Received 22 May 2026, accepted 15 September 2026, date of publication 30 September 2026
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.25.3.4856

Fault-Tolerant HR Evaluation Methodology Using Multi-Stage Local LLM Pipelines

VITALII PAVLIUK, VOLODYMYR DREVETSKYI
National University of Water and Environmental Engineering (NUWEE), Rivne, Ukraine

Corresponding author: Vitalii Pavliuk (e-mail: v.v.pavlyuk@nuwm.edu.ua).

ABSTRACT Personnel screening for safety-critical industrial roles demands simultaneously high evaluative accuracy, strict data sovereignty, and auditable decision rationale — three requirements that cloud-based artificial intelligence deployments often struggle to satisfy concurrently under regulatory constraints on personal-data egress. This paper proposes and empirically evaluates a fault-tolerant, privacy-preserving pipeline for automated candidate evaluation in high-hazard industrial environments. The methodology integrates a five-criterion Multi-Criteria Decision Analysis (MCDA) model with a weighted Likert aggregation formula, Behaviourally Anchored Rating Scales (BARS) embedded in structured rationale generation inspired by step-by-step prompting, asynchronous task orchestration via a Redis/Celery message broker, and a locally deployed Llama 3 8B Instruct model served through vLLM with PagedAttention and prefix caching. Evaluation comprised two experiments: a controlled 50-record cross-model comparison against GPT-4o-mini, yielding a Spearman rank correlation of 0.745 and 78% score agreement within one Likert point; and a retrospective observational method-performance evaluation on an anonymised dataset of 1,534 historical questionnaire records, producing a dataset-wide distribution of model output bands and a complete audit trail. A first-author manual review of ten sampled cases identified one unsupported evidence citation among the five inspected Recommended-category outputs (1/5, 20%; equivalently 1/10, 10% of all inspected cases), together with role-track conflation, evidence under-citation, and credential over-weighting. The system is proposed for further evaluation as a human-supervised decision-support method; Experiment B did not form part of any recruitment decision, and the small author-reviewed sample does not support an accuracy claim or a dataset-level error-rate estimate.

KEYWORDS large language models; HR screening; multi-criteria decision analysis; fault-tolerant pipelines; data sovereignty; privacy-preserving AI; structured rationale generation; behaviourally anchored rating scales; critical infrastructure; explainable AI.

I. INTRODUCTION

The operational safety of high-hazard industrial facilities is inseparably linked to the quality and competency of the personnel entrusted with their operation and maintenance. Such facilities operate under a regulatory framework that mandates rigorous qualification standards for every worker admitted to safety-critical operational areas [1]. The traditional human-resources workflow — manual review of paper or electronic questionnaires by specialist recruiters — scales poorly in the face of mass-recruitment events or facility expansions and introduces systematic inter-rater variability which is fundamentally incompatible with the safety culture and human-AI oversight requirements relevant to high-stakes decision-support systems [2].

The emergence of large language models (LLMs) as capable evaluators of unstructured textual data presents a compelling avenue for automating the preliminary screening phase of the recruitment pipeline [3], [4]. LLMs can simultaneously parse educational credentials, interpret free-form experience descriptions, and reason about regulatory compliance in a manner that extends beyond classical keyword-matching approaches used in earlier recruitment systems [3], [4]. However, the deployment of commercial cloud-based LLM services within safety-critical industrial contexts is constrained by data-sovereignty requirements: sensitive personnel records, including psychometric profiles, security-clearance histories, and health declarations, may raise compliance challenges under GDPR and related data-protection regulations, national data-protection legislation, as well as human-AI interaction and national AI-governance recommendations [2], [5]. This constraint defines the "zero-egress" architectural boundary within which any viable AI-assisted HR system must operate.

Beyond the privacy constraint, a second fundamental challenge concerns reliability. LLM inference is subject to non-deterministic failure modes — malformed structured output,

hallucinated credentials, latency spikes under load — which are inadmissible in an administrative system whose outputs directly inform safety-critical hiring decisions [6]. A production-grade pipeline must therefore implement multi-layer fault tolerance that guarantees decision continuity even when individual inference calls partially or completely fail. A third dimension of the problem is explainability. Regulatory auditors and HR compliance officers require not merely a numerical score but a documented rationale that demonstrates how each sub-criterion was evaluated. Opaque embeddings-based similarity systems provide no such trace [7]. Systems that explicitly generate structured reasoning traces can improve traceability and support this requirement.

The present paper proposes a unified methodology that resolves these three tensions simultaneously. The principal scientific contributions are: (i) a formally specified MCDA scoring model with domain-informed criteria aligned with industrial safety staffing standards; (ii) a prompt-engineering architecture combining BARS rubrics with structured rationale generation; (iii) a fault-tolerant asynchronous pipeline with mathematically bounded availability; (iv) a locally deployed vLLM inference stack with prefix-caching optimisation supporting privacy-preserving deployment without significant performance trade-offs; (v) an empirical classification of deployment-relevant LLM failure modes in regulated HR workflows. The architecture is evaluated against a cloud-based baseline and assessed with respect to ISO/IEC 25059:2023 quality attributes [8].

The remainder of the paper is structured as follows. Section 2 reviews related work across four interconnected literature domains. Section 3 describes the materials and methods in detail. Section 4 presents results and discussion. Section 5 concludes with implications for safety-critical industrial HR practice and future research directions.

II. RELATED WORK

A. AUTOMATED RECRUITMENT AND AI-POWERED SCREENING

Gan et al. [3] introduced a multi-agent framework decomposing résumé analysis into extraction, evaluation, and formatting sub-agents — a structural precedent directly applicable to the criterion-parallel architecture proposed in this work. Møller et al. [9] demonstrated that LLM-augmented data can support low-resource classification workflows, which is relevant to recruitment contexts where labelled expert datasets are often limited. Jiang et al. [10] showed that ensemble strategies combining multiple LLM outputs can improve answer selection and generation quality, supporting the use of comparative or multi-model evaluation strategies. Tam et al. [11] analysed how LLMs behave under structured-format constraints, highlighting the importance of reliable JSON-like output for downstream automation pipelines. Raghavan et al. [4] examined bias-mitigation claims in algorithmic hiring systems, thereby motivating the need for auditable, human-supervised evaluation frameworks such as the BARS-anchored approach adopted in this work.

B. PROMPT ENGINEERING AND STRUCTURED RATIONALE GENERATION

Wei et al. [12] introduced Chain-of-Thought prompting as a technique for eliciting intermediate reasoning steps before producing a final answer. In the present work, this idea is adapted not as unrestricted hidden reasoning, but as structured rationale generation in which each score must be accompanied by an explicit CLAIM, EVIDENCE, and VERDICT. This property is particularly important in HR evaluation, where each numerical score should be traceable to candidate-specific evidence.

Reynolds and McDonell [13] showed that prompt programming can guide model behaviour beyond the few-shot paradigm, supporting the use of explicit task instructions and output schemas. Yao et al. [14] extended step-by-step reasoning into Tree of Thoughts, illustrating how branching deliberation can support complex multi-step evaluation. Huang et al. [6] systematised hallucination risks in LLMs, reinforcing the need for evidence-grounded prompts and validation layers. Recent work on LLM-based autonomous agents [15] provides a complementary framework for analysing error propagation across chained evaluation steps, applicable to the five-criterion pipeline proposed in this paper.

C. DATA SOVEREIGNTY AND PRIVACY-PRESERVING LOCAL INFERENCE

The zero-egress architectural constraint for sensitive data processing is well established in adjacent high-assurance domains. Miranda et al. [16] systematise privacy threats and mitigation strategies for large language models, including risks related to training-data leakage, prompt disclosure, model inversion, and sensitive-data exposure. This supports the motivation for zero-egress deployment in workflows involving personnel records and other regulated personal data. Agrawal et al. [17] demonstrated few-shot clinical information extraction with large language models, illustrating the feasibility of structured extraction from sensitive records with compact prompting strategies. Federated learning approaches [18] offer a pathway toward inter-facility model improvement without raw-data sharing. Moor et al. [19] discuss foundation models for generalist medical AI, illustrating both the potential and sensitivity of LLM-based decision support in high-stakes domains. Human-AI interaction guidelines emphasise that AI systems used in high-stakes workflows should support user control, transparency, appropriate reliance, and recoverability from errors [2]. Kwon et al. [20] introduced PagedAttention within vLLM, enabling near-linear throughput scaling with concurrent inference requests — a principal enabler of the parallel five-criterion evaluation design.

D. FAULT-TOLERANT DISTRIBUTED AI PIPELINES

Bass et al. [21] formalise system resilience as the capacity to maintain integrity under partial failure, furnishing the probabilistic basis for the retry-logic modelling presented in Section 3.5. Liu et al. [22] discuss trustworthy AI from a computational perspective, including reliability, robustness, safety, privacy, and accountability concerns that are directly relevant to automated decision-support pipelines. Recent surveys on LLM inference serving [23] describe batching, scheduling, memory management, and prefill/decode optimisation as key mechanisms for improving throughput in production LLM systems. Kreps et al. [24] describe distributed messaging architectures for scalable log and task processing, supporting the use of broker-based orchestration in fault-tolerant pipelines. Avizienis et al. [25] define dependability in terms of reliability, availability, safety, integrity, and maintainability, providing the conceptual foundation for the fault-tolerance model adopted in this work. Pinto et al. [26] survey machine-learning-based intrusion-detection systems for

critical infrastructure, providing adjacent evidence on dependable automated analysis under operational constraints.

E. FOUNDATIONAL RESEARCH ON INFORMATION SUFFICIENCY AND CRITICAL-INFRASTRUCTURE RESILIENCE

The present work is informed by prior research on software-quality assessment, fault-tolerant multiprocessor modelling, and explainable LLM architectures. Pomorova and Hovorushchenko [7] investigated the use of artificial neural networks for software quality evaluation and prediction, supporting the broader principle that structured input characteristics can influence the reliability of downstream assessment outputs. The present system operationalises this idea through an information-sufficiency pre-processing stage that verifies whether candidate records contain the required information elements before LLM evaluation. Romankevich et al. [27] developed graph-logical models for simulating the failure behaviour of non-basic fault-tolerant multiprocessor systems by combining auxiliary models for different operating conditions, demonstrating an explicit reliability-modelling approach for complex system states. Breja [28] presented an explainable Why-QA prototype that integrates a domain-enriched LLM with a knowledge graph, a causal-inference engine, and an explainability layer to produce detailed explanations, providing a cross-domain example of structured explanatory generation.

III. MATERIALS AND METHODS

A. SYSTEM ARCHITECTURE OVERVIEW

The proposed system, hereafter referred to as Talent Finder, is structured as a four-tier processing architecture (T1) input ingestion and validation; (T2) asynchronous task orchestration; (T3) local LLM inference and structured parsing; (T4) result aggregation and delivery. Tier boundaries are enforced through message-passing interfaces with clearly defined failure-handling semantics, such that faults are contained within their immediate processing boundary rather than propagating to downstream components. The full infrastructure is containerised via Docker Compose, encompassing PostgreSQL, Redis, Django (web layer), Celery workers, vLLM (OpenAI-compatible endpoint), and MailHog. This containerised topology supports deployment in air-gapped environments with no external network dependencies, supporting the data-sovereignty requirement established in Section 1.

Input data are accepted in Excel (.xlsx/.xls) or CSV formats — the standard data exchange formats for enterprise applicant-tracking systems. The parser normalises encodings (UTF-8 with cp1251 fallback), converts date cells to ISO-8601 strings, and normalises data types to ensure schema consistency. Each input record is represented as a structured key-value object, whose first field is a mandatory record identifier; the remaining keys correspond to questionnaire fields. The parser requires that identifier in the first column and raises a validated exception with a user-visible diagnostic if the requirement is violated. In Experiment B, this field contained only a random, study-specific record ID with no identity or applicant-tracking-system crosswalk (Section III.G).

B. Multi-Criteria Decision Analysis (MCDA) Scoring Model

The evaluation model is specified as a weighted additive value function over five domain criteria $C = \{c_1, c_2, c_3, c_4, c_5\}$, aligned with industrial safety staffing standards for high-hazard environments [1]. Each criterion c_i is assigned a weight w_i satisfying the normalisation constraint:

$$\Sigma_i w_i = 1, \quad w_i > 0, \quad i \in \{1, 2, 3, 4, 5\}. \tag{1}$$

For each criterion, the LLM evaluator produces an integer score s_i defined on a five-point Likert scale:

$$s_i \in \{1, 2, 3, 4, 5\} \tag{2}$$

where s = 1 denotes minimal compliance and s = 5 denotes exceptional compliance exceeding baseline requirements. The aggregated candidate suitability score S is computed as the weighted sum:

$$S = \Sigma_i w_i \cdot s_i, \quad S \in [1, 5] \tag{3}$$

The monotonicity property $\partial S/\partial s_i = w_i > 0$ for all i ensures that any improvement in a single criterion score strictly increases S, supporting consistent ordinality for candidate ranking. Table 1 specifies the five criteria, assigned weights, and regulatory rationales. Educational compliance ($w_1 = 0.35$) represents a statutory licensing prerequisite; industry experience ($w_2 = 0.30$) validates practical domain competency; technical certifications ($w_3 = 0.20$) confirm formal credential compliance; safety readiness ($w_4 = 0.10$) operationalises the safety culture of high-hazard industrial operations; regulatory fit ($w_5 = 0.05$) provides necessary but not sufficient motivational confirmation. The five criteria were informed by the personnel-qualification considerations in IAEA SSG-75 [1]. The numerical weights 0.35 / 0.30 / 0.20 / 0.10 / 0.05 are domain-informed heuristic assignments reflecting the intended relative importance of the criteria; they are not prescribed by SSG-75 and have not been psychometrically calibrated.

Table 1. MCDA Criteria and Weight Justification

Criterion ID	Domain Description	Weight (w_i)	Rationale
education_compliance	Educational qualifications vs. vacancy requirements	0.35	Mandatory licensure prerequisite
industry_experience	Sector-specific and role-relevant professional experience	0.30	Domain competency verification
qualification_certifications	Technical permits, certificates, regulatory approvals	0.20	Formal credential compliance
safety_readiness	Safety culture, responsibility, behavioural competencies	0.10	Safety culture alignment
overall_regulatory_fit	Motivational and regulatory alignment	0.05	Baseline eligibility

C. INFORMATION SUFFICIENCY PRE-PROCESSING

Prior to dispatching an evaluation task to the LLM, the system applies a structural completeness check inspired by prior work on neural-network-based software quality evaluation and prediction [7]. The procedure verifies that each candidate record contains non-null values for the field subsets required by each criterion c_i. Formally, let $F_i \subseteq F$ denote the required field subset for criterion i, let c denote a candidate record and

let v(f) indicate whether field f is populated. The sufficiency predicate for criterion i is:

$$Suff_i(c) = \Lambda_{f \in F_i} (v(f) = 1), \quad (4)$$

when $Suff_i(c)$ evaluates to 0, the LLM prompt for criterion c_i is augmented with an explicit constraint: the evaluator is instructed to confine its reasoning strictly to documented attributes and to assign a conservative score $s_i = 2$ (marginal) instead of inferring unsupported qualifications. This strategy reflects the broader principle that input completeness and structure affect the reliability of automated evaluation outputs [7].

D. PROMPT ENGINEERING ARCHITECTURE

D1. BEHAVIOURAL ANCHORED RATING SCALES (BARS)

Each criterion prompt incorporates a complete BARS rubric specifying concrete behavioural anchors for all five integer scores. The anchors translate abstract constructs such as "safety awareness" into observable, verifiable attributes derived from candidate questionnaire fields. Table 2 illustrates the BARS structure for the safety-readiness domain (criterion c_4).

Table 2. Illustrative BARS Rubric — Safety Readiness Criterion (c_4)

Score	Behavioural Anchor — Process Safety Domain (Illustrative)
1 — Unacceptable	Fails to identify standard process-safety protocol violations in simulated exercises; cannot locate emergency response mechanisms.
2 — Marginal	Recognises protocol deviations only when explicitly prompted; requires supervisor intervention for standard safety checks.
3 — Fully Successful	Consistently monitors safety-critical parameters and reports deviations within the prescribed protocol window.
4 — Exceeds Standard	Independently cross-validates sensor readings against control-room logs; provides guidance to peers on safety procedures.
5 — Exceptional	Proactively identifies a cascading process anomaly before primary alarm activation; initiates containment protocol.

By providing these anchors to the Llama 3 8B agent, the system maps underspecified trait descriptions into concrete observation targets, helping to reduce inter-evaluation variance — consistent with the original development of behaviourally anchored rating scales [29].

D2. STRUCTURED RATIONALE ENFORCEMENT

All criterion prompts are constrained to produce a two-field JSON output schema via an explicit structural directive in the system prompt. The schema is specified as: { "reasoning": "<free-form deliberation>", "score": <integer in [1,5]> }. The prompt design draws on the step-by-step reasoning paradigm introduced by Wei et al. [12], but constrains it into a structured rationale format suitable for audit. A low sampling temperature ($\tau = 0.2$) helps constrain output variance, thereby reducing the likelihood of score variability across semantically equivalent prompts. Reynolds and McDonell [13] demonstrate that carefully structured prompts can significantly influence model behaviour and improve output consistency, supporting the use of rubric-based prompt design in professional-domain tasks. The reasoning field follows a structured rationale format: CLAIM, EVIDENCE, and VERDICT. CLAIM states the evaluated candidate attribute, EVIDENCE cites the source profile fragment or explicitly records missing evidence, and VERDICT maps the evidence to the corresponding BARS anchor and numerical score.

E. ASYNCHRONOUS PIPELINE AND FAULT-TOLERANT ORCHESTRATION

E1. TASK QUEUE ARCHITECTURE

Upon receipt of a valid candidate batch, the web layer (Django) creates one Celery task per candidate and enqueues each task in Redis. A separate aggregation task polls the completion states of all candidate tasks at fixed intervals (5 seconds in the current implementation), with a maximum ceiling of 30 minutes, before reading committed results from PostgreSQL and dispatching the summary email. Celery workers consume tasks from the Redis queue with automatic acknowledgement only upon successful task completion. Failed tasks are subject to exponential-backoff retry up to k attempts. The probability of task-level availability under this scheme is formally approximated using an independent-trial model. Let p denote the probability of successful single-attempt inference. Then the probability of failure after k independent retries is:

$$P(failure \mid k) = (1 - p)^k \quad (5)$$

and the effective task availability is:

$$SA(k) = 1 - (1 - p)^k. \quad (6)$$

For the analytical illustration, a single-attempt success probability $p = 0.95$ is assumed rather than empirically estimated. With $k = 3$, $SA(3) = 1 - 0.05^3 = 0.999875$. This is an illustrative task-completion probability under independent and identically distributed attempts; it is not a measured service availability, a batch-level reliability estimate, or a lower bound under correlated hardware, broker, or power failures.

E2. PARALLEL CRITERION INFERENCE

The five LLM calls required per candidate are issued concurrently via asyncio.gather with a single AsyncOpenAI client pointed at the local vLLM endpoint. If τ_i denotes the inference latency for criterion c_i, sequential execution would incur total latency $T_{seq} = \Sigma_i \tau_i$. Parallel execution reduces this to:

$$T_{par} = max\{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}. \quad (7)$$

Under the assumption of approximately equal per-criterion latency ($\tau_i \approx \tau$ for all i), the speedup factor is $S_par \approx 5$. This is consistent with the continuous-batching throughput model of vLLM PagedAttention [20]: the five concurrent requests from a single candidate are dynamically batched at the GPU scheduler level, maximising arithmetic intensity without application-level batch management. For a processing load of N candidates, the aggregate wall-clock time saving is approximately:

$$\Delta T = N \cdot 4\tau \ \ (savings\ versus\ sequential\ processing). \quad (8)$$

For $N = 500$ candidates and $\tau = 3$ s per criterion, $\Delta T \approx 100$ minutes — operationally significant for mass-recruitment scenarios. This aligns with LLM serving research that treats batching, scheduling, and memory-efficient execution as central mechanisms for improving inference throughput [23], [20].

E3. GRACEFUL DEGRADATION STRATEGY

The system implements a five-stage deterministic fallback cascade when structured output fails validation: (1) Markdown-fence stripping; (2) full JSON parsing and Pydantic schema validation; (3) enforcement of $s_i \in \{1,...,5\}$; (4) regex partial recovery of a malformed "score": N field; and (5) assignment of $s_i = 3$ after complete parse failure. The final value is a continuity fallback, not a criterion-neutral estimate. Relative to a true extreme score, its maximum contribution error is $2w_i$; for education ($w_1 = 0.35$), this is 0.70 aggregate-score points and can cross a recommendation threshold. Any result containing a fallback score must therefore be flagged for human review.

Table 3. Fault-Tolerance Matrix — Failure Modes and Recovery Strategies

Failure Mode	Detection Mechanism	Recovery Strategy	Max Impact
Malformed JSON output	Pydantic schema validation	Regex partial-extract → fallback score = 3	Up to $2w_i$ in aggregate score (0.70 for education); may cross a category threshold
LLM inference timeout	AsyncIO timeout guard (30 s)	Celery task retry (max k = 3)	~0.013% task loss
Worker node crash	Redis heartbeat / dead-letter	Message re-queued to healthy worker	Zero data loss
Full GPU OOM event	vLLM PagedAttention OOM signal	Task suspended; re-queued after cooldown	Latency spike only
Aggregation timeout	30-minute ceiling polling loop	Partial results committed; email flagged	Partial report

F. PRIVACY AND INFERENCE OPTIMISATION

The evaluated configuration used a locally deployed Llama 3 8B Instruct model with AWQ weight-only quantisation, served through the vLLM runtime. PagedAttention was used for memory management and concurrent request serving [20]. The retained experimental record identifies the model family, quantisation method, serving runtime, and hardware platform: Experiment B was conducted on a MacBook Pro with an Apple M4 Max SoC, a 32-core integrated GPU, and 36 GB of unified memory. The record does not preserve the exact quantised checkpoint identifier, quantisation bit width, vLLM version, or execution-backend configuration. This missing deployment metadata limits exact computational replication and is acknowledged as a study limitation; no quantisation-specific performance claim is therefore made. Appendix A reproduces the complete system message, common prompt template, criterion-specific instructions, and BARS rubrics for all five criteria. Let l_prefix denote the shared-prefix length and l_total the full prompt length; their ratio is defined in Equation (9).

$$\eta_{cache} = l_{prefix} / l_{total}. \qquad (9)$$

With a system message of approximately 300 tokens and a vacancy-requirements block of approximately 500 tokens in an approximately 1,200-token prompt, the estimated shared-prefix fraction is η_cache ≈ 0.67. This value describes the potentially reusable share of prefill tokens; it does not imply a two-thirds reduction in end-to-end inference cost because decoding tokens, cache misses, scheduling, and other overheads remain. No Experiment B research record traversed an external network interface. This architecture reduces exposure to privacy risks identified for LLM-based systems [16] and supports human oversight [2] and AI-system quality considerations [8].

G. ETHICS, DATA PROVENANCE, AND DATA PROTECTION

Experiment B was a retrospective observational evaluation of method behaviour on an anonymised research dataset derived from historical questionnaires. The original identifiable questionnaires had been collected and processed by the industrial partner for its own recruitment activities before, and separately from, this study. Before any study evaluation, direct identifiers were removed and each research record and associated audit output was assigned a random, study-specific ID. No lookup table or other crosswalk linking those IDs to identities or to an applicant-tracking system is held or available to either the authors or the partner; neither party can identify an individual from the research records, whether considered alone or together with information reasonably available to it. The anonymised dataset remained within locally isolated infrastructure and was accessible only to authorised study personnel. Experiment B did not form part of a recruitment decision and did not identify, shortlist, reject, or select any individual. Its outputs were retained for methodological analysis and are proposed to the partner only for consideration in a possible future evaluation. No formal academic ethics-committee approval or study-specific data-protection impact assessment was obtained. The article does not characterise or certify the legal basis for the partner's earlier processing of identifiable recruitment records. These statements do not constitute a certification of legal compliance.

IV. RESULTS AND DISCUSSION

A. EVALUATION STRATEGY AND EXPERIMENTAL DESIGN

The evaluation of the proposed pipeline proceeds across two complementary experiments. Experiment A is a controlled cross-model comparison on a 50-record benchmark sample, assessing the degree to which the local zero-egress configuration (Llama 3 8B Instruct via vLLM) approximates the scoring and ranking behaviour of GPT-4o-mini under identical prompt, BARS rubric, and MCDA weight conditions. Experiment B is a retrospective observational study in which the pipeline was applied to an anonymised research dataset derived from 1,534 historical questionnaire records for a skilled role in a regulated high-hazard setting, followed by a first-author source-grounding review of ten cases drawn from the two highest model-output bands. Experiment B characterises method behaviour under dataset-scale processing and does not constitute deployment, recruitment use, or validation against independent ground truth.

B. EXPERIMENT A: CROSS-MODEL COMPARISON ON BENCHMARK SAMPLE

B1. SCORE DISTRIBUTION

Table 4 presents the descriptive statistics of aggregate scores S in [1, 5] produced by each configuration across the 50-profile

benchmark. The local model produces a marginally higher mean score (3.048 vs. 2.925, delta = 0.123), indicating a slightly more permissive overall tendency. More substantively, it exhibits considerably lower score dispersion (std = 0.869 vs. 1.281), confirming that GPT-4o-mini employs a wider effective scoring range and discriminates more aggressively between candidates. The local model is correspondingly compressed toward the mid-scale region, assigning more candidates to the intermediate Review band.

Table 4. Descriptive Statistics - Local Llama 3 8B vs. GPT-4o-mini (n = 50)

Statistic	Local Llama 3 8B	GPT-4o-mini	Difference
Mean	*3.048*	*2.925*	*+0.123*
Std. Deviation	*0.869*	*1.281*	*-0.412*
Median	*3.40*	*2.80*	*+0.60*
Mean Absolute Difference (MAD)	—	—	*0.659*
Agreement within +/-0.5 pts	—	—	*42%*
Agreement within +/-1.0 pt	—	—	*78%*

The mean absolute difference of 0.659 points is non-trivial but operationally bounded. Practical agreement thresholds contextualise this figure: 78% of candidates receive scores within one point of one another across configurations, indicating that broad evaluative alignment is substantially stronger than exact agreement. The 42% rate of near-exact agreement (within +/-0.5 points) confirms that fine-grained numerical equivalence should not be assumed.

B2. SCORE-LEVEL AND RANKING-LEVEL CORRELATION

Table 5 presents the correlation statistics between local and GPT-4o-mini aggregate scores. The Pearson correlation r = 0.791 ($p < 0.001$) indicates strong linear score-level agreement. The Spearman rank correlation rho = 0.745 ($p < 0.001$) corroborates this at the ordinal level: the local model largely preserves the candidate ordering produced by GPT-4o-mini, which is of primary importance for pre-screening and shortlisting workflows where ranking consistency outweighs exact score equality.

Table 5. Score-Level and Ranking-Level Agreement Statistics

Metric	Value	Interpretation
Pearson r (aggregate score)	*0.791*	*Strong linear agreement*
Spearman rho (rank order)	*0.745*	*Strong-to-moderate rank consistency*
Recommendation agreement	*52% (26/50)*	*Moderate categorical alignment*

Recommendation-level agreement is markedly lower at 52%. This disparity between strong continuous correlation and moderate categorical agreement is attributable to threshold sensitivity: when aggregate scores are discretised into three recommendation categories, small numerical differences near boundary values produce categorical label divergence even when the underlying scores are close [6]. This is a system calibration issue, not a structural model failure.

B3. RECOMMENDATION DISTRIBUTION AND POLARISATION

GPT-4o-mini exhibits a markedly more polarised recommendation pattern (28% Recommended, 54% Not Recommended, 18% Review), whereas the local model channels 54% of candidates into the Review category (10% Recommended, 36% Not Recommended). This reflects the score-compression effect noted above. In the context of safety-critical industrial HR screening, a higher Review rate may be epistemically appropriate rather than a deficiency: a system that defers more borderline cases to human reviewers is less likely to automate consequential mis-categorisations, consistent with the human-in-the-loop deployment model advocated for regulated environments [2].

B4. CRITERION-LEVEL DIVERGENCE ANALYSIS

Table 6 disaggregates agreement to the level of individual MCDA criteria, revealing substantial heterogeneity. Safety readiness (c4) achieves the highest Pearson correlation (r = 0.903), indicating that both models identify safety-related evidence in a highly consistent manner -- an important positive finding given the primacy of safety culture in high-hazard industrial environments. Industry experience (c2) also demonstrates good agreement (r = 0.710, delta = 0.06).

Table 6. Criterion-Level Score Comparison (n = 50)

Criterion	Local Mean	GPT Mean	Delta	Pearson r	Level
Education compliance	*2.96*	*3.36*	*-0.40*	*0.577*	*Moderate*
Industry experience	*3.08*	*3.02*	*+0.06*	*0.710*	*Good*
Qualifications / certs.	*3.62*	*2.46*	*+1.16*	*0.106*	*Very weak*
Safety readiness	*2.14*	*2.38*	*-0.24*	*0.903*	*Very strong*
Regulatory fit	*3.00*	*2.26*	*+0.74*	*0.607*	*Moderate*

The qualifications and certifications criterion (c3) is the primary source of inter-model divergence: the local model scores this criterion substantially higher (mean 3.62 vs. 2.46, delta = +1.16) with near-absent correlation (r = 0.106). This suggests the two models interpret the construct of "qualification" in fundamentally different ways: the local model appears to treat broad technical skills or adjacent training as qualifying evidence, whereas GPT-4o-mini applies a stricter standard requiring explicit formal certification. This divergence represents the most critical methodological finding of Experiment A and motivates the rubric recalibration discussed in Section 4.4.

Both configurations demonstrate strong consensus on extreme candidates. The five highest-scoring candidates receive consistently high averaged scores (4.275 to 4.475), whilst the lowest-ranking candidates receive scores of 1.20 (local) and 1.00 (GPT) - confirming that the local model reliably identifies both well-qualified and clearly unsuitable profiles. The largest individual disagreement (2.05 points) occurred on an ambiguous profile where evidential uncertainty was resolved in opposing directions; such cases are designated as mandatory human-review items.

C. EXPERIMENT B: RETROSPECTIVE EVALUATION ON AN ANONYMISED HISTORICAL QUESTIONNAIRE DATASET

C1. DATASET AND EVALUATION CONTEXT

Experiment B evaluated the pipeline retrospectively on an anonymised research dataset derived from 1,534 historical questionnaire records concerning a skilled role in a regulated high-hazard setting. Direct identifiers had been removed before evaluation. Records and audit outputs were keyed only by random, study-specific IDs; no lookup table or crosswalk to identities or to an applicant-tracking system was available to either the authors or the partner. All processing was completed in a locally isolated zero-egress configuration without external API calls. The pipeline generated scores, model-output categories, and per-criterion rationales for methodological analysis. The category names were experimental output labels only: no individual was identified, shortlisted, rejected, or selected on their basis.

C2. DATASET-WIDE OUTPUT DISTRIBUTION

Table 7 presents the dataset-wide model-output distribution.

Table 7. Dataset-Wide Output Distribution (n = 1,534)

Category	Count	Share	Score Threshold
Recommended (model-output label)	15	1.0%	S >= 4.00
Review (model-output label)	152	9.9%	3.00 <= S < 4.00
Not Recommended (model-output label)	1,367	89.1%	S < 3.00
Dataset mean score	2.35 / 5.00	—	—
Primary low-score driver	Absence of relevant-sector experience (~96% of low-band outputs)	—	—

The dataset mean of 2.35 and the concentration of 89.1% of records in the lowest output band reflect the specificity of the encoded role requirements rather than an observed recruitment outcome. Approximately 77% of the anonymised records lacked documented experience in the target high-hazard sector or an adjacent regulated setting, and a further 19% contained no verifiable trade-level qualification relevant to the target role. Applying the two upper output bands retrospectively would have surfaced 167 of 1,534 records for possible further review. The corresponding 89.1% reduction is therefore a computed reduction in a notional first-pass review set, not an observed reduction in recruiter workload and not evidence that any recruitment action occurred.

C3. FIRST-AUTHOR MANUAL REVIEW OF TEN SAMPLED CASES

To inspect the source grounding of the experimental outputs, the first author manually reviewed ten anonymised records: five from the Recommended output band and five from the Review output band. The review was not blinded, and the model scores and rationales were visible. Record text was compared with each rationale to determine whether the cited evidence was present and whether the score appeared source-supported, over-estimated, under-estimated, or unsupported. This convenience review is a preliminary error-analysis exercise, not an accuracy study or independent expert validation. Table 8 summarises the outcomes.

Table 8. First-Author Manual Review Results -- Ten Sampled Cases

Case	Category	LLM Score	Source-Grounding Outcome	First-Author Assessment
C-01	Recommended	4.40	Source-supported	4.40 -- supported by source text
C-02	Recommended	4.15	Source-supported	4.15 -- supported by source text
C-03	Recommended	4.05	Source-supported	4.05 -- supported by source text
C-04	Recommended	4.25	Partially source-supported	~3.2 -- operations track != maintenance role
C-05	Recommended	4.35	Critical error: hallucination	1.5-2.0 -- fabricated evidence citation
C-06	Review	3.35	Source-supported	3.35 -- supported by source text
C-07	Review	3.40	Source-supported	3.40 -- supported by source text
C-08	Review	3.40	Source-supported	3.40 -- supported by source text
C-09	Review	3.35	Under-scored	3.8-4.0 -- 15+ yrs evidence missed
C-10	Review	3.40	Over-scored	2.5-2.8 -- degree without sector practice

Of the ten cases, six were assessed as source-supported (60%), three exhibited apparent scoring biases (30%), and one contained an unsupported factual claim (10% of the ten inspected cases). These are descriptive observations from a small, deliberately selected, non-blinded sample and must not be interpreted as accuracy estimates or dataset-level error rates. A larger, prospectively sampled and blinded assessment by multiple domain-qualified reviewers, including inter-rater agreement analysis, is required for further validation.

C4. Characterisation of Observed Error Modes

The verification exercise identified four distinct error modes, each with different implications for method development and possible future use.

Evidence hallucination (critical severity). Case C-05 exhibited the most severe failure mode: the model cited specific employment history absent from the anonymised record, placing that record in the highest output band without supporting evidence. This was one of the five inspected Recommended-band outputs (1/5, 20%) and one of all ten inspected outputs (1/10, 10%). These are descriptive proportions for the selected review sample, not estimates for the full research dataset. Automated evidence-grounding and human verification would be required in any future evaluation. Role-track conflation (moderate severity). Case C-04 illustrated an inability to distinguish the operations track (process control and shift management) from the maintenance track (instrument repair and trade certification). These tracks may share facility labels but represent distinct competency domains. Case C-08 was assessed as source-supported in the

first-author review and is therefore not counted in this error mode. The C-04 error motivates an explicit role-type classification step before scoring.

Evidence under-citation (moderate severity). Case C-09 demonstrated the converse of hallucination: a candidate with over 15 years of documented facility employment received a low experience score because the model concluded the affiliation was "unclear", despite the facility abbreviation appearing explicitly in the profile text. This under-citation error results from the model's conservative anchoring strategy interacting poorly with non-standard abbreviation formats in free-text profiles. It is addressable through a facility-name normalisation step in the pre-processing pipeline.

Academic credential over-weighting (low severity). Case C-10 received a Review-band score despite possessing no documented practical experience in the relevant domain. The model awarded partial scores on the qualifications criterion based on academic degree subject alignment, without penalising the complete absence of trade certification. This is consistent with the qualifications criterion divergence identified in Experiment A (r = 0.106) and reinforces the finding that the qualifications BARS rubric requires explicit evidence-type anchors distinguishing formal trade certification from adjacent academic credentials.

D. Synthesis and Prospective Applicability Assessment

The two experiments characterise the method's behaviour and limitations without establishing operational effectiveness. Experiment A measures cross-model agreement, while Experiment B shows how the pipeline behaves on a dataset-scale collection of anonymised historical records, including its output distribution, auditability, and observable failure modes.

The findings support further independent evaluation of the method for possible future human-supervised use. They do not establish present suitability for candidate shortlisting, prioritisation, or selection. Any future operational assessment would require prospectively defined procedures, independent domain review, ground-truth annotation, and safeguards against unsupported evidence, role-track conflation, and over-reliance on weak credentials.

Table 9 summarises the prospective applicability assessment across the principal dimensions.

Table 9. Prospective Applicability Assessment -- Summary Across Both Experiments

Evaluation Dimension	Key Finding	Prospective Implication
Candidate ranking	Spearman rho = 0.745 (Exp. A)	Supports further evaluation; operational use not validated
Score agreement	78% within +/-1.0 pt (Exp. A)	Cross-model agreement only; not an accuracy result
Safety readiness	r = 0.903; 3/3 source-supported in the sampled review	Promising signal; requires independent validation
Industry experience	r = 0.710; consistent (Exp. B)	Requires validation; inspect non-standard record formats
Qualifications criterion	r = 0.106; recurrent over-scoring	Rubric recalibration required
Evidence hallucination	1/10 overall; 1/5 Recommended cases inspected	Human verification required in any future evaluation
Role-track conflation	1/10 cases in Exp. B (moderate)	Add role-type classifier to prompt pipeline
Notional workload concentration	167/1,534 records in the two upper output bands	Scenario estimate; no observed recruiter-workload reduction
Data sovereignty	Zero egress; no external API calls	Architecturally aligned with data-sovereignty and AI-system quality considerations

The retrospective output distribution demonstrates potential workload concentration rather than an observed operational saving. Under the experimental band thresholds, 167 of 1,534 anonymised records fell within the two upper output bands, corresponding to an 89.1% reduction in a notional first-pass review set. No recruiter workload was measured, and no individual was identified or subjected to a recruitment action. The criterion-level and sampled source-grounding results provide behavioural evidence for further validation, not evidence of screening accuracy.

Four findings constrain any prospective use: the unsupported evidence citation, weak cross-model agreement for the qualifications criterion, observed role-track conflation, and the small non-blinded first-author review. The study therefore supports further methodological validation only; it does not establish suitability for operational triage or hiring decisions. Future validation should use a larger prospectively sampled set, multiple blinded domain reviewers, inter-rater agreement, ground-truth annotation, and psychometric calibration of the MCDA weights.

V. CONCLUSIONS

This paper has proposed and empirically evaluated a methodology for building fault-tolerant, privacy-preserving, and explainable automated candidate evaluation pipelines for possible future use in regulated high-hazard settings. The methodology was studied through two experiments: a controlled 50-record cross-model comparison against GPT-4o-mini and a retrospective observational method-performance evaluation on an anonymised dataset derived from 1,534 historical questionnaire records, followed by a first-author source-grounding review of ten sampled cases. The findings characterise the method's behaviour and limitations rather than its accuracy or operational effectiveness.

First, the five-criterion MCDA scoring model used criteria informed by IAEA SSG-75 personnel-qualification considerations [1], while the numerical weights were assigned using domain-informed heuristics rather than derived from the standard. It produced candidate rankings with strong cross-model consistency (Spearman rho = 0.745, $p < 0.001$ against GPT-4o-mini). The first-author case review provided preliminary error-analysis evidence but does not constitute independent expert validation. Psychometric calibration of the weight assignments remains future work.

Second, the prompt-engineering architecture combining Behaviourally Anchored Rating Scales with structured rationale generation produced criterion-level evaluations of demonstrably unequal quality. The safety readiness criterion, whose BARS rubric was the most precisely specified, achieved the highest cross-model correlation (r = 0.903), and all three

relevant sampled cases were assessed as source-supported. Conversely, the qualifications criterion, whose rubric lacked explicit evidence-type anchors, exhibited near-zero inter-model correlation ($r = 0.106$) and apparent over-scoring in retrospective Experiment B. This contrast suggests that rubric specificity may materially influence criterion-level behaviour; BARS anchor quality should therefore be treated as a first-class engineering concern.

Third, under the illustrative assumptions $p = 0.95$, independent attempts, and $k = 3$, the analytical single-task completion probability is $SA(3) = 1 - (1 - p)^3 = 0.999875$. This is not a measured availability claim and does not model correlated infrastructure failures or batch completion. The five-stage parse cascade maintains processing continuity, but its fallback score of 3 is not criterion-neutral and can change the recommendation; fallback-containing outputs must be flagged for human review.

Fourth, the locally deployed AWQ-quantised Llama 3 8B Instruct model served through vLLM operated without research-record egress in Experiment B. The estimated 67% shared-prefix fraction applies to potentially reusable prefill tokens and is not an observed two-thirds reduction in total inference cost. The exact quantised checkpoint identifier, quantisation bit width, vLLM version, and execution-backend configuration were not retained; consequently, no quantisation-specific performance claim is made.

Fifth, the first-author manual review identified four methodologically distinct failure modes: evidence hallucination (critical; 1/5 or 20% of the inspected Recommended-band outputs, equivalently 1/10 of all inspected cases), role-track conflation (moderate; Case C-04), evidence under-citation (moderate), and academic credential over-weighting (low). The sample design and lack of independent blinded review prevent dataset-level generalisation.

The primary observed methodological limitation is the unsupported evidence citation found in one of the five manually inspected Recommended-band outputs (20%). Across all ten inspected outputs, the same observation represents 10%. Neither proportion is a dataset-wide error-rate estimate. Additional limitations are the absence of an independent domain-review panel, the small and non-random manual sample, the absence of a ground-truth dataset annotated by domain specialists, and incomplete retained metadata for exact inference-stack replication.

The results show that privacy-preserving local processing, structured rationale generation, and fault-bounded orchestration can be evaluated together on a dataset-scale collection of anonymised historical records. Under the experimental output thresholds, 167 of 1,534 records fell within the two upper bands, implying an 89.1% reduction in a notional first-pass review set. This is a retrospective scenario estimate, not a measured reduction in recruiter workload or evidence of recruitment use. The method requires independent, prospectively designed validation before it can be considered for operational candidate screening or hiring support.

Future work is prioritised by the observed limitations. The first priority is an automated evidence-grounding check against the source profile. The second is a role-type classification step addressing the conflation observed in Case C-04. The third is recalibration of the qualifications BARS rubric to distinguish formal trade certification from adjacent academic credentials. A larger prospectively sampled, blinded, multi-reviewer study should then quantify inter-rater agreement, calibration, and population-level error rates. Federated learning may support multi-site improvement without raw-data transfer [18], while multi-sample perturbation may support uncertainty-aware thresholds [15].

References

[1] IAEA, "Recruitment, Qualification and Training of Personnel for Nuclear Power Plants," IAEA Safety Standards Series No. SSG-75, Vienna, 2022.

[2] S. Amershi et al., "Guidelines for Human-AI Interaction," *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'19)*, 2019, pp. 1-13. https://doi.org/10.1145/3290605.3300233.

[3] C. Gan, Q. Zhang, and T. Mori, "Application of LLM Agents in Recruitment: A Novel Framework for Automated Resume Screening," Journal of Information Processing, vol. 32, pp. 881–893, 2024. https://doi.org/10.2197/ipsjjip.32.881.

[4] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 469–481, 2020. https://doi.org/10.1145/3351095.3372828.

[5] Ministry of Digital Transformation of Ukraine, "White Paper on Artificial Intelligence Regulation in Ukraine: Vision of the Ministry of Digital Transformation of Ukraine," Version for Consultation, June 2024. [Online]. Available: https://storage.thedigital.gov.ua/files/c/fc/36c4cae89deedfbf3781ec6bceddffcc.pdf.

[6] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Transactions on Information Systems*, vol. 43, issue 2, article 42, 2025. https://doi.org/10.1145/3703155.

[7] O. Pomorova and T. Hovorushchenko, "Research of Artificial Neural Network's Component of Software Quality Evaluation and Prediction Method," in Proceedings of the 2011 IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), Prague, Czech Republic, vol. 2, pp. 959–962, 2011. https://doi.org/10.1109/IDAACS.2011.6072916.

[8] ISO/IEC 25059:2023, "Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems," ISO/IEC, 2023.

[9] A. G. Møller, A. Pera, J. A. Dalsgaard, and L. M. Aiello, "The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), pp. 179–192, 2024. https://doi.org/10.18653/v1/2024.eacl-short.17.

[10] D. Jiang et al., "LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 14165–14178, 2023. https://doi.org/10.18653/v1/2023.acl-long.792.

[11] Z. R. Tam et al., "Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models," arXiv preprint arXiv:2408.02442, 2024. DOI: 10.48550/arXiv.2408.02442.

[12] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint arXiv:2201.11903, 2022. DOI: 10.48550/arXiv.2201.11903.

[13] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," in Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '21), 2021. https://doi.org/10.1145/3411763.3451760.

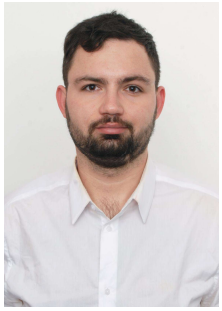
[14] S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 2023. https://doi.org/10.52202/075280-0517.

[15] L. Wang et al., "A Survey on Large Language Model based Autonomous Agents," Frontiers of Computer Science, vol. 18, no. 6, 2024. https://doi.org/10.1007/s11704-024-40231-1.

[16] M. Miranda et al., "Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions," arXiv preprint arXiv:2408.05212, 2024. DOI: 10.48550/arXiv.2408.05212.

[17] K. Agrawal et al., "Large Language Models are Few-Shot Clinical Information Extractors," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130.

[18] Q. Yang et al., “Federated Machine Learning: Concept and Applications,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019. https://doi.org/10.1145/3298981.
[19] M. Moor et al., “Foundation models for generalist medical artificial intelligence,” Nature, vol. 616, pp. 259–265, 2023. https://doi.org/10.1038/s41586-023-05881-4.
[20] W. Kwon et al., “Efficient Memory Management for Large Language Model Serving with PagedAttention,” in Proceedings of the 29th Symposium on Operating Systems Principles (SOSP ’23), 2023. https://doi.org/10.1145/3600006.3613165.
[21] L. Bass, P. Clements, and R. Kazman, Software Architecture in Practice, Addison-Wesley Professional, 2021.
[22] H. Liu et al., “Trustworthy AI: A Computational Perspective,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 14, pp. 1–59, 2023. https://doi.org/10.1145/3546872.
[23] B. Li, Y. Jiang, V. Gadepally, and D. Tiwari, "LLM Inference Serving: Survey of Recent Advances and Opportunities," in Proceedings of the 2024 IEEE High Performance Extreme Computing Conference (HPEC), 2024, pp. 1–8. https://doi.org/10.1109/HPEC62836.2024.10938426.
[24] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A Distributed Messaging System for Log Processing," in Proceedings of the 6th International Workshop on Networking and Databases (NetDB 2011), 2011.
[25] A. Avizienis et al., “Basic Concepts and Taxonomy of Dependable and Secure Computing,” IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11–33, 2004. https://doi.org/10.1109/TDSC.2004.2.
[26] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, "Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure," Sensors, vol. 23, no. 5, p. 2415, 2023. https://doi.org/10.3390/s23052415.
[27] V. Romankevich, K. Morozov, A. Romankevich, P. Malezhyk, and L. Zacharioudakis, "A Generalized Method for Constructing Graph-Logical Models of Non-Basic Fault-Tolerant Multiprocessor Systems via Model Combination," International Journal of Engineering and Manufacturing, vol. 16, no. 3, pp. 254–269, 2026. https://doi.org/10.5815/ijem.2026.03.15.
[28] M. Breja, "A Novel Explainable LLM-based Why-QA Framework for Climate Resilient and Sustainable Smart Agriculture," International Journal of Engineering and Manufacturing, vol. 16, no. 4, pp. 302–323, 2026. https://doi.org/10.5815/ijem.2026.04.21.
[29] P. C. Smith and L. M. Kendall, "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales," Journal of Applied Psychology, vol. 47, no. 2, pp. 149–155, 1963. https://doi.org/10.1037/h0047060.



VITALII PAVLIUK is a Senior Backend Engineer and a PhD student at the National University of Water and Environmental Engineering. He is focusing on the application of Artificial Intelligence and Large Language Models for personnel selection in critical infrastructure. His research interests include Multi-Criteria Decision Analysis, Behaviourally Anchored Rating Scales, and secure AI systems for the industrial sector.

VOLODYMYR DREVETSKYI *is a Doctor of Technical Sciences, professor, and vice-president of the Engineering Academy of Ukraine. He leads the Department of Automation, Electrical, and Computer-Integrated Technologies. Author of over 200 scientific works, his research focuses on developing methods and devices for automated control of liquid parameters. He has mentored candidates and masters of technical sciences.*

APPENDIX A. PROMPT SPECIFICATION AND BARS RUBRICS

Industry, organisational, facility, regulator, standard-body, and facility-specific degree/specialisation identifiers are replaced by bracketed placeholders to protect the industrial partner's identity. The prompt structure, decision rules, output schema, criterion instructions, and score anchors are otherwise reproduced. The placeholders {requirements_text} and {candidate_data} are populated at run time.

Common system and user prompt

SYSTEM MESSAGE (verbatim, organisational identifier redacted):

You are a certified psychometric assessor for personnel qualification at a [INDUSTRY NAME] facility ([ANONYMIZED FACILITY]). You apply Evidence-Based Assessment (ISO 10667) and [INDUSTRY STANDARD BODY] [INDUSTRY STANDARD] safety culture standards. Core rule: evaluate ONLY what is explicitly documented in the provided CV and vacancy text. When evidence is absent, score low — absence of evidence is not evidence of competence. Think step-by-step before scoring. Write the reasoning content in Ukrainian (українською); keep the CLAIM/EVIDENCE/VERDICT labels in English. The "reasoning" field MUST follow this exact structure — three labelled parts, no exceptions: "CLAIM: [one precise, falsifiable statement about the candidate on this criterion]. EVIDENCE: '[verbatim quote or exact fact copied from the input — if nothing relevant is present, write: NO EVIDENCE FOUND]'. VERDICT: [one sentence: map the evidence to the rubric level and state why it earns that score]." LANGUAGE: keep the labels CLAIM/EVIDENCE/VERDICT in English, but write all of the reasoning CONTENT in Ukrainian (українською мовою). The EVIDENCE quote stays verbatim in its original language. Recruiters at the [FACILITY NAME] read this field in Ukrainian. PROHIBITIONS — do NOT do any of the following: • Use phrases like "experienced", "knowledgeable", "safety-conscious" as evidence (these are marketing language, not verifiable facts). • Infer, assume, or extrapolate qualifications beyond what is explicitly stated. • Grant equivalence credit without a direct documentary statement. Your response MUST be exactly one JSON object with exactly two keys: "reasoning" (string, following the CLAIM/EVIDENCE/VERDICT structure above) and "score" (integer 1–5). No other text before or after the JSON. Example (labels in English, content in Ukrainian): {"reasoning": "CLAIM: Кандидат має ступінь магістра за відповідною спеціальністю. EVIDENCE: 'Освіта: [ANONYMIZED INSTITUTION], спеціальність [ANONYMIZED SPECIALISATION], магістр, 2018'. VERDICT: Рівень освіти та спеціальність прямо відповідають вимозі вакансії.", "score": 5}

USER TEMPLATE (verbatim):

{criterion_part}

{rubric}

- Vacancy / requirements:

{requirements_text}

MANDATORY INSTRUCTIONS — read before scoring:

1. GOLDEN THREAD: Your score must be traceable to a specific, verbatim fact from the candidate data below. Every claim in your reasoning must cite an exact quote. If no

relevant evidence exists, say so.

2. REASONING FORMAT — three labelled parts, in this order (labels in English, content in Ukrainian):

CLAIM: [one precise statement about the candidate on this criterion — Ukrainian]

EVIDENCE: "[exact quote or fact from the candidate data — copy verbatim; if none: NO EVIDENCE FOUND]"

VERDICT: [one sentence linking evidence → rubric level → score justification — Ukrainian]

3. INFERENCE CONTROL:

- Do NOT treat soft phrases ("experienced", "knowledgeable", "responsible") as evidence.
- Do NOT assume a qualification exists because the role typically requires it.
- Do NOT grant equivalence between fields or roles unless the document states it explicitly.
- Absence of evidence defaults to score 1, NOT score 3.

4. OUTPUT — STRICT SCHEMA (two keys, no others):

{{"reasoning": "CLAIM: ... EVIDENCE: '...' VERDICT: ...", "score": 4}}

Rules:

- "score" MUST be a separate integer field — do NOT only write the score inside the reasoning text.
- Do NOT include any other keys (no "mandatory_present", "sc_indicators", "analysis", etc.).
- The VERDICT line in reasoning must NOT contain "score N" — put the score only in the score field.

Now score the following candidate against the vacancy and rubric above.

- Candidate data:

{candidate_data}

Education compliance ($w_1 = 0.35$)

CRITERION-SPECIFIC INSTRUCTION:

You are a certified workforce qualification specialist for [INDUSTRY NAME] energy facilities (ISO 10667).

Task:

Assess whether the candidate's formal education meets the educational requirements stated in the vacancy.

Instructions:

1. DEGREE LEVEL: verify the level (молодший бакалавр / бакалавр / магістр / PhD).

A lower degree level than required is a hard disqualification — do not compensate with experience.

2. SPECIALIZATION CHECK — three tiers:

Tier A (facility-specific): [ANONYMIZED SPECIALISATION GROUP A] → automatically relevant for facility-specific roles.

Tier B (directly matching): [ANONYMIZED SPECIALISATION GROUP B] → strong match for the target role family.

Tier C (partially relevant): [ANONYMIZED SPECIALISATION GROUP C] → acceptable only if the vacancy explicitly lists the field as qualifying.

3. NO EQUIVALENCE RULE: do NOT treat one specialization as equivalent to another unless the
candidate's document explicitly states it (e.g., a diploma supplement noting equivalence,
or a Ukrainian MOH re-qualification certificate).

4. INSTITUTION: note whether the institution is Ukrainian state-accredited.
Foreign degrees require Ukrainian MOH nostrification — check if it is mentioned.

5. ADDITIONAL EDUCATION: postgraduate courses, professional development, or second degree may
add context but do NOT substitute for a missing primary qualification.

In your EVIDENCE: quote the exact degree name, field, institution, and year from the CV.
If education is not mentioned at all, write: NO EVIDENCE FOUND and score 1.

BARS RUBRIC:

BARS scale — Education Compliance (integer 1–5, base score on documented facts only):

1 = No relevant education, or degree level/field completely mismatches the requirement.
Example: secondary school only; or degree in unrelated humanities/social sciences.

2 = Partially relevant education: degree exists but specialization does not match.
Example: a degree in [ANONYMIZED PARTIALLY RELEVANT FIELD] where a directly matching field is required.
Subnote: a "related" field is NOT equivalent without an explicit document stating equivalence.

3 = Meets the stated minimum: correct degree level AND broadly relevant specialization.
Example: Bachelor's in [ANONYMIZED RELEVANT TECHNICAL FIELD].

4 = Full match: degree level AND specialization directly align with vacancy text.
Example: Bachelor's or Master's in [ANONYMIZED DIRECTLY MATCHING FIELD] from an accredited technical university.

5 = Exceeds requirement: [INDUSTRY NAME]-specific education OR advanced degree in matching field.
Qualifying specialisations: [ANONYMIZED FACILITY-SPECIFIC FIELDS], or a PhD/Master's in a fully matching technical discipline.

Industry experience (w_2 = 0.30)

CRITERION-SPECIFIC INSTRUCTION:

You are a senior workforce qualification expert for the [INDUSTRY NAME] and energy sector ([INDUSTRY STANDARD BODY] [INDUSTRY STANDARD]).

Task:

Assess whether the candidate's professional experience meets the industry and role requirements.

Instructions:

1. EXPERIENCE TIER — classify each position in the CV into the highest applicable tier:

Tier 1 — On-site [FACILITY NAME] operational: control room operator, I&C engineer directly on a unit,
shift supervisor, safety-critical protection systems technician. Employer must be an [FACILITY NAME]
(an operating [INDUSTRY NAME] facility or a unit of the national [INDUSTRY NAME] operator).

Tier 2 — [INDUSTRY NAME]-adjacent / industry supply chain: [INDUSTRY NAME] research institute or national-operator division, fuel
fabrication, regulated waste processing, decommissioning, [INDUSTRY NAME] regulatory authority ([ANONYMIZED REGULATORY AUTHORITY]).

Tier 3 — General energy / safety-critical non-[INDUSTRY NAME]: thermal power plant (ТЕС, ТЕЦ), hydro
(ГЕС), high-voltage substation (ПС 110 kV+), petrochemical I&C, aviation maintenance,
industrial automation in continuous-process industries.

Tier 4 — Unrelated: IT software, construction, retail/FMCG, general manufacturing
with no documented instrumentation or automation function.

2. DURATION: count only DOCUMENTED years (start date – end date explicitly stated in CV).
Do NOT estimate or infer duration from vague phrases like "extensive experience".

3. ROLE FUNCTION: verify that the candidate's stated responsibilities overlap with the vacancy
function (automation, I&C maintenance, PLC/SCADA, process control — not just "engineer").

4. HARD EVIDENCE RULE: the following phrases are NOT evidence of experience and must be ignored:
"vast experience", "deep knowledge of", "proficient in", "successfully implemented",
"expert-level skills". Only count: named employer + role title + dates + responsibilities.

5. GENERAL vs OPERATIONAL DISTINCTION (critical for [INDUSTRY NAME] roles):
"Worked in energy sector" ≠ "Worked at an [FACILITY NAME]".
"Familiar with [INDUSTRY NAME] regulations" ≠ "Operated under [INDUSTRY NAME] regulatory oversight".
Score Tier 1 only when the employer is verifiably an [FACILITY NAME] unit.

In your EVIDENCE: quote the exact employer name, job title, and dates from the CV.
If no relevant work history is present, write: NO EVIDENCE FOUND and score 1.

BARS RUBRIC:

BARS scale — Industry Experience (integer 1–5, verified work history only):

Experience tiers (highest applicable tier determines the base):

Tier 1 — Operational [FACILITY NAME] experience: directly on a safety-critical operational unit (control room, I&C maintenance,

shift supervisor, safety-critical unit operator). Highest relevance.

Tier 2 — [INDUSTRY NAME] industry supply chain / [INDUSTRY NAME]-adjacent: regulated-material fabrication, regulated waste, decommissioning,

[INDUSTRY NAME] research facilities, supporting divisions of the national [INDUSTRY NAME] operator.

Tier 3 — General energy / safety-critical non-[INDUSTRY NAME]: thermal power plants, hydro, substations,

aviation maintenance, industrial automation in chemical/oil-gas.

Tier 4 — Unrelated or unclear industry: IT, construction, retail, generic manufacturing

with no instrumentation/automation component.

Score:

1 = Tier 4 only, or no documented work history at all.

2 = Tier 4 with <2 years, or Tier 3 tangential with no automation/I&C role.

3 = Tier 3 (general energy/safety-critical), 2–5 years, role matches the vacancy function.

4 = Tier 2 OR Tier 3 with 5+ years in automation/instrumentation/I&C role.

5 = Tier 1 (operational [FACILITY NAME] experience), any duration. OR Tier 2 with 8+ years.

Qualifications and certifications (w_3 = 0.20)

CRITERION-SPECIFIC INSTRUCTION:

You are a compliance specialist verifying professional qualifications for regulated industry roles.

Task:

Assess whether the candidate holds the required formal qualifications, certifications, and permits

documented in the vacancy. Apply a binary check: either the document exists, or it does not.

Instructions:

1. MANDATORY vs OPTIONAL — classify each required qualification:

Mandatory (M): explicitly required in the vacancy text. Absence = disqualification or score ≤ 2.

Optional/Desirable (O): adds value but absence does not disqualify.

Score is primarily determined by mandatory qualification coverage.

2. [INDUSTRY NAME] / REGULATORY-SPECIFIC QUALIFICATIONS (highest value if present):

- Допуск до самостійної роботи — [FACILITY NAME]-issued work permit for the specific role
- Кваліфікаційна картка національного оператора — national-operator qualification card
- Дозвіл [ANONYMIZED REGULATORY AUTHORITY] — state [INDUSTRY NAME] regulatory permit
- Допуск до роботи з [ANONYMIZED REGULATED-SOURCE PERMIT] — permission to work with regulated hazardous sources
- [INDUSTRY STANDARD BODY] training certificate (TC programme)
- Посвідчення з ядерної та радіаційної безпеки

3. TECHNICAL QUALIFICATIONS (commonly required for automation/I&C roles):

- Група з електробезпеки: II (basic) / III / IV / V (highest — required for 1 kV+ work)
- PLC programming: Siemens TIA Portal / Step 7 / S7-1500, ABB AC500, Allen-Bradley
- SCADA/HMI: Siemens WinCC, Wonderware InTouch, GENESIS64, iFIX
- Industrial fieldbus: Profibus-DP, Profinet, Modbus RTU/TCP, HART, Foundation Fieldbus
- Метрологічна атестація — metrology attestation for measurement instruments
- ISO 9001 / ISO 45001 internal auditor certificate (if in vacancy)

4. VERIFICATION RULES:

- A certificate with an expiry date: check if it is still valid (compare to vacancy date).
- "Knows Siemens PLC" is NOT a certification — only count named training documents.
- Course completions listed on LinkedIn are NOT official certifications unless supported

by a named issuing institution and document number.

- Do NOT assume a qualification is held because the job normally requires it.

In your EVIDENCE (inside the reasoning text): quote the exact certification name, issuing body,

and year (if stated). Within the reasoning text, list separately which mandatory items are

present and which are absent. Do NOT add extra JSON fields — output only "reasoning" and "score".

If no qualifications are mentioned at all, write: NO EVIDENCE FOUND and score 1.

BARS RUBRIC:

BARS scale — Qualifications & Certifications (integer 1–5, documents only):

1 = All required qualifications missing; no relevant certifications documented.

2 = Some certifications present but key mandatory ones are absent.

Example: has general electrical safety group (I–II) but no PLC/SCADA-specific training.

3 = Core mandatory qualifications documented; optional or desirable ones partially met.

Example: electrical safety group III+, basic PLC training (Siemens/ABB), no SCADA cert.

4 = All mandatory qualifications documented AND most optional ones present.

Example: electrical safety group IV–V, TIA Portal / Step 7, WinCC / InTouch,

industrial networks (Profibus, Modbus), internal [FACILITY NAME] or enterprise training certificates.

5 = Complete qualification set including [INDUSTRY NAME]/regulatory-specific credentials.

Qualifying additions: [ANONYMIZED REGULATORY AUTHORITY]-recognised training, [INDUSTRY STANDARD BODY] training certificate,

national-operator qualification card, regulated-source safety permit (допуск до роботи з [ANONYMIZED REGULATED-SOURCE PERMIT]),
or vendor certification (Siemens, ABB, Emerson) in a relevant product line.

Safety readiness (w_4 = 0.10)

CRITERION-SPECIFIC INSTRUCTION:

You are a safety culture assessor applying [INDUSTRY PEER ORGANISATION] safety-culture standards and [INDUSTRY STANDARD BODY] [INDUSTRY STANDARD] criteria.

Task:

Assess whether the candidate's documented background demonstrates readiness to work in a
safety-critical, high-consequence environment ([INDUSTRY NAME] facility).

Instructions:

1. SEARCH FOR [INDUSTRY PEER ORGANISATION] SAFETY CULTURE INDICATORS ([INDUSTRY STANDARD]):

For each indicator below, note whether a CONCRETE, VERIFIABLE instance appears in the CV.

Generic claims do not count. Required: specific event, role, date, or named procedure.

SC-1 Industry Safety Priority

Look for: explicit prioritisation of safety over production or schedule in a documented context.

SC-2 Conservative Decision-Making / Questioning Attitude

Look for: documented escalation of a concern; "stop-and-think" decision; raising a
safety issue to management; refusal to proceed without authorisation.

SC-3 Procedure Adherence

Look for: references to working strictly per written procedures (НД, РЕ, ТІ, ПБЯ);
participation in procedure review or update; deviation reporting (відхилення від НД).

SC-4 Stop-Work Authority

Look for: any mention of stopping work, delaying an operation, or challenging an
instruction due to a safety concern.

SC-5 Cross-checking / Independent Verification

Look for: described habit of verifying own or others' work before an irreversible action;
participation in two-person rule procedures.

SC-6 Learning Organisation

Look for: post-incident analysis participation; application of lessons learned;
involvement in corrective action programmes (КАП / CAP).

SC-7 Personal Accountability

Look for: ownership statements with outcomes; self-reporting of near-misses or errors;
safety-related recognition or disciplinary-free record explicitly noted.

2. INFERENCE CONTROL — PROHIBITED PHRASES (score these as ZERO evidence, do not cite them):

"safety-conscious", "responsible worker", "attentive to safety", "safety first",
"committed to safety culture", "follows all rules", "careful and precise",
"understands the importance of safety". These are self-reported adjectives with no
evidentiary value. A CV full of these phrases and nothing else scores 2 (maximum).

3. ENVIRONMENT WEIGHT:

Evidence from [INDUSTRY NAME] / aviation / military / offshore contexts carries more weight
than evidence from general manufacturing or construction.

4. ABSENCE OF EVIDENCE:

If the CV contains no documented safety-specific actions or regulated-environment roles,
score 1 regardless of the candidate's claimed years of experience.

In your EVIDENCE (inside the reasoning text): quote the exact phrase, role, or event from the CV
that demonstrates a [INDUSTRY PEER ORGANISATION] indicator, and note which SC indicator (SC-1 through SC-7) it supports.

Do NOT add extra JSON fields — output only "reasoning" and "score".

If no qualifying evidence exists, write: NO EVIDENCE FOUND and score 1.

BARS RUBRIC:

BARS scale — Safety Readiness / Safety Culture (integer 1–5):

[INDUSTRY PEER ORGANISATION] Safety Culture Indicators ([INDUSTRY STANDARD]) — search explicitly for:

SC-1 Industry Safety Priority: statements placing safety above schedule or production.

SC-2 Conservative Decision-Making: "when in doubt, don't" / escalation / questioning attitude.

SC-3 Procedure Adherence: documented work per written procedures; deviation reporting.

SC-4 Stop-Work Authority: any mention of halting or challenging an unsafe condition.

SC-5 Cross-checking / Independent Verification: habit of verification before action.

SC-6 Learning Organisation: post-incident analysis, lessons-learned, corrective actions.

SC-7 Personal Accountability: ownership of outcomes, incident/near-miss self-reporting.

INFERENCE CONTROL — the following phrases are MARKETING LANGUAGE and score ZERO evidence:

"safety-conscious", "responsible", "attentive to detail", "safety-first attitude",
"committed to safety", "follows all safety rules". Do NOT count these as evidence.

Score:

1 = No regulated-environment exposure; background shows no safety-critical work.

2 = Only marketing-language safety claims; no specific, datable, verifiable safety events.

3 = Documented experience in a regulated environment with procedure-based work;

at least one SC indicator supported by a concrete role or responsibility description.

4 = Clear evidence of ≥2 [INDUSTRY PEER ORGANISATION] SC indicators with specifics (dates, roles, named procedures,

or documented safety actions). OR: work in [INDUSTRY NAME]/aviation/military safety context.

5 = Strong, documented track record of ≥3 [INDUSTRY PEER ORGANISATION] SC indicators in safety-critical environment

([INDUSTRY NAME] preferred). Concrete examples with outcomes. OR: formal safety culture role

(safety inspector, regulated-environment safety officer, [INDUSTRY NAME] shift supervisor).

Overall regulatory fit (w_5 = 0.05)

CRITERION-SPECIFIC INSTRUCTION:

You are a senior regulatory reviewer preparing an expert conclusion for a staffing commission

at [ANONYMIZED HIGH-HAZARD INDUSTRIAL FACILITY]. This criterion is the synthesis score.

Task:

Based on the evidence already evaluated across Education, Experience, Qualifications, and Safety

Readiness criteria, provide an overall regulatory and professional fitness assessment.

This is the commission's final recommendation — apply the most conservative defensible interpretation.

Instructions:

1. SYNTHESIS — do not re-evaluate individual criteria; instead, assess the overall regulatory picture:

- Are there any hard disqualifiers (missing mandatory degree, expired permit, prohibited background)?
- Does the cumulative profile meet the eligibility threshold for the stated [FACILITY NAME] role?
- Is the candidate suitable for immediate appointment, conditional appointment, or ineligible?

2. CONSERVATIVE APPROACH ([INDUSTRY STANDARD BODY] [INDUSTRY STANDARD]):

- When evidence is ambiguous or incomplete, score down.
- "Probably meets requirements" is not a basis for a positive recommendation.
- Benefit of the doubt does NOT apply in a [INDUSTRY NAME] safety-critical context.

3. REGULATORY FRAMEWORK AWARENESS:

Award higher scores for documented familiarity with or work under:

- Ukrainian [INDUSTRY NAME] regulatory framework: [ANONYMIZED REGULATORY AUTHORITY AND NORMS] ([INDUSTRY NAME] norms)
- the national [INDUSTRY NAME] operator's internal qualification system
- [INDUSTRY STANDARD BODY] safety standards
- [INDUSTRY PEER ORGANISATION] peer review or self-assessment processes

4. SCORE MAPPING:

Use the integer 1–5 scale (see rubric below). Do NOT use legacy compliance labels

such as "Fully Compliant" or "Non-Compliant" — these are superseded by the 1–5 scale.

5. COMMISSION RECOMMENDATION — end your VERDICT with one of:

- "Recommended for appointment" (score 4–5)
- "Conditionally recommended / requires additional training" (score 3)
- "Not recommended" (score 1–2)

In your EVIDENCE: cite the single most impactful positive and (if any) the single most

impactful negative factor from the full candidate profile.

BARS RUBRIC:

BARS scale — Overall Regulatory Fit (integer 1–5):

This is the synthesis criterion. Evaluate the candidate's overall alignment with the

regulatory, organisational, and professional context of an [FACILITY NAME] staffing commission.

Use a conservative approach: when evidence is ambiguous, score down.

1 = Background is incompatible with [INDUSTRY NAME] regulatory requirements.

Example: background in unrelated field; disqualifying factors in work history.

2 = Partial alignment; significant regulatory or qualification gaps not addressable

without substantial additional training or certification.

3 = Generally suitable; no disqualifying factors; meets baseline eligibility.

Candidate could be considered with standard onboarding.

4 = Clear regulatory alignment: documented compliance work, familiarity with Ukrainian

[INDUSTRY NAME] regulatory framework ([ANONYMIZED REGULATORY AUTHORITY]), or national-operator procedures.

5 = Candidate is immediately eligible from a regulatory standpoint: holds relevant

permits, has passed prior [FACILITY NAME] qualification review, or has equivalent documented

qualification pathway. Commission recommendation: Eligible for appointment.