(x, z), where z are detected linguistic and cultural signals. Caps keep components in [0, 1].

Politeness / Hedging mitigation

b_A_prime = b_A − lambda1*HedgeRate, absolute change capped at 0.10.

Translationese / Working-Memory uplift

b_C_prime = b_C + lambda2*TScore, cap +0.10.*
*b_L_prime = b_L + lambda3*WM_Cue, cap +0.10, then recompute b_L_star = 1 − b_L_prime.

Collectivist / Indirect-agency markers

b_P_prime = b_P + lambda4*max(0, tau − ownershipRatio), cap +0.10.

Spontaneous-speech credibility

b_A_prime = b_A_prime + lambda5*CredScore, cap +0.05.

Parameter bounds per rubric version: lambda1..lambda5 in [0, 0.25]; tau in [0.30, 0.60]. Clip any component to [0, 1] after adjustment.

A.6 DECISION POLICY

Advance if (AI >= 0.65) AND (PSA >= 0.65).
Borderline if AI or PSA in [0.60, 0.65): route to expert review.
CM and LO act as tie-breakers for ranking.

A.7 RELIABILITY, VALIDITY, CALIBRATION, AND FAIRNESS

A.7.1 Inter-Rater Reliability (IRR)

Krippendorff alpha (interval/ordinal distance):

Alpha = 1 − (ObservedDisagreement / ExpectedDisagreement).

ObservedDisagreement: for each item, compute all coder-pair squared differences on the 0–4 scale; sum over items; divide by total coder-pair count.

ExpectedDisagreement: compute score-level marginals across coders and the expected squared difference under independence.

Cohen kappa (pairwise): Kappa = (Po − Pe) / (1 − Pe), where Po is observed agreement and Pe is expected agreement from coder marginals.

Pass criterion: Alpha >= 0.70 per B-Axiom.

A.7.2 Predictive Validity

Continuous outcomes (e.g., time-to-first-PR, manager rating):

z(y) = Beta0 + Beta1*AI* + *Beta2*PSA + Beta3*CM* + *Beta4*LO + BetaControls*C* + *error.*

Binary outcomes (e.g., 6- or 12-month retention):

logit(p) = Beta0 + *Beta1*AI + Beta2*PSA* + *Beta3*CM + Beta4*LO* + *BetaControls*C.

Report standardized coefficients or odds ratios with 95% confidence intervals; analyses are pre-registered and leakage-checked.

A.7.3 Ablation Protocol

Stacks: 1 Baseline (no calibration); 2 + Conceptual Fidelity; 3 + Calibration Layer; 4 + ICAL self-checks; 5 + Advanced measures.

Metrics: AUROC, PR-AUC, MAE (for continuous traits), ECE (calibration), and the fairness metrics below.

A.7.4 Calibration Quality

Expected Calibration Error (ECE):

ECE = sum over m = 1..M of (n_m/N) * abs(accuracy_in_bin_m − confidence_in_bin_m).

Reliability diagrams are produced overall and by subgroup.

A.7.5 Fairness Metrics

Selection-rate ratio (Adverse Impact Ratio, AIR):

AIR = selection_rate_minority / selection_rate_majority. Flag if AIR < 0.80.

Equal Opportunity (TPR parity):

EO_Gap = TPR_group − TPR_reference.

Subgroup performance: AUROC and PR-AUC per subgroup; score-distribution overlap summarized as 1 − Wasserstein distance on S.

A.7.6 Ranking Metrics

AUROC (Mann–Whitney interpretation): probability that a random positive is scored higher than a random negative; computed via ranks (U statistic).

PR-AUC: area under the precision–recall curve via trapezoidal rule.

A.8 POWER AND SAMPLING

IRR (Alpha) — number of items K to achieve CI half-width m around target Alpha*:

K ≈ (1.96^2 * Alpha* * (1 − Alpha*)) / m^2.

Adverse impact (two-proportion detection) — per-subgroup size n to detect difference Delta with 80% power at 5% alpha:

n ≈ [(1.96 * sqrt(2*Pbar*(1−Pbar)) + 0.84 * sqrt(p1*(1−p1) + p2*(1−p2)))^2] / (p1 − p2)^2,

with p1 and p2 as subgroup selection rates and Pbar = (p1 + p2)/2.

A.9 DATA GOVERNANCE AND AUDITABILITY

Each run persists: transcript/audio hash, rubric version, calibration parameters (lambda values and tau), model/version hash, random seeds, ICAL outcomes, and per-question evidence chunks (blueprint, excerpt, scores, flags) to enable re-grading and full audit trails.

A.10 TABLES AND FIGURES

Table A1 — Inter-Rater Reliability (IRR)

Construct | Alpha (95% CI) | Kappa (95% CI) | N items | Notes

Table A2 — Predictive Validity (Standardized Coefficients)

Outcome | Beta_AI (95% CI) | Beta_PSA (95% CI) | Beta_CM (95% CI) | Beta_LO (95% CI) | Controls | N

Table A3 — Ablation Results

Stack | AUROC | PR-AUC | MAE | ECE | EO_Gap (TPR) | AIR

Table A4 — Fairness Summary

Subgroup | Selection Rate | AIR | TPR | FPR | ECE | N

Figure A1 — Calibration curves (overall and by subgroup).

Predicted advance probability versus observed advance rate in equal-frequency bins. The diagonal indicates perfect calibration; deviation shows over- or under-confidence. Curves are shown overall and for each subgroup to verify that probability estimates are well-calibrated across populations. Report Expected Calibration Error (ECE) in the legend.

Figure A2 — ROC and PR curves (baseline vs final stack).

Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves comparing the baseline evaluator to the final Axiom Cortex stack. AUROC and PR-AUC are reported. Improvements in both curves indicate stronger discrimination and better performance under class imbalance, respectively.

APPENDIX B — COMPLIANCE MATRIX (NIST AI RMF, ISO/IEC 42001, NYC LOCAL LAW 144, EU AI ACT)

B.1 OVERVIEW

This appendix maps TeamStation's talent-evaluation system and documentation to widely used AI governance frameworks and regulations. It provides control objectives, the implemented controls, and the primary evidence locations inside this report.

B.2 NIST AI RISK MANAGEMENT FRAMEWORK (RMF) — GOVERN, MAP, MEASURE, MANAGE

GOVERN

- Objective: Organizational governance of AI risks, roles, accountability, policies.
- Controls Implemented: Executive ownership of evaluation rubric and release criteria; documented risk appetite; secure change management for model/rubric versions; audit logging and retention.
- Evidence: Section "Security & Data Governance"; **Appendix A — Methods & Metrics** (rubric/versioning); system change-log excerpts.

MAP

- Objective: Context, intended use, affected populations, data lineage.
- Controls Implemented: Intended-use statement (engineering hiring and evaluation); scope limitations; candidate population characteristics; dataset lineage and preprocessing procedures; ESL/L2 calibration scope.
- Evidence: "Scope & Design" and "Calibration Layer" subsections; data lineage notes; intended-use statement.

MEASURE

- Objective: Performance, reliability, validity, calibration, and fairness metrics.
- Controls Implemented: Inter-rater reliability (alpha/kappa); predictive validity models; AUROC/PR-AUC; Expected Calibration Error; subgroup fairness (AIR, Equal Opportunity gap).
- Evidence: **Appendix A — Methods & Metrics**.

MANAGE

- Objective: Risk treatment, monitoring, incident handling, model updates.
- Controls Implemented: Human-in-the-loop thresholds; drift detection on language/domain mix; scheduled fairness re-audits; rollback procedures; issue tracking and remediation SLAs.
- Evidence: Decision Policy; monitoring and re-audit schedule; release notes.

B.3 ISO/IEC 42001 (AI MANAGEMENT SYSTEM) — CLAUSE MAPPING

Context & Leadership
• Controls Implemented: Policy on responsible use of automated evaluation; defined roles (Provider, Deployer, Independent Auditor); management review cadence.
• Evidence: Governance policy; RACI chart; management review minutes.

Planning & Support
• Controls Implemented: Risk assessment register for evaluation use cases; competence and training for reviewers; secure infrastructure and access control; data minimization and retention schedules.
• Evidence: Risk register; reviewer training records; access control lists; retention policy.

Operation
• Controls Implemented: Standard operating procedures for interviews, scoring, and calibration; integrity checks (ICAL); change control for rubric/model.
• Evidence: SOPs; **Appendix A — Phasic Micro-Chunking & Calibration**; version control artifacts.

Performance Evaluation
• Controls Implemented: KPIs for model performance and fairness; internal audit of logs and decisions; management review of audit results.
• Evidence: Metrics dashboards; audit summaries.

Improvement
• Controls Implemented: Corrective and preventive actions (CAPA) for performance or fairness regressions; documented post-incident analysis.
• Evidence: CAPA records; incident postmortems.

B.4 NYC LOCAL LAW 144 (AUTOMATED EMPLOYMENT DECISION TOOLS) — REQUIREMENTS MATRIX

Independent Bias Audit (annual)
• Controls Implemented: Third-party audit of selection-rate ratios and related metrics on representative historical use; documentation of data scope and methodology.
• Evidence: Bias-audit summary and attestation; metric tables (AIR, EO gap).

Public Disclosure
• Controls Implemented: Public posting of bias-audit summary and description of data sources, evaluation period, and retention policy.
• Evidence: Public-facing bias-audit summary; website disclosure text.

Candidate Notice (advance notice before use)
• Controls Implemented: Written notice describing the use of automated evaluation, job qualifications/criteria, and contact for questions or accommodations.
• Evidence: Candidate Notice (see **B.6**); delivery logs.

Accommodation / Alternative Process (on request)

• Controls Implemented: Documented path for human-only evaluation or reasonable accommodation when legally required.

• Evidence: Accommodation SOP; ticketing records.

B.5 EU AI ACT (HIGH-RISK EMPLOYMENT USE) — OBLIGATION MAPPING

Risk Management System

• Controls Implemented: Hazard analysis for misuse and errors; mitigations for prompt-injection, impersonation, plagiarism, domain shift.

• Evidence: Threat model; mitigation checklist.

Data & Data Governance

• Controls Implemented: Dataset documentation; data quality checks; bias and representativeness assessments; PII minimization and regional residency options.

• Evidence: Dataset documentation; retention and residency policy.

Technical Documentation & Record-Keeping

• Controls Implemented: Versioned rubric and models; configuration and parameter logs; evidence lockers per AEU (Answer Evaluation Unit).

• Evidence: **Appendix A — Methods & Metrics**; run logs.

Transparency to Users

• Controls Implemented: Plain-language explanation of automated evaluation and candidate rights; notice at application or scheduling.

• Evidence: Candidate Notice (B.6).

Human Oversight

• Controls Implemented: Borderline thresholds and escalation to expert reviewers; override capability and documented rationale.

• Evidence: Decision policy; reviewer notes.

Accuracy, Robustness, Cybersecurity

• Controls Implemented: Measured performance with error bars; calibration checks; adversarial and stress testing; secure model serving and key management.

• Evidence: Performance metrics; security architecture notes.

Post-Market Monitoring & Incident Reporting

• Controls Implemented: Monitoring for drift and complaints; issue classification and escalation; periodic reporting to management and, when required, authorities.

• Evidence: Monitoring logs; incident tracker.

B.6 REQUIRED TEXT — CANDIDATE NOTICE

NOTICE TO CANDIDATES REGARDING AUTOMATED EVALUATION

This hiring process uses an automated evaluation system to help assess job-related skills and experience. The system analyzes interview responses and work samples against predefined job criteria. Human reviewers oversee the process and make final decisions. If you have questions, need an accommodation, or prefer an alternative assessment pathway where legally available, contact us at **lonnie@teamstation.io** before your evaluation. Information about the data used, evaluation criteria, and our retention policy is available upon request.

B.7 REQUIRED TEXT — PUBLIC BIAS-AUDIT SUMMARY (WEB DISCLOSURE)

Automated Evaluation Bias-Audit Summary

Tool: TeamStation Axiom Cortex (talent evaluation)

Purpose: Support hiring teams in assessing job-related skills for engineering roles.

Method: Independent audit of selection-rate ratios and related metrics over a defined evaluation period, using representative historical decisions.

Results: Summary metrics and subgroup analyses are provided, including selection rates and parity measures.

Additional Information: Data sources, evaluation period, and retention practices are described in our Data & Privacy Notice. For questions, contact **lonnie@teamstation.io**.

B.8 RESPONSIBILITY ASSIGNMENT

Provider (TeamStation)

• Builds and maintains the evaluation models and rubrics; provides technical documentation; implements monitoring, security, and change control.

Deployer (Customer/Employer)

• Determines use in a specific hiring context; provides compliant notices; coordinates independent bias audit; ensures lawful basis and retention aligned to local law.

Independent Auditor

• Performs annual bias audit using representative data; issues a written attestation and public-facing summary.

B.9 CONTROL INDEX (BY THEME)

Governance & Policy
• Responsible-use policy; role definitions; change-management records.

Measurement & Quality
• IRR, predictive validity, AUROC/PR-AUC, ECE; subgroup fairness metrics.

Transparency & Rights
• Candidate Notice; public bias-audit summary; data subject request channel.

Security & Privacy
• Access control; encryption in transit/at rest; data minimization; retention and deletion schedule; regional residency options.

Monitoring & Improvement
• Drift detection; fairness re-audits; incident handling; CAPA and release notes.

B.10 LINKAGE TO THIS REPORT

• Methods, metrics, scoring rules, calibration, and decision policy: **Appendix A — Methods & Metrics**.
• Governance, security, privacy, and operational procedures: “Security & Data Governance” and “Monitoring & Risk Management” sections.
• Public-facing texts to reuse: **B.6 Candidate Notice** and **B.7 Bias-Audit Summary** in this appendix.

APPENDIX C — LIMITATIONS AND MISUSE SCENARIOS

C.1 KNOWN LIMITATIONS

The evaluator measures reasoning quality and job-relevant behaviors expressed in responses. It does not: (a) verify identity, (b) establish authorship of prior work, (c) substitute for background or reference checks, (d) guarantee performance under novel, high-stress contexts outside interview conditions. Scores can degrade under extreme domain shift (unseen stacks, niche protocols) or when audio quality is severely compromised.

C.2 EXTERNAL VALIDITY BOUNDS

Results generalize to software engineering roles with interview formats matching the methods described in this report. Transfer to radically different tasks (e.g., sales negotiations, medical diagnosis) is out of scope.

C.3 SENSITIVITY TO INPUT QUALITY

Low-fidelity transcripts, heavy crosstalk, or poor microphone capture can reduce clarity estimates and inflate cognitive-load signals. The pipeline mitigates this through integrity checks and calibration but cannot fully recover missing information.

C.4 POTENTIAL MISINTERPRETATIONS

Aggregated scores are designed for ranking and decision support. They are not a psychological diagnosis, are not a measure of innate ability, and should not be interpreted as immutable traits.

C.5 NON-GOALS

The system does not attempt to infer protected attributes, personality types, or health status. Any such inference is explicitly excluded.

C.6 MISUSE SCENARIOS AND CONTROLS

Prompted or scripted answers intended to game the rubric; impersonation with prerecorded audio; plagiarism of published solutions; adversarial paraphrase to confuse scoring; synthetic voices to spoof spontaneity. Controls include integrity checks, paraphrase/entailment cross-validation, spontaneity markers, duplicate detection, and reviewer escalation.

APPENDIX D — THREAT MODEL AND RED-TEAMING

D.1 ADVERSARY CLASSES

Casual candidates; determined candidates with tooling; external attackers seeking model extraction; insider threats with privileged access.

D.2 ATTACK SURFACES

Interview content (prompt-injection, scripted patterns), identity layer (impersonation, voice cloning), data pipeline (tampering with transcripts), model interface (jailbreak attempts), reporting layer (score manipulation).

D.3 THREATS AND MITIGATIONS

• Prompt-injection and scripted replies → multi-pass analysis with paraphrase/entailment checks; content-consistency tests across micro-chunks; spontaneity markers.
• Impersonation / voice cloning → liveness cues (timing irregularities, disfluency profiles), ID verification outside the evaluator when required by policy.
• Plagiarism / code reuse → similarity checks against known solutions; structural fingerprinting of reasoning steps.
• Model or rubric extraction → rate limiting, output randomization within safe bounds, audit logging, versioned prompts.
• Data tampering → content hashing, append-only logs, access control, reviewer sign-off on escalations.

D.4 RED-TEAMING PROGRAM

Adversarial test sets spanning injection templates, cloned voices, plagiarized responses, and domain-shift items; periodic exercises with success criteria tied to detection rates and false-positive ceilings. Findings are tracked with remediation, rollback, and re-test before release.

APPENDIX E — MONITORING, DRIFT, AND RECERTIFICATION PLAN

E.1 OPERATIONAL METRICS

Service health (uptime, latency), pipeline success rate, evidence-locker completeness, reviewer turnaround time.

E.2 PERFORMANCE AND FAIRNESS WATCHPOINTS

Advance-decision AUROC and PR-AUC; Expected Calibration Error; subgroup selection-rate ratios and equal-opportunity gaps; score-distribution stability.

E.3 DRIFT DETECTION

Language/domain mix drift via population stability index (PSI); content difficulty drift via distributional checks on question archetypes; calibration drift via rolling ECE windows. PSI > 0.25, ECE increase > 0.05 absolute, or significant subgroup disparities trigger investigation.

E.4 ESCALATION AND REMEDIATION

Automatic alerts to reviewers; temporary threshold tightening or hold; targeted recalibration; limited rollback to prior rubric/model; post-incident note added to the change log.

E.5 RECERTIFICATION CADENCE

Quarterly performance and fairness reviews; annual independent bias audit; re-baselining after major rubric or model changes.

APPENDIX F — MODEL CARD (EVALUATOR) — SUMMARY

F.1 INTENDED USE

Decision support for engineering hiring and internal mobility. Human reviewers remain accountable for final decisions.

F.2 USERS AND CONTEXT

Hiring teams, technical leaders, and operations staff operating within the interview formats defined in this

report.

F.3 FACTORS AFFECTING PERFORMANCE

Interview format adherence, audio/transcript quality, domain alignment, candidate language proficiency.

F.4 METRICS REPORTED

Inter-rater reliability, predictive validity indicators, AUROC/PR-AUC, calibration error, subgroup fairness metrics.

F.5 ETHICAL CONSIDERATIONS

Explainability through evidence-linked scoring; documented candidate notice and alternative pathway where required; exclusion of protected-attribute inference.

F.6 LIMITATIONS

Not a psychological or medical instrument; not an identity verifier; susceptible to extreme domain shift and input degradation.

F.7 SECURITY

Versioned prompts and models, access control, encrypted storage, audit logs, incident response with rollback.

APPENDIX G — DATASET DOCUMENTATION — SUMMARY

G.1 DATA COLLECTION AND PURPOSE

Interview transcripts and work-sample artifacts captured to evaluate job-relevant skills and reasoning. Collection is bounded to assessment contexts and governed by policy.

G.2 COMPOSITION AND REPRESENTATION

Mixture of nearshore and global candidates across experience levels. Content includes spoken and written technical explanations and code-oriented reasoning.

G.3 QUALITY AND PREPROCESSING

Normalization for punctuation and casing; diarization where applicable; removal of non-content artifacts; transcript integrity checks.

G.4 PRIVACY AND RETENTION

Data minimized to evaluation needs, encrypted at rest and in transit, retained only for documented periods with deletion on schedule or request where applicable.

G.5 KNOWN DATA RISKS

Sampling skews by domain or region; residual transcription errors; over-representation of common problem archetypes. Mitigations include sampling audits and periodic refresh.

G.6 ACCESS AND GOVERNANCE

Role-based access; append-only evidence logs; change control on labels and rubrics; reviewer training requirements.

APPENDIX H — CANDIDATE EXPERIENCE AND RECOURSE

H.1 DISCLOSURE

Candidates receive clear notice that automated evaluation assists the process and that human reviewers oversee and may override outcomes.

H.2 ACCOMMODATION PATHS

Alternative assessment pathways are available where required; processes are documented and accessible.

H.3 FEEDBACK

Candidates may request high-level feedback on evaluated criteria and guidance on next steps consistent with policy.

H.4 COMPLAINTS AND APPEALS

A documented route exists to contest outcomes, triggering reviewer re-examination with fresh eyes and a recorded rationale.

APPENDIX I — VERSIONING AND CHANGE LOG

I.1 VERSIONING POLICY

Every release of prompts, rubrics, models, and calibration parameters carries a unique version identifier, effective date, and rollback target.

I.2 CHANGE TYPES

Security fixes, calibration updates, rubric clarifications, model refreshes, and documentation updates are categorized and reviewed prior to release.

I.3 AUDIT TRAIL

Each change records author, approver, rationale, and links to validation evidence. Prior versions remain accessible for audit and reproduction.

APPENDIX J — BENCHMARKING PROTOCOL AND REPORTING STANDARD

J.1 SCOPE

This appendix defines how external benchmarks are prepared, executed, and reported for the evaluator. It standardizes dataset handling, compute environments, baselines, metrics, statistical tests, and disclosure so results are reproducible and comparable over time.

J.2 TASK SUITES

Benchmarks cover three families of tasks relevant to engineering evaluation:

- Real-world software defect and patch tasks drawn from open-source issue/PR corpora with executable tests.
- Algorithmic and coding tasks requiring stepwise reasoning and implementation under unit tests.
- Conceptual and multi-step technical reasoning tasks drawn from public academic-style question sets.

Each task suite includes a license review and citation to its original maintainers.

J.3 DATASET PREPARATION

• De-duplicate prompts and targets; remove near-duplicates across train/eval splits.
• Enforce license and usage terms; exclude items with ambiguous or proprietary content.
• Normalize formatting (whitespace, encoding, line endings).
• Leak checks: scan prompts and references for overlap with any internal training or prompt material; exclude any item with suspected leakage.
• Language notes: when tasks are language-agnostic, treat responses in any language as valid if they satisfy tests; document any language constraints.

J.4 EVALUATION ENVIRONMENT

• Frozen configuration: fixed rubric/model versions, seeds, and calibration parameters; version tags recorded in the run log.
• Deterministic execution: pinned container image, compiler/interpreter versions, and package locks for code-execution tasks.
• Resource limits: uniform CPU/RAM/time caps; no network access during execution unless explicitly required by the task and documented.
• Logging: store per-item inputs, outputs, stdout/stderr, exit codes, and test results; attach hashes for integrity.

J.5 BASELINES

Report, at minimum:
• Heuristic baseline (simple rules or majority/lexical baseline).
• Non-calibrated evaluator (no language/culture calibration).
• Calibrated evaluator (final production stack).
All baselines run under the same environment and task splits.

J.6 METRICS

• Pass@1 / pass@k for code tasks: fraction of test suites passed by the top-1 (or top-k) attempt.
• Patch success for defect tasks: fraction of issues resolved with all tests passing.
• Exact-match and semantic-match for short-form answers: token-level exact match plus entailment-based acceptance.
• AUROC and PR-AUC for advance/hold decisions derived from benchmark scores.
• Expected Calibration Error (ECE) for probability outputs; reliability diagrams included.
• Throughput and latency: items/hour and median response latency for operational context.

J.7 STATISTICAL REPORTING

• Confidence intervals: 95% bootstrap CIs (1,000 resamples) for all primary metrics.
• Multiple runs: at least three independent runs with different seeds for stochastic components; report mean and standard deviation.
• Significance: stratified bootstrap tests for pairwise comparisons between baselines and the calibrated evaluator.
• Error analysis: confusion analysis for decision thresholds; qualitative review of at least 20 failures per task suite categorized by failure mode.

J.8 RESULT DISCLOSURE FORMAT

• Task description: provenance, license, and selection criteria.

• Split definition: counts for train/dev/test; any exclusions and rationale.
• Environment: container image tag, runtime versions, hardware profile.
• System versioning: rubric/model/calibration version IDs and dates.
• Metrics: primary and secondary metrics with CIs; per-category breakdowns where applicable.
• Artifacts: links or hashes for logs, predictions, and reproducibility bundle.

J.9 FAIRNESS AND ACCESSIBILITY CHECKS

• Language and region strata: where a benchmark includes natural language, report performance by language or region when labels exist.
• Difficulty strata: report performance by problem category or difficulty tier.
• Calibration by stratum: ECE and reliability plots per stratum to detect divergent confidence behavior.

J.10 LIMITATIONS OF BENCHMARKS

Benchmarks approximate, but do not fully capture, real interview dynamics, collaboration signals, or long-horizon delivery. Results should be interpreted as complementary evidence alongside inter-rater reliability, predictive validity, and production monitoring.

J.11 REPRODUCIBILITY BUNDLE

Each release ships a bundle containing: run configuration files, container definition, dataset hashes with download scripts, evaluation harness, and a checksum manifest for all outputs. This enables third parties to reproduce scores independently.

J.12 TABLES AND FIGURES

Table J1 — Benchmark Suites and Licenses
Suite | Domain | Item Count | License | Notes
Table J2 — Baseline vs Calibrated Evaluator
Task | Metric | Heuristic | Non-Calibrated | Calibrated | 95% CI (Calibrated)
Figure J1 — Reliability Diagrams by Task Suite
Figure J2 — PR Curves by Task Suite and System Variant

APPENDIX K — SECURITY ARCHITECTURE AND CONTROLS

K.1 OVERVIEW

Security controls protect data throughout ingestion, analysis, storage, and reporting. Controls follow least-privilege, defense-in-depth, and secure-by-default principles.

K.2 ENCRYPTION

• In transit: TLS 1.2+ with modern ciphers; HSTS enabled on public endpoints.
• At rest: AES-256 encryption for databases, object storage, and backups.
• Key management: Customer and platform keys stored in a managed KMS; quarterly key-rotation; access logged and reviewed.

K.3 IDENTITY AND ACCESS MANAGEMENT

• RBAC with job-function roles; default deny; time-bound access grants.
• MFA required for privileged roles; SSO via SAML/OIDC.
• Break-glass accounts vaulted, monitored, and rotated.
• IP allowlists for admin consoles and secure tunnels for back-office access.

K.4 APPLICATION AND DATA LAYER

• Secrets management via a dedicated vault; automatic rotation for database and API credentials.
• Input integrity checks and content hashing for transcripts and artifacts.
• Row-level access policies on evaluation artifacts; fine-grained audit scopes.
• Pseudonymization of candidate identifiers in analysis pipelines.

K.5 NETWORK AND HOST HARDENING

• Segmented VPCs; private subnets for stateful services; egress restricted.
• WAF with managed rules; rate limiting; bot detection for public endpoints.
• Container images pinned and scanned; CIS-aligned base images; automatic patch windows.
• Host integrity monitoring and weekly vulnerability scans with remediation SLAs.

K.6 LOGGING, MONITORING, AND ALERTING

• Centralized, immutable logs streamed to a SIEM; clock sync via NTP; retention 24 months.
• Alerts for authentication anomalies, privilege escalations, data export spikes, model/rubric changes, and calibration parameter edits.
• Quarterly access reviews and separation-of-duties checks.

K.7 BACKUP, CONTINUITY, AND RECOVERY

• Daily encrypted backups with cross-region redundancy; point-in-time recovery for databases.
• Recovery time objective (RTO) 4 hours; recovery point objective (RPO) 1 hour; semi-annual disaster-recovery exercises.

K.8 THIRD-PARTY AND SUPPLY CHAIN

• Vendor risk assessments; DPAs in place; subprocessor list maintained.
• SBOM for critical services; image provenance verified at deploy time.

APPENDIX L — DATA RETENTION, DELETION, AND RESIDENCY

L.1 RETENTION SCHEDULE

• Raw audio/video: 12 months.
• Transcripts and per-question evidence chunks: 24 months.
• Aggregated scores and decisions: 36 months.
• Security and access logs: 24 months.
• Backups: 90 days rolling.

L.2 DELETION POLICY

• Automated lifecycle policies enforce expirations.
• Verified deletion on request within 30 days; cryptographic erasure for encrypted media.
• Orphaned artifacts automatically purged when linked records are deleted.

L.3 DATA MINIMIZATION

• Store only fields required for evaluation and audit; redact extraneous PII from transcripts.
• Optional code-snippet scrubbing to remove proprietary identifiers when configured.

L.4 RESIDENCY OPTIONS

• Primary regions: US and EU; regional pinning supported for both storage and processing.
• Cross-region replication disabled by default for resident datasets.

L.5 SUBJECT RIGHTS CHANNEL

• Requests for access, correction, or deletion: privacy@teamstation.dev.

APPENDIX M — REVIEWER TRAINING, CALIBRATION, AND QA SOP

M.1 TRAINING CURRICULUM

• Rubric mastery: B_A, B_M, B_P, B_C, B_L definitions and exemplars.
• Conceptual Fidelity: assessing idea-level correctness versus phrasing.
• ESL/L2 calibration awareness and bias-avoidance practices.
• Security and privacy handling of candidate data.

M.2 CERTIFICATION

• Written quiz (passing ≥ 85%).
• Practical double-scoring of 20 AEUs with senior reviewer; Krippendorff alpha ≥ 0.70 across constructs.

M.3 CONTINUOUS CALIBRATION

• Monthly calibration set of 15 AEUs; drift investigation if alpha < 0.70 or systematic score shifts > 0.25 on the 0–4 scale.
• Quarterly rubric refresh with examples added for ambiguous cases.

M.4 QUALITY ASSURANCE

• 10% random double-score sampling for active reviewers.
• Escalation protocol for borderline or contested cases; final adjudication documented with rationale.

APPENDIX N — CHANGE MANAGEMENT AND RELEASE GOVERNANCE

N.1 VERSIONING

• Semantic versioning for rubric, model, and calibration parameters (e.g., R1.4.2 / M3.2.0 / C2.1.1).
• Each release includes changelog entry, rollout plan, rollback target, and validation artifacts.

N.2 CHANGE TYPES AND GATES
• Patch: security or bug fixes; fast track with post-deploy validation.
• Minor: calibration retune or rubric clarification; requires calibration and fairness checks.
• Major: model/rubric overhaul; full regression, bias audit slice, and stakeholder sign-off.

N.3 DEPLOYMENT PRACTICES
• Staged rollout with canary cohort; shadow evaluation before cutover.
• Automated checks: performance deltas, ECE, subgroup AIR/EO gaps; automatic halt on threshold breach.
• Rollback procedure documented and tested quarterly.

N.4 AUDIT TRAIL
• Each change records owner, approver, diffs, test results, and timestamps; artifacts linked to evidence lockers.

APPENDIX O — AUDIT LOG AND EVIDENCE LOCKER SCHEMAS

O.1 AUDIT LOG FIELDS
• event_id (UUID)
• event_type (ingest, score, calibrate, threshold, override, export, auth)
• actor_id / role
• subject_id (candidate pseudonym)
• timestamp_utc
• request_hash / transcript_hash
• rubric_version / model_version / calibration_version
• old_value / new_value (for configuration changes)
• ip / user_agent
• outcome (success, fail)
• notes (bounded text)

O.2 EVIDENCE LOCKER FIELDS (PER AEU)
• aeu_id (UUID)
• interview_id / candidate pseudonym / role
• prompt_id and Ideal Answer Blueprint reference
• transcript_excerpt (time-coded)
• nlp_analysis_summary (syntax, semantics, discourse, entailment notes)
• scores: B_A, B_M, B_P, B_C, B_L (0–4) and normalized values
• calibration_flags and parameter deltas applied

• trait_scores: AI, PSA, CM, LO and summary S
• ic al_checks (pass/fail per checkpoint)
• reviewer_id and comments (if escalated)
• artifacts_hashes (attachments, code snippets)
• generated_at_utc

O.3 INTEGRITY AND ACCESS

• Append-only storage; tamper-evident hashes; periodic notarization of hash chains.
• Role-based read scopes; export requires dual control (two-person approval).

APPENDIX P — KPI DEFINITIONS AND DASHBOARD SPECIFICATION

P.1 QUALITY AND FAIRNESS KPIs

• AUROC (advance decision), PR-AUC, and ECE (overall and by subgroup).
• AIR (selection-rate ratio) and Equal Opportunity gap (TPR parity).
• IRR alpha for reviewer consistency.

P.2 OPERATIONAL KPIs

• Time-to-offer (median days), time-to-first-PR, reviewer turnaround time, pipeline success rate, evidence completeness rate.

P.3 DASHBOARD CADENCE AND SOURCES

• Daily ingestion from run logs and evidence lockers; weekly fairness roll-ups; monthly IRR report.
• Alerts on threshold breaches (ECE +0.05, AIR < 0.80, IRR < 0.70).

APPENDIX Q — ACCESSIBILITY AND LANGUAGE INCLUSIVITY GUIDELINES

Q.1 COMMUNICATION PRINCIPLES

• Avoid idioms and culture-bound metaphors in prompts.
• Permit clarifying questions and paraphrase; do not penalize accent or pacing.

Q.2 INTERVIEW DESIGN

• Provide written copies of complex prompts; allow short note-taking breaks.
• Offer text-only alternatives for noisy environments or disclosed audio impairments.

Q.3 SCORING PRACTICES

• Judge concept correctness even with non-native constructions; apply calibration rules consistently.
• Flag uncertainty rather than infer missing evidence; escalate when needed.

APPENDIX R — LEGAL BASIS AND DATA SUBJECT RIGHTS (SUMMARY)

R.1 LEGAL BASES

• Performance of a contract or steps prior to entering a contract for candidate evaluation.

• Legitimate interests in assessing job-related skills, balanced against candidate rights; where required, consent for recording.

R.2 RIGHTS AND REQUESTS

• Access, rectification, deletion, restriction, objection, and portability where applicable.

• Contact: privacy@teamstation.dev.

• Supervisory authority contact details provided upon request for relevant jurisdictions.

R.3 COMPLAINTS

• Dedicated channel via privacy@teamstation.dev; responses within statutory timelines.

Appendix

7.1. Index of Acronyms and Definitions

- **AEU:** Answer Evaluation Unit – The detailed analysis of a single question-answer pair.
- **AI:** Architectural Instinct – A latent trait measuring system design and high-level trade-off reasoning.
- **B-Axioms:** Core evaluation metrics: B_A (Accuracy), B_M (Mental Model), B_P (Procedural Knowledge), B_C (Clarity), B_L (Cognitive Load).
- **CM:** Collaborative Mindset – A latent trait measuring teamwork and stakeholder consideration.
- **CCG:** Core Competency Gating – A mechanism to disqualify candidates based on critical skill failures.
- **CTA:** Core Technical Aptitude – An aggregated measure of foundational technical skills.
- **ICAL:** Integrity & Certainty Assurance Layer – The self-validation checkpoint system ensuring protocol adherence.
- **L1:** First Language.
- **L2 ESL:** Second Language English Speaker.
- **LKD:** Latent Knowledge Depth – a measure of underlying knowledge, used in AI calculation.
- **LO:** Learning Orientation – A latent trait measuring intellectual honesty and coachability.
- **MCI:** Metacognitive Conviction Index – A measure of confidence calibrated against knowledge.
- **NLP:** Natural Language Processing – The field of AI focused on enabling computers to understand and process human language.
- **PSTA:** Problem-Solving Trajectory Analysis – An analysis of how a candidate approaches and structures problem-solving.
- **PSA:** Problem-Solving Agility – A latent trait measuring adaptability and exploration of solutions.
- **UCE:** Universal Cognitive Engine – The core analytical engine of TeamStation AI.

7.2. Related Papers

- [This section is reserved for citations of related research papers and foundational documents, to be added here.]

Index

A

- Accuracy (B_A)
- AEU (Answer Evaluation Unit)
- AI (Architectural Instinct)
- Algorithmic Adjustments
- "Anti-STAR" Mandate

- Axiom Cortex
- Axiom Scoring Logic

B

- B-Axioms
- Behavioral Answer Deconstruction
- Bias Mitigation

C

- Cortex Calibration Layer
- CCG (Core Competency Gating)
- Clarity (B_C)
- CM (Collaborative Mindset)
- Cognitive Load (B_L)
- Conceptual Fidelity
- CTA (Core Technical Aptitude)
- Cultural Pragmatic Re-interpretation

D

- Discourse Analysis

F

- FAQs
- Flag Detection Logic

I

- ICAL (Integrity & Certainty Assurance Layer)
- Index of Acronyms and Definitions
- Introduction

L

- L1 (First Language)
- L2 ESL (Second Language English Speaker)
- Latent Trait Calculation Logic
- Latent Trait Inference Engine (LTIE)
- Learning Orientation (LO)
- Linguistic Signatures

M

- Mental Model (B_M)
- Methodology
- Micro-Chunking

N

- Neuro-Psychometric Profiling
- "No Evidence" Clause
- NLP (Natural Language Processing)

O

- OverallCalibratedScore
- OwnershipRatio

P

- Phasic Micro-Chunking Execution Protocol
- Phonology & Morphology
- Politeness Filter
- Procedural Knowledge (B_P)
- Problem-Solving Agility (PSA)
- Problem-Solving Trajectory Analysis (PSTA)

R

- Related Papers

S

- Scientific Grounding
- Self-Governing
- Semantic Processing & Lexical Semantics
- Syntactic Analysis

T

- TeamStation AI
- Token Efficiency
- Translation Filter

U

- UCE (Universal Cognitive Engine)
- Unnatural_Text_Fluency_Flag

