

Retention-Aware Match Tracing for Auditable Allocation of Rural Youth Talent to Village Posts under Generative AI-Assisted Profiling

Yang Lu

Department of Electronic and Information Engineering, Tongji Zhejiang College, Jiaxing, Zhejiang 314051, China

Abstract—County programmes that recruit young people into rural industry and village governance increasingly pair generative-AI profiling of unstructured application materials with algorithmic person-post matching, yet the matchers in use optimise immediate placement, explain themselves post hoc if at all, and are silent about the outcome the whole talent chain is built around: whether the young person is still in the village two years later. This paper proposes Retention-Aware Match Tracing (RAMT), a matching engine that is auditable by construction. The scorer is additive over a pre-registered ledger of person-post features, each passed through a piecewise-linear spline, so every match ships an evidence ledger that reconstructs its score exactly, with zero residual rather than an attribution estimate. The score feeds a discrete-time retention hazard trained on logged administrative matching records with inverse-propensity weights, and offers are formed by capacity-constrained deferred acceptance, with low-margin matches abstained to human review. Because no county authority can release youth-level records, evaluation uses a simulated labour micro-market whose marginals are anchored to published national statistics and whose ground-truth retention process is deliberately outside the model family — regime-switching, threshold-bearing and interactive. Across 20 market replications, RAMT converts 28.1 of every 100 offers into matches still in place at 24 months, 4.3 times the administrative wage-rank rule and 88.2 per cent of the oracle ceiling, while its ledger passes counterfactual-flip audits that sampled Shapley trails on equally accurate black-box baselines fail. The principal warning is distributional: because village wages sit below graduate expectations, a retention-optimal matcher rationally deprioritises the most educated youth, assigning them at 0.65 times the rate of their less-educated peers — the engine quietly gives up on exactly the people the policy is trying to attract, and repairing this costs a small, quantifiable share of total retention.

Index Terms—Auditable artificial intelligence, generative artificial intelligence, person-post matching, rural revitalization, youth employment.

I. INTRODUCTION

Rural revitalisation runs on young people, and China's counties now compete for them. National monitoring counts nearly 300 million rural migrant workers [1], and county policy now works to draw part of that cohort back toward

county and village employment. What a returning young person finds is a thin, fragmented market: village posts in modern agriculture, rural e-commerce, agritourism and grassroots governance are scattered across dozens of administrative villages, capacities are small, wages sit below urban expectations, and the match between what a post needs and what a young person can do is judged by whoever happens to administer the scheme. Field studies of returnee entrepreneurship show how sensitive these trajectories are to local embeddedness, prior experience and capability endowments [2], [3], [4], and county governments have responded by building talent-chain programmes that recruit, train, place and — they hope — retain rural youth.

Generative artificial intelligence is entering this chain at two points. Upstream, large language models turn unstructured application materials, training records and post descriptions into structured profiles — the digital portraiture that county programmes now build [5], and a task at which structured extraction from free text is by now well demonstrated [6]. Downstream, the structured profiles feed person-post matching, for which the technical shelf offers learned person-job fit models of considerable sophistication [7], [8]. A county talent office that wires these two components together has, in effect, an automated employment decision system for a public programme.

Three things are wrong with the default way of wiring it. First, the objective is wrong: recommender-style matchers are trained on immediate outcomes — clicks, applications, offer acceptance — whereas the talent chain's binding constraint is retention. The predictable failure mode follows: young people are placed, and then leave, because the placement optimised the placement rate. Second, the models are opaque: a two-tower network scores a pair and the office cannot state why this youth was routed to that post, which matters because an automated employment decision about a public programme is precisely the class of decision that regulation now expects to be explained and audited — the European Union's AI Act classifies employment-related systems as high-risk [9], New York City already mandates bias audits of automated employment decision tools [10], and China's interim measures on generative AI impose transparency duties on exactly the profiling layer upstream [11]. Third, the allocation is unstable: greedy platform-style filling leaves pairs who would both prefer each other unmatched to their assigned partners, and in a small two-sided market such instability is not a technicality

This work was supported by the 2024 Zhejiang Provincial Philosophy and Social Sciences Planning "Provincial-Municipal Cooperation" Research Project: "A Study on Pathways and Institutional Innovations for the Participation of New Rural Elders in Grassroots Governance in the Context of Comprehensive Rural Revitalization: A Case Study of Jiaxing, Zhejiang" under Grant No. 24SSHZ122YB.

Yang Lu is with the Department of Electronic and Information Engineering, Tongji Zhejiang College, Jiaxing, Zhejiang 314051, China (e-mail: guhs225@163.com).

but churn.

This paper proposes Retention-Aware Match Tracing (RAMT), a matching engine built so that the three failures cannot occur by construction rather than by good intentions. The scorer is additive over a pre-registered ledger of person-post features, each passed through a one-dimensional piecewise-linear spline in the tradition of intelligible additive models [12], [13]; because the score is a sum, every match ships an evidence ledger whose entries reconstruct the score exactly, with zero residual, rather than approximately in the manner of post-hoc attribution [14], [15]. The score parameterises a discrete-time retention hazard [16], trained not on placements but on what became of them, using logged records of the county's historical matching rule with inverse-propensity weights to undo the logging policy's selection [17], [18]. Offers are formed by capacity-constrained deferred acceptance [19], [20], which leaves no blocking pair with respect to the stated preferences, and matches whose score margin is small under an ensemble are abstained to a human case worker [21], following the deference designs of human-AI decision making [22].

Evaluating such an engine raises an obstacle that is itself methodologically interesting: no county authority can release youth-level administrative records with retention outcomes, and any evaluation on real logs would in any case be unable to observe the counterfactual — what would have happened under the assignment the engine did not make. We therefore evaluate on a simulated labour micro-market, stated as such in every section of this paper, in the tradition of simulation-based evaluation of decision systems [23], [24]. The simulator's marginal distributions are anchored to published national statistics [1], [25] and its ground-truth retention process is deliberately constructed outside the model family of every engine evaluated: hazards switch regimes between a settling-in phase and a consolidation phase, carry a hard wage-expectation threshold, and contain multiplicative interactions. An additive scorer cannot win on this ground truth by specification identity, and a further validity check re-runs the entire evaluation under a second, structurally different generator.

The contributions are fourfold. (1) An auditable matching architecture that couples an exactly decomposable additive scorer to an off-policy retention hazard and a stable capacity-constrained assignment, with abstention as a first-class output. (2) A decision-level audit methodology: beyond the reconstruction residual, every evidence trail is subjected to sufficiency, minimality and counterfactual-flip checks against the runner-up post, criteria in the spirit of rationale evaluation [26] and formal explanation minimality [27], applied here to allocation decisions and equally to the sampled Shapley trails of black-box baselines. (3) A quantified account, on the simulated market, of what each design choice buys: the retention objective, the additive scorer, the stable assignment and the propensity correction are ablated one at a time. (4) Two findings of direct policy relevance: optimising immediate acceptance sacrifices retention among formed matches and

vice versa, so the mixing weight between the two is a policy dial rather than an engineering constant; and a retention-optimal matcher rationally deprioritises the most educated applicants because village wages fall short of their expectations — an equity failure that arises precisely when the engine is working as intended, and that a single audited priority offset repairs at a measured cost.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 fixes notation and reviews the components the architecture builds on. Section 4 presents RAMT. Section 5 describes the simulated micro-market, the logged behaviour policy, the baselines and the metrics. Section 6 reports matching performance, audits, ablations, distributional analysis and validity checks. Section 7 discusses deployment and limitations, and Section 8 concludes.

II. RELATED WORK

The engine sits at the intersection of four literatures: digital technology in rural revitalisation, learned person-job matching, two-sided market design, and interpretable machine learning under algorithmic accountability.

A. Digital Technology and the Rural Youth Talent Chain

Province-level panels associate rural digital-economy development with composite revitalisation outcomes [28], and conceptual work maps where generative AI can enter the countryside's organisational, industrial and talent processes [5]. The talent link is the least automated. Studies of returnee entrepreneurship identify who returns and who succeeds: migrant work experience shifts returnees toward transformational rather than subsistence ventures [3], local embeddedness and bridging networks condition survival in marginal regions [2], and large questionnaire studies decompose the entrepreneurial capability of returning workers [4]. The distributional stakes are documented too: whether rural e-commerce narrows or widens local income gaps has been shown to turn on where its gains concentrate [29]. What this literature does not contain is an allocation mechanism: given profiles and posts, who should go where. That is the gap this paper addresses, and the recent field-experimental finding that algorithmic assistance upstream of hiring shifts real labour-market outcomes [30] argues that the allocation layer deserves the same scrutiny.

B. Learned Person-Job Matching

Neural person-job fit models score resume-post pairs with hierarchical attention and its successors [7], and e-recruitment recommendation has accumulated a survey-scale literature of its own [8]. Industrial retrieval rests on two-tower architectures trained on logged engagement [31], and large language models now both generate recommendations directly [32] and screen resumes — with measured intersectional bias when they do [33]. Vendor claims about de-biased algorithmic hiring have been examined and found wanting [34]. Two properties are near-universal in this line and both are liabilities for a public talent programme: the training signal is an immediate engagement outcome rather than what became of

the placement, and the scorer is a black box explained, if at all, after the fact. Both choices are reversed here.

C. Two-Sided Matching and Market Design

Deferred acceptance is the canonical mechanism for two-sided markets with capacities [19]; its stability guarantee — no youth-post pair prefers each other to their assignment — has carried it from college admissions to residency matching and school choice [20], [35], [36]. Machine learning has recently entered matching markets from the learning side, with bandit learners converging to stable matchings [37] and fairness analysed when the qualities being matched on are themselves uncertain predictions [38]. Entropy-regularised optimal transport offers a continuous alternative to deferred acceptance for capacity-respecting assignment [39], [40], and inverse optimal transport has been used to learn matching costs from observed matches [41]. RAMT uses deferred acceptance as its default allocation layer and reports the transport alternative as an ablation, because stability is an institutional virtue: an assignment that leaves obvious mutually preferred pairs unrealised invites exactly the informal side-dealing that a transparent programme is meant to replace.

D. Interpretable Models, Audits and Accountability

The case that high-stakes decisions call for models interpretable by construction rather than explained post hoc has been made forcefully [42], and additive models are its workhorse: generalised additive models with selective interactions match black-box accuracy on tabular problems [12], [43], and neural additive models carry the idea into deep learning [13], with statistical foundations reviewed in [44]. The post-hoc alternative — LIME, SHAP and their descendants [14], [15] — approximates the model rather than reporting it, and the faithfulness of such explanations is itself a contested measurement problem [45], [26], [46], [47]. Formal explanation research supplies sharper criteria — minimal sufficient reason sets computed abductively [27] and counterfactual explanations stating what would have had to differ [48]. On the accountability side, fairness in ranking and allocation is a mature field [49], [50], [51], [52], and the first years of mandated hiring-algorithm audits have been studied empirically, with sobering compliance rates [10], [53]. This paper’s contribution to the line is to move the audit from the score to the decision: the checks in Section 4 ask whether a trail justifies the allocation actually made, against the alternative actually foregone.

III. PRELIMINARIES

This section fixes notation and reviews the three components on which the engine builds: two-sided matching with capacities, discrete-time survival, and learning from logged allocation data.

A. Problem Formulation

A matching round presents a set of young people $i \in [N]$, each described by a structured profile produced by the

upstream generative-AI layer, and a set of village posts $j \in [M]$ with capacities c_j . An engine must output an assignment $\mu : [N] \rightarrow [M] \cup [\emptyset]$ together with, for every assigned pair, an evidence trail. Writing A_{ij} for the probability that youth i accepts an offer of post j and R_{ij} for the probability that the formed match survives the full horizon of $T = 24$ months, the quantity the talent chain cares about is

$$\max_{\mu} \sum_{i : \mu(i) \neq \emptyset} A_{i\mu(i)} R_{i\mu(i)} \quad \text{s. t.} \quad |\mu^{-1}(j)| \leq c_j \quad (1)$$

where the summand is the probability that the offer to i both forms and lasts, which we call the stay-yield of the offer. Neither A nor R is observable at decision time; the engine works from estimates, and the estimates come from logs.

B. Stability in Capacitated Matching

Given youth-side preferences $u_i(\cdot)$ over posts and post-side priorities $s(\cdot, j)$ over youth, an assignment μ is stable if no youth-post pair (i, j) would jointly deviate from it, that is, if no pair satisfies

$$u_i(j) > u_i(\mu(i)) \wedge [|\mu^{-1}(j)| < c_j \vee s(i, j) > s_{\min}(j)] \quad (2)$$

where $s_{\min}(j) = \min_{i' \in \mu^{-1}(j)} s(i', j)$ is the weakest priority currently holding a seat at j , and $u_i(\emptyset)$ is ranked below every acceptable post. Youth-proposing deferred acceptance produces a stable assignment in this sense and is strategy-proof for the proposing side [19], [20]. Stability is reported below as a measurable property — the rate of blocking pairs an assignment leaves — rather than assumed.

C. Discrete-Time Survival and Logged Data

Retention over a horizon observed monthly is a discrete-time survival problem [54], [16]: a formed match either survives month t or ends in it, so a per-month hazard h_t determines the survival probability $\prod_t (1 - h_t)$, and censored spells contribute only the months they were observed. Machine-learning survival modelling is a mature field [55], [56], and employee-turnover prediction its best-studied industrial instance [57]. The complication here is that training data are logs of a historical allocation policy: outcomes exist only for matches that policy formed, so naive fitting inherits its selection. The standard correction weights each logged episode by the inverse of the probability the policy showed that pair [17], [18], the counterfactual-learning machinery that also underlies doubly robust evaluation [58] and policy learning from observational data [59]; long-horizon outcomes raise further identification questions of their own [60].

IV. RETENTION-AWARE MATCH TRACING

Figure 1 shows the pipeline the engine is deployed in: the generative-AI layer turns unstructured materials into structured profiles, RAMT scores and allocates, offers go out with their evidence ledgers, low-margin cases are routed to a case worker, and observed acceptance and retention flow back as the next round’s training data. Figure 2 shows the engine’s

internal structure, which the rest of this section walks through.

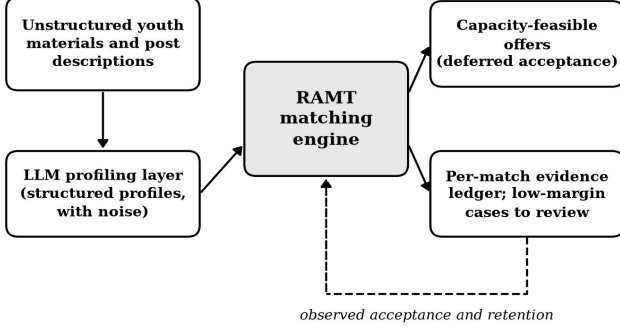


Figure 1. The deployed pipeline. The generative-AI profiling layer feeds RAMT, whose offers carry evidence ledgers and abstentions; observed acceptance and retention feed the next round of training.

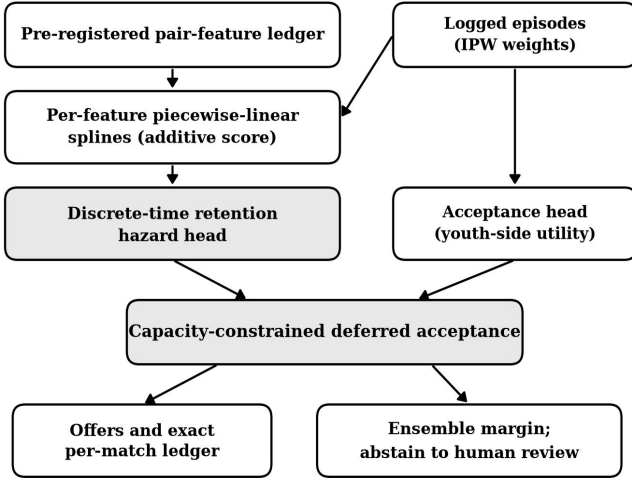


Figure 2. The RAMT architecture: additive spline scoring over the pre-registered feature ledger, retention and acceptance heads mixed by the dial α , capacity-constrained deferred acceptance, and ensemble-margin abstention.

A. A Pre-Registered Feature Ledger

Every scored pair is described by a fixed vector of $D = 15$ person-post features

$$z(i, j) = (z_1(i, j), \dots, z_D(i, j)) \quad (3)$$

comprising four requirement-weighted skill-fit scores (agronomy, digital commerce, operations, governance), the wage offered minus the wage expected, travel distance to the home village, mentor pairing, housing support, the youth's schooling level and return-migrant status, a home-village tie, the youth's generative-AI literacy interacted with whether the post is a digital one, and three interaction features fixed in advance on domain grounds: distance interacted with having children, village amenities interacted with having children, and mentor pairing interacted with skill shortfall. The ledger is pre-registered in the exact sense that its composition was frozen before any model was estimated and no term was added, removed or reweighted afterwards; the interaction whitelist is the entire concession the additive family makes to non-additivity, and it is declared rather than searched.

B. An Additive Spline Scorer with an Exact Ledger

Each feature passes through its own one-dimensional piecewise-linear spline, and the score is their sum. With hat functions B_{kp} on quantile knots and weights w_{kp} ,

$$\begin{aligned} f_k(z) &= \sum_{p=1}^{P_k} w_{kp} B_{kp}(z), \quad s(i, j) \\ &= b + \sum_{k=1}^D f_k(z_k) \end{aligned} \quad (4)$$

which is the classical shape of an intelligible additive model [12], [13]: flexible within each feature, additive across them, and linear in its parameters, so the training problems below are smooth and convex-like and run in CPU seconds. Defining each contribution against the population-mean reference profile $\bar{z}$,

$$\begin{aligned} \varphi_k(i, j) &= f_k(z_k(i, j)) - f_k(\bar{z}_k), \quad s(i, j) \\ &= b + \sum_{k=1}^D \varphi_k(i, j) \end{aligned} \quad (5)$$

holds as an identity, not an approximation: the per-feature entries φ_k are the evidence ledger, and they reconstruct the score with zero residual because they are the score. This is the property post-hoc attribution methods estimate for black boxes [15] and that comes free by construction, which is the architectural bet of the paper: Section 6 measures what, if anything, the bet costs in accuracy.

C. A Retention Hazard Trained Off-Policy

The score parameterises a discrete-time hazard with regime-block intercepts $a_{r(t)}$ for months 1–6, 7–12 and 13–24,

$$\begin{aligned} h_t(i, j) &= \sigma(a_{r(t)} - s(i, j)), \quad \hat{S}(i, j) \\ &= \prod_{t=1}^T (1 - h_t(i, j)) \end{aligned} \quad (6)$$

where σ is the logistic function; a higher score means a lower hazard in every month, so s orders pairs by predicted retention while the intercepts absorb the calendar shape. The training data are logged episodes of the county's historical matching rule: pairs shown, offers accepted or declined, and for formed matches the observed months survived m_e with a censoring flag κ_e . Each episode carries the exact probability π_e with which the logging rule showed that pair, and the fitting criterion is the inverse-propensity-weighted negative log-likelihood

$$\begin{aligned} L_{\text{ret}} &= - \frac{1}{\sum_e \omega_e} \sum_e \omega_e \left[\sum_{t=1}^{m_e} \ln(1 - h_t^e) \right. \\ &\quad \left. + (1 - \kappa_e) \ln h_{m_e+1}^e \right] \end{aligned} \quad (7)$$

with weights $\omega_e = \min(1/\pi_e, W)$ and the clip $W = 20$ in the usual way [17]. Nothing in the training loop ever touches the simulator's latent process; the engine sees exactly what a county system would see, which is its own logs.

D. An Acceptance Head and a Mixing Dial

A second head with the same basis models the youth side: the probability an offer is accepted,

$$g(i, j) = \sigma \left(b_a + \sum_{k=1}^D f_k^a(z_k(i, j)) \right) \quad (8)$$

trained on all logged offers by weighted logistic regression. Its score serves two roles: it supplies the youth-side preference

orderings that deferred acceptance requires, and it provides the other end of a dial. Standardising both scores and mixing,

$$s_\alpha(i, j) = (1 - \alpha) \tilde{s}(i, j) + \alpha \tilde{g}(i, j) \quad (9)$$

interpolates between a matcher that optimises predicted retention ($\alpha = 0$, the engine's default) and one that optimises predicted acceptance ($\alpha = 1$, the objective of deployed platforms). Section 6 traces the frontier and argues the choice of α is a policy decision that should be taken in the open.

E. Assignment by Deferred Acceptance

Offers are formed by youth-proposing deferred acceptance: each youth proposes down their acceptance-utility ordering, each post tentatively holds its c_j best proposals by the engine score, and displaced youth continue proposing. The outcome is stable in the sense of (2) with respect to the stated preferences (g, s) . As a continuous alternative we also evaluate an entropy-regularised transport plan [39], [40] obtained by Sinkhorn scaling of

$$K_{ij} = \exp(s(i, j) / \varepsilon), \quad u \leftarrow a / K v, \quad v \leftarrow b / K^\top u \quad (10)$$

with row masses a uniform over youth, column masses b proportional to capacities and the rounded plan used as an assignment. Transport fills every seat by construction; what it gives up, the ablations of Section 6 measure, is stability.

F. Decision-Level Audits and Abstention

An evidence trail should justify the decision actually taken, against the alternative actually foregone. For an assigned pair (i, j^*) with runner-up j' , the audited object is the ledger of the score difference, $\Delta \varphi_k = \varphi_k(i, j^*) - \varphi_k(i, j')$, and three checks are applied. Sufficiency asks whether a small set of entries carries the preference:

$$\begin{aligned} \min |S| \quad \text{s.t.} \quad & \sum_{k \in S} \Delta \varphi_k \\ & \geq \frac{1}{2} \sum_{k=1}^D \Delta \varphi_k > 0 \end{aligned} \quad (11)$$

in the spirit of rationale sufficiency [26]; minimality requires the selected set to be small in the sense of abductive explanation [27], here at most a third of the ledger; and the counterfactual-flip check neutralises the largest entry $k^\top$ to the reference value and requires the recomputed margin to move as the ledger predicted:

$$|\Delta s_{k^\top \leftarrow \bar{z}} - (\Delta s - \Delta \varphi_{k^\top})| \leq 0.25 |\Delta s| \quad (12)$$

which is a counterfactual criterion in the sense of [48]: a trail whose largest stated reason, when removed, does not move the decision by its stated share is a rationalisation rather than a reason. For the additive scorer the identity (5) makes all three checks pass mechanically; the point of running them is that the same checks can be applied to any engine whose trails are sampled Shapley attributions, which is how the black-box baselines are audited in Section 6.

Finally, the engine declines to decide alone where its own evidence is thin. An ensemble of five hazard heads trained on bootstrap resamples yields a score dispersion for each assigned pair,

$$v(i) = \text{std}_b(s^{(b)}(i, \mu(i))) \quad (13)$$

and the most uncertain share $1 - q$ of matches is routed to

human review rather than auto-approved, the selective-prediction pattern [21], [61] in its learning-to-defer form [22]. Section 6 traces retention against coverage q ; calibration of the underlying probabilities is checked with standard tools [62], [63].

V. EXPERIMENTAL SETUP

To evaluate a matching engine one must know what would have happened under assignments the historical policy never made, and no administrative dataset contains that. The evaluation therefore runs on a simulated labour micro-market. The corpus is simulated; no young person and no village post described below exists, and every number in Section 6 is a property of the simulation. What discipline the exercise carries comes from three commitments: marginals anchored to published statistics, a ground-truth process deliberately outside every evaluated model family, and engines that see only what a deployed system would see — their own logs.

A. The Simulated Micro-Market

One market draw comprises 2,400 young people and 360 posts across 60 villages, with post capacities of two to eight and total capacity below demand, so allocation is genuinely competitive. Youth profiles carry age, schooling, a twelve-skill vector in four groups, digital and generative-AI literacy, wage expectation, home village, family ties and children; posts carry a category (modern agriculture, rural e-commerce, agritourism, village governance, digital agriculture), requirement profiles, wage, mentor pairing, housing support and village amenities. The education structure of the youth population follows the published schooling distribution of young rural migrant workers [1]; digital-literacy levels are a declared modelling shape consistent with rural internet penetration of roughly two-thirds [25]; wage scales for posts and expectations sit below the published national migrant-worker average monthly wage of 4,961 CNY [1], which is what makes staying in the county an economic decision at all. Every remaining marginal — capacities, category shares, amenity levels, the share with children — is a stated modelling choice, and the full anchor table ships with the code.

The ground-truth retention process is built to be unwinnable by specification. Monthly hazards switch regime at month six: in the settling-in regime, hazard is driven by the expectation shortfall, by distance interacted with children, and by a hard threshold that fires when the offered wage falls below 80 per cent of the youth's expectation; in the consolidation regime, by skill fit, community ties, amenities interacted with children, mentor pairing interacted with fit, and generative-AI literacy on digital posts. Mentorship helps misfits far more than good fits, and none of this structure — the switch, the threshold, the interactions — is in the hypothesis class of any engine below. Acceptance follows a separate two-sided utility with its own noise. A second, structurally different generator (a single smooth logistic hazard with no regimes, no threshold and no interactions) is held out for the validity checks of Section 6.

B. Logged Behaviour Policy

Engines train on logs of the administrative rule the programme is imagined to have run for four annual rounds: an eligibility filter followed by wage-rank greedy assignment, with a 15 per cent exploration rate, the kind of wage-led allocation county practice actually reports. Logging yields 9,600 episodes per market, of which 2,140 formed matches with observed survival; the logged acceptance rate is 0.223 and 0.440 of formed matches survive the 24-month horizon. Every episode records the exact probability with which the rule showed that pair, which is what the propensity weights in (7) consume. No engine ever queries the generator.

C. Baseline Methods

Five comparison engines span current practice.

Administrative rule: the logging policy without exploration — eligibility screen plus wage-rank greedy fill. This is the incumbent.

First-come queue: youth in random order each take their most-preferred feasible post under the acceptance model; the floor that any coordination should beat.

Platform matcher: a two-hidden-layer neural scorer of the two-tower kind [31] trained on offer acceptance, deployed greedily, as engagement-optimising platforms deploy them; audited post hoc with sampled Shapley values [15].

Interpretable acceptance matcher: the additive acceptance head (8) feeding deferred acceptance — identical machinery to RAMT with the objective swapped, so the comparison isolates the objective.

Boosted retention matcher: gradient-boosted shallow trees trained on 24-month retention with the same propensity weights, deployed greedily; the strongest tabular learner in the pool, audited with sampled Shapley values. A variant feeding deferred acceptance is reported in the ablations, so scorer and allocator are never confounded.

An oracle upper bound runs deferred acceptance directly on the generator’s true retention probabilities. It is not a competitor; it prices the headroom.

D. Metrics

Deployment metrics are computed exactly under the generator: offer acceptance, 24-month retention among formed matches, and the headline stay-yield per 100 offers — the expected number of offers that both form and survive the horizon, the objective in (1) scaled to a round of one hundred. Skill fit of assigned pairs, the blocking-pair rate of (2) estimated on sampled pairs against the retention-priority preferences, and an assignment-parity ratio — the assignment rate of youth with college education and above over that of the rest — complete the deployment picture. Audit metrics follow the checks of Section 4: reconstruction residual as a share of the score difference, and pass rates of the sufficiency, minimality and counterfactual-flip checks, computed on samples of assigned matches. Ranking quality of each scorer is summarised by pairwise AUC against true retention on random pairs. All results aggregate 20 independent market replications (audits over 8), reported as mean and standard

deviation across markets.

E. Implementation Details

Every model, the simulator, the assignment algorithms, the audits and the boosting and neural training loops are implemented from first principles in NumPy for this study; there is no external learning framework, no GPU and no pretrained component. Spline scorers use six quantile knots per continuous feature; the platform matcher has hidden widths 24 and 12; boosting uses 120 depth-two trees; Sinkhorn uses regularisation 0.08. Deferred acceptance on a 2,400-by-360 market runs in under a second, a full market replication in about 20 CPU seconds, and the entire experimental programme in 29.2 minutes on commodity hardware. Code, seeds and the anchor table reproduce every number in this paper.

VI. RESULTS AND ANALYSIS

A. Matching Performance

Table 1 reports the deployment metrics. The incumbent administrative rule converts 6.48 of every 100 offers into matches still in place at 24 months; it offers everyone a post, but the posts are chosen by wage rank, accepted at only 0.164, and abandoned quickly. Every learned engine beats it by a wide margin, which is the cheapest finding here: almost any signal beats none. The informative comparisons are among the learned engines. RAMT attains a stay-yield of 28.13 (1.57) per 100 offers — 4.3 times the incumbent and 88.2 per cent of the oracle ceiling of 31.88 (1.17) — with 24-month retention among formed matches of 0.605 (0.010), the highest of any learned engine, and a blocking-pair rate statistically indistinguishable from zero. Mean skill fit of assigned pairs is nearly flat across the pool (from 0.421 to 0.438), so no engine buys its yield by mismatching skills, and Table 1 omits the column. One baseline deserves a sentence of its own: the platform-style neural acceptance matcher, deployed greedily as such systems are, yields 21.59 (1.22) and does not clear the first-come floor of 22.64 (0.84) — engagement-style ranking without coordination reproduces what an uncoordinated queue already achieves.

Table 1. Deployment Performance on the Simulated County Micro-Market (Mean (SD) over 20 Market Replications)

Engine	Accept. rate	Retent. 24 m	Stay-yield /100 offers	Block. rate	Parity (high/low)
Admin. rule (wage rank)	0.164 (0.010)	0.395 (0.008)	6.48 (0.51)	0.4996 (0.0176)	1.00 (0.03)
First-come queue	0.429 (0.012)	0.528 (0.008)	22.64 (0.84)	0.1836 (0.0223)	1.00 (0.02)
Platform matcher [31]	0.421 (0.015)	0.513 (0.014)	21.59 (1.22)	0.2116 (0.0228)	1.01 (0.03)
Interpretable acceptance matcher	0.511 (0.013)	0.570 (0.008)	29.13 (1.01)	0.1260 (0.0283)	0.57 (0.05)

Engine	Accept. rate	Retent. 24 m	Stay-yield /100 offers	Block. rate	Parity (high/low)
Boosted retention matcher	0.357 (0.045)	0.533 (0.032)	19.06 (2.67)	0.2523 (0.0534)	1.00 (0.03)
RAMT (ours)	0.465 (0.022)	0.605 (0.010)	28.13 (1.57)	0.0005 (0.0006)	0.65 (0.13)
Oracle (upper bound)	0.522 (0.012)	0.611 (0.010)	31.88 (1.17)	0.0827 (0.0195)	0.50 (0.04)

Two structural lessons sit inside the table. First, the objective matters exactly as the design argued: the interpretable acceptance matcher — RAMT’s machinery with the objective swapped — fills slightly more seats (acceptance 0.511 (0.013) against 0.465 (0.022)) but holds them less well (retention 0.570 (0.008) against 0.605 (0.010)); on the stay-yield product the two land within one standard deviation of each other (29.13 (1.01) against 28.13 (1.57)), because forming more matches and keeping each formed match longer trade off almost exactly under this wage structure. The mixing dial of Section 6 spans the entire frontier between the two objectives, so where a county sits on it is a published policy parameter, not a property of the engine. Second, the allocator matters as much as the scorer: the boosted-tree scorer is the best ranker in the pool (pairwise AUC 0.947 against 0.864 for the additive scorer) yet deployed greedily it yields only 19.06 (2.67), because greedy filling spends the best posts on whoever is processed first and leaves a 0.252 blocking-pair rate behind. A better prediction, allocated worse, loses; the ablations return to this.

B. Auditability

Table 2 applies the decision-level audits of Section 4 to every engine that ships a trail. For RAMT the trail is the ledger; for the black-box engines it is sampled Shapley attribution over the same features, with the same checks applied to the same decisions. The reconstruction residual is deliberately uninformative: the ledger reconstructs its score because (5) is an identity, and a permutation-sampled Shapley trail also telescopes exactly to the score difference, so every row reports zero and summing to the decision cannot separate a trail that is the computation from one that merely balances. The informative columns are the decision checks. The ledger passes the counterfactual-flip check at 1.000; the platform matcher’s Shapley trail passes at 0.188 and the boosted retention scorer’s at 0.360 — trails that read plausibly but, when their largest stated reason is removed, do not move the decision by the stated amount, because the underlying functions are not additive and the attribution pretends they are. An office that acted on such a trail — waiving a distance concern here, a wage top-up there — would be acting on reasons the model does not actually have.

Table 2. Decision-Level Audit Outcomes for Every Engine That Ships a Trail (Mean (SD) over 8 Market Replications)

System (trail)	Recon. residual	Sufficiency pass	Minimality pass	Counterf. flip pass
----------------	--------------------	---------------------	--------------------	------------------------

System (trail)	Recon. residual	Sufficiency pass	Minimality pass	Counterf. flip pass
RAMT evidence ledger (exact)	0.000 (0.000)	1.000 (0.000)	0.949 (0.059)	1.000 (0.000)
RAMT, neural scorer + sampled Shapley [15]	0.000 (0.000)	1.000 (0.000)	0.922 (0.052)	0.114 (0.067)
Platform matcher + sampled Shapley	0.000 (0.000)	1.000 (0.000)	0.904 (0.067)	0.188 (0.098)
Boosted retention matcher + sampled Shapley	0.000 (0.000)	1.000 (0.000)	0.945 (0.036)	0.360 (0.153)

Figure 3 shows what the ledger looks like for a single median-margin match: each bar is one feature’s exact contribution to the preference for the offered post over the best alternative, and the bars sum to the score difference to machine precision. This is the artefact a case worker receives, and every claim in it is checkable against the profile fields it names.

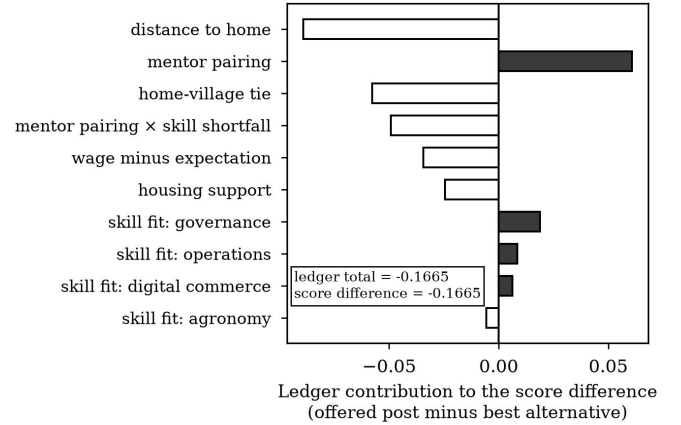


Figure 3. The exact evidence ledger for one median-margin match (market replication 1): the ten largest signed contributions to the score difference between the offered post and the best alternative. Filled bars favour the offered post.

C. Ablation Study

Table 3 removes one design commitment at a time. Swapping the retention objective for acceptance costs 3.55 points of 24-month retention among formed matches; replacing deferred acceptance with greedy filling collapses stay-yield from 28.13 (1.57) to 19.97 (2.55) and instability jumps two orders of magnitude, the single largest effect in the table and the clearest argument that allocation, not prediction, is where deployed systems lose most of their value. Dropping the propensity weights costs 0.79 points of stay-yield: the logging rule showed high-wage posts disproportionately, and a hazard fitted naively to its selection inherits the distortion. Replacing the additive splines with an unconstrained neural scorer — the interpretability bet — does not pay for itself: the black-box variant yields 26.26 (2.06) against the additive 28.13 (1.57), so on logged samples of this size the inspectable model is not merely free, it is better, consistent with the tabular record of additive models [43]. The boosted scorer fed

through deferred acceptance recovers everything greedy filling had cost it — 29.13 (2.64) against the additive 28.13 (1.57), a gap inside one standard deviation — which confirms once more that the allocator was the bottleneck; but it arrives without a ledger, fails the counterfactual-flip audit of Table 2, and leaves a blocking-pair rate two orders of magnitude above RAMT’s, so its point of mean stay-yield buys neither stability nor an account a case worker can check. Sinkhorn transport fills every seat but gives up 2.43 points of stay-yield (25.70 (3.69)) and a measurable blocking-pair rate; where seat coverage outranks institutional stability, it remains a defensible alternative.

Table 3. Ablation Study: Effect of Removing One Design Commitment at a Time (Mean (SD) over 20 Market Replications)

Variant	Stay-yield /100 offers	24 m retention	Blocking rate	Parity (high/low)
RAMT (full)	28.13 (1.57)	0.605 (0.010)	0.0005 (0.0006)	0.65 (0.13)
Boosted-tree scorer + deferred acceptance	29.13 (2.64)	0.595 (0.040)	0.0771 (0.0316)	0.65 (0.14)
Acceptance objective ($\alpha = 1$)	29.13 (1.01)	0.570 (0.008)	0.1260 (0.0283)	0.57 (0.05)
Neural scorer in place of additive splines	26.26 (2.06)	0.565 (0.023)	0.1128 (0.0332)	0.66 (0.10)
Greedy fill in place of deferred acceptance	19.97 (2.55)	0.547 (0.017)	0.1864 (0.0322)	1.01 (0.02)
No inverse-propensity weighting	27.34 (1.02)	0.595 (0.010)	0.0159 (0.0068)	0.78 (0.08)
Sinkhorn transport in place of deferred acceptance	25.70 (3.69)	0.594 (0.020)	0.1348 (0.0305)	0.92 (0.05)

D. Whom the Engine Leaves Behind

The distributional finding is the one a deploying county most needs to hear. Youth with college education and above are assigned at 0.65 times the rate of their less-educated peers under RAMT, and the pattern is no artefact of the proposed engine. It appears in every engine that allocates competitively on a learned objective — the acceptance-trained matcher shows it more strongly still (0.57) — and most starkly in the oracle itself (0.50), which allocates on the true retention probabilities; the greedy engines in Table 1 sit near parity only because they never let a predicted stayer outbid anyone for a scarce seat. Better prediction makes the abandonment sharper, not milder. The mechanism is visible in the ledger: village wages cluster well below graduate expectations, the hard threshold in the settling-in regime makes a large expectation shortfall the strongest single predictor of early exit, and the engine — correctly, on its own terms — concludes that graduates will not stay and stops spending scarce good posts on them. The graduates who are assigned receive the leftovers: their matches’ true retention averages 0.202 against 0.575 for everyone else. An optimiser told to maximise staying gives up

on exactly the complex-skill applicants the talent programme was created to attract — an equity failure that emerges when the system works, not when it breaks, and a concrete instance of the general concern that fairness in matching runs through the uncertainty of the predictions themselves [38].

Because the engine is a mechanism, the repair can be a mechanism too. Adding a single audited priority offset to the graduate stratum’s scores — visible in every affected ledger as its own entry — and raising it until assignment parity reaches 0.95 costs 1.45 points of stay-yield per 100 offers (from 28.13 to 26.69). That is the measured price of not abandoning the graduates, and the point of an auditable engine is that a county can see the price, debate it, and pay it deliberately — or fix the wage structure that generates it, which the ledger identifies just as plainly.

E. Two Dials: the Objective and the Right to Decide

Figure 4(a) traces the mixing dial of (9). Pure acceptance training ($\alpha = 1$) accepts 0.510 of offers but holds 0.571 of formed matches to 24 months; pure retention training ($\alpha = 0$) accepts 0.461 and holds 0.607. Stay-yield peaks at $\alpha = 0.50$ with 29.58 per 100 offers, because forming and keeping are both necessary and neither objective alone maximises their product. There is no engineering answer to where on this frontier a county should sit — a programme judged on annual placement counts and a programme judged on settled talent want different points — and the contribution of the dial is to make the choice explicit rather than buried in a loss function.

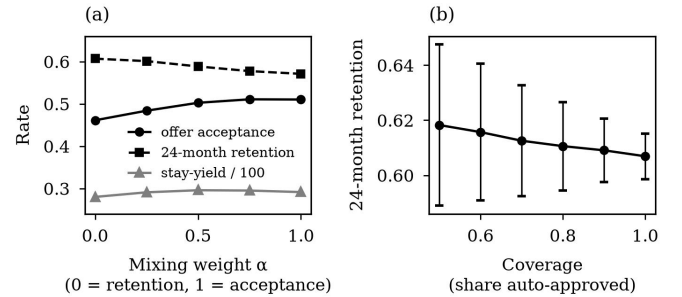


Figure 4. The two policy dials. (a) Acceptance, retention and stay-yield against the mixing weight α (means over 10 market replications). (b) 24-month retention among auto-approved matches against automated coverage, with one standard deviation across replications.

Figure 4(b) traces the abstention dial. Auto-approving the most certain half of matches yields 24-month retention of 0.618 among them, against 0.607 at full automation: uncertainty concentrates real risk, so routing the uncertain minority to a case worker buys retention on the automated remainder while keeping a human in exactly the cases where the evidence is thin — the division of labour the learning-to-defer literature argues for [22] and that employment-algorithm governance increasingly expects [9], [53].

F. Validity: A Mismatched Generator and a Noisy Profiler

Two checks probe whether the conclusions are artefacts of the simulator. First, the entire evaluation was re-run under the held-out generator of Section 5 — a single smooth logistic

hazard with none of the regime, threshold or interaction structure the engines could not represent. Table 4 shows the engine ordering is preserved essentially intact (rank correlation 1.00 across the seven engines), with the two interpretable deferred-acceptance engines again at the top of the learned pool; the acceptance-trained variant gains the most under the smooth generator, which removes the hard wage-expectation cliff that made early churn expensive, and the design conclusions — allocator first, objective as a policy dial, audits only from the ledger — do not depend on which wrong model of the world generated the data. Second, the generative-AI profiling layer was corrupted by replacing each extracted field, with probability rising to certainty, by the corresponding field of another randomly drawn profile — the plausible-but-wrong error an extraction pipeline actually makes. The response is a threshold, not a slope. Stay-yield is essentially flat through 40 per cent field error (28.42 clean against 27.86 at 0.4), because the additive scorer pools fifteen fields and deferred acceptance needs only the relative order of candidates to survive; but as corruption approaches totality RAMT falls to 21.89 — the level of the uncoordinated queue (22.43) — while the acceptance matcher barely moves (29.27 to 29.11), because acceptance is driven by administrative fields the profiler does not have to extract — distance, wage, housing — whereas retention lives in exactly the fields it must: skills, expectations, ties. The entire advantage of retention-aware matching is purchased by the extraction layer, and a county that does not audit the profiler with the same seriousness as the matcher forfeits precisely that advantage — a concrete reading of the transparency duties the generative-AI measures already impose [11].

Table 4. Stay-Yield per 100 Offers under the Regime-Switching Generator and the Held-Out Smooth Logistic Generator

Engine	Regime-switching generator	Smooth logistic generator
Admin. rule (wage rank)	6.48 (0.51)	7.99 (0.75)
First-come queue	22.64 (0.84)	27.51 (1.04)
Platform matcher [31]	21.59 (1.22)	26.15 (1.38)
Interpretable acceptance matcher	29.13 (1.01)	34.73 (0.88)
Boosted retention matcher	19.06 (2.67)	21.89 (2.85)
RAMT (ours)	28.13 (1.57)	32.23 (2.22)
Oracle (upper bound)	31.88 (1.17)	35.55 (0.86)

VII. DISCUSSION AND FUTURE WORK

A. What a County Talent Office Receives

The practical claim of the architecture is narrow and can be stated exactly. An office running RAMT receives, each matching round, a capacity-feasible offer list that no youth-post pair would jointly overturn; for every offer, a ledger whose entries sum to the decision and name the profile fields behind it; a short queue of low-margin cases with the reasons for the uncertainty attached; and one number — the position

on the acceptance-retention frontier — that its principals chose in the open. Nothing in that bundle requires trusting a model’s self-report: the ledger is the computation, the stability is checkable against the logs, and the frontier position is a published policy parameter. That is what auditable means here, and it is considerably less than trustworthy-by-assertion and considerably more than a SHAP plot appended to a black box.

B. The Direction of the Errors

The equity finding deserves restating as a general warning, because nothing about it is specific to this engine, this simulator or even this country: any allocator optimising a long-horizon outcome against a short-horizon budget will learn to abandon the applicants whose staying is expensive, and the abandonment will look, in every evaluation the optimiser is judged by, like performance. The graduates in this market are not mis-scored; they are correctly scored and therefore rationally deprioritised, and only an explicitly distributional audit surfaces it. The mandated bias audits now appearing in employment law [10], [53] look for disparate treatment of protected classes; the failure mode here — disparate abandonment of the policy’s own target population, induced by the objective — is not in their templates, and it should be.

C. From a Simulated Market to a County Programme

The corpus is simulated, and this is a genuine limitation as well as a deliberate methodological choice. What the simulation establishes is internal: that the architecture extracts most of the achievable retention from its logs, that its audits mean what they claim, that its known failure modes have measurable prices, and that these conclusions survive a structurally different generator. What it cannot establish is external: the anchored marginals do not make the market real, the acceptance and retention processes of any actual county will differ in ways no held-out generator anticipates, and the logged propensities that the off-policy correction consumes are clean here in a way administrative records rarely are. A deployment would begin exactly where the simulation ends: run shadowed alongside the incumbent rule for one or two annual rounds, log its counterfactual offers, and let the observed gap between predicted and realised retention calibrate both the hazard and the abstention threshold before any offer follows the engine. The upstream profiling layer, represented here by a noise channel, must in deployment be evaluated as its own component, with extraction accuracy measured against hand-labelled profiles; the noise sweep bounds how much its errors cost downstream. And consequential decisions about people carry obligations no architecture discharges: consent to profiling, contestability of ledger entries, and periodic distributional audits are governance work [9], [10], not engineering residue.

D. Extensions

Three extensions follow naturally. The first is dynamic: matching rounds repeat, retention outcomes arrive with delay, and the exploration a learning matcher needs sits in tension with the stability an institution needs — the bandit-matching

literature [37] meets long-horizon off-policy evaluation [60] exactly here. The second is the training side: the county controls not only who is matched but what training precedes the match, so the ledger’s skill-shortfall entries define targeted curricula, coupling this engine to the cultivate link of the talent chain. The third is richer preference elicitation: the youth side is represented here by a fitted acceptance model, and replacing it with elicited or LLM-assisted preference structures [32] — with the profiling caveats of Section 6 — would let the stability guarantee bind to preferences the young people actually state.

VIII. CONCLUSION

County talent programmes are beginning to allocate rural youth to village posts with learned matchers fed by generative-AI profiles, and the defaults they inherit from the recommendation stack — engagement objectives, opaque scorers, greedy allocation — are the wrong defaults for a public programme whose success is measured in years, not clicks. This paper built the opposite defaults into one engine: a pre-registered additive scorer whose evidence ledger reconstructs every decision exactly, a retention hazard trained off-policy from administrative logs, stable capacity-constrained assignment, decision-level audits, and abstention where the evidence is thin. On a simulated micro-market anchored to published statistics and governed by a deliberately mismatched ground truth, the engine converted 28.1 of every 100 offers into matches still in place at two years — 4.3 times the administrative wage-rank rule and 88.2 per cent of the oracle ceiling — while the audits that its ledger passes exposed the post-hoc trails of equally accurate black boxes as rationalisations.

The findings that matter beyond the engine are the two dials and the warning. Forming matches and keeping them are different objectives whose mixture is a policy choice, and the frontier between them should be chosen in public. Deciding automatically and deferring to a case worker are different acts whose boundary should follow the evidence, not the throughput target. And a retention optimiser working exactly as intended will quietly abandon the educated returnees a talent programme most wants, at a price the ledger makes visible and a county can choose to pay. Simulated evidence cannot certify a deployment; it can, and here does, establish that an allocation engine for the rural talent chain can be accurate, stable, honest about its reasons and honest about its failures at the same time — and that none of these properties needs to be traded for another.

REFERENCES

- [1] National Bureau of Statistics of China, “2024 年农民工监测调查报告 (2024 national monitoring survey report on rural migrant workers),” *Nat. Bur. Stat. China, Beijing*, Apr. 2025. [Online]. Available: https://www.stats.gov.cn/sj/zxfb/202504/t20250430_1959523.html
- [2] Y. Wang, Y. Jiang, B. Geng, B. Wu, L. Liao, “Determinants of returnees’ entrepreneurship in rural marginal China,” *Journal of Rural Studies*, vol. 94, pp. 429-438, 2022, doi: 10.1016/j.jrurstud.2022.07.014.
- [3] H. Ding, W. Dong, C. Qin, “Does migrant working experience stimulate returnees’ entrepreneurship: Evidence from rural China,” *Journal of Economic Behavior & Organization*, vol. 229, pp. 106865, 2025, doi: 10.1016/j.jebo.2024.106865.
- [4] Y. Lu, Y. Zhou, P. Liu, “Improving the entrepreneurial ability of rural migrant workers returning home in China: Study based on 5,675 questionnaires,” *Humanities and Social Sciences Communications*, vol. 10, pp. 150, 2023, doi: 10.1038/s41599-023-01663-5.
- [5] X. Liu, “Empowering rural revitalization in China with GAI: the application value, challenges, and optimization paths,” *Cogent Food & Agriculture*, vol. 11, pp. 2502389, 2025, doi: 10.1080/23311932.2025.2502389.
- [6] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, “Structured information extraction from scientific text with large language models,” *Nature Communications*, vol. 15, pp. 1418 (article number), 2024, doi: 10.1038/s41467-024-45563-x.
- [7] C. Qin, H. Zhu, T. Xu, C. Zhu, C. Ma, E. Chen, H. Xiong, “An Enhanced Neural Network Approach to Person-Job Fit in Talent Recruitment,” *ACM Transactions on Information Systems*, vol. 38, no. 2, pp. 1-33, 2020, doi: 10.1145/3376927.
- [8] Y. Mashayekhi, N. Li, B. Kang, J. Lijffijt, T. De Bie, “A Challenge-based Survey of E-recruitment Recommendation Systems,” *ACM Computing Surveys*, vol. 56, no. 10, pp. 1-33, 2024, doi: 10.1145/3659942.
- [9] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” *Official Journal of the European Union, L series*, vol. OJ L, 2024/1689, 2024.
- [10] L. Wright, R. M. Muenster, B. Vecchione, T. Qu, P. Cai, A. Smith, J. Metcalf, J. N. Matias, “Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability,” *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, Rio de Janeiro, Brazil, 2024, doi: 10.1145/3630106.3658998.
- [11] S. Migliorini, “China’s Interim Measures on Generative AI: Origin, Content and Significance,” *Computer Law & Security Review*, vol. 53, pp. 105985, 2024, doi: 10.1016/j.clsr.2024.105985.
- [12] Y. Lou, R. Caruana, J. Gehrke, “Intelligible models for classification and regression,” *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’12)*, pp. 150-158, 2012, doi: 10.1145/2339530.2339556.
- [13] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, G. Hinton, “Neural Additive Models: Interpretable Machine Learning with Neural Nets,” *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

- [14] M. T. Ribeiro, S. Singh, C. Guestrin, ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135-1144, 2016, doi: 10.1145/2939672.2939778.
- [15] S. M. Lundberg, S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4765-4774, 2017.
- [16] J. D. Singer, J. B. Willett, “It’s About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events,” *Journal of Educational Statistics*, vol. 18, no. 2, pp. 155-195, 1993, doi: 10.3102/10769986018002155.
- [17] A. Swaminathan, T. Joachims, “Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization,” *Journal of Machine Learning Research*, vol. 16, no. 52, pp. 1731-1755, 2015.
- [18] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, “Recommendations as Treatments: Debiasing Learning and Evaluation,” *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR, vol. 48, pp. 1670-1679, 2016.
- [19] D. Gale, L. S. Shapley, “College Admissions and the Stability of Marriage,” *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9-15, 1962, doi: 10.1080/00029890.1962.11989827.
- [20] A. E. Roth, “Deferred acceptance algorithms: history, theory, practice, and open questions,” *International Journal of Game Theory*, vol. 36, no. 3, pp. 537-569, 2008, doi: 10.1007/s00182-008-0117-6.
- [21] Y. Geifman, R. El-Yaniv, “Selective Classification for Deep Neural Networks,” *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, vol. 30, pp. 4878-4887, 2017.
- [22] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, D. Sontag, “Who Should Predict? Exact Algorithms For Learning to Defer to Humans,” *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, PMLR, vol. 206, pp. 10520-10545, 2023.
- [23] J. J. Horton, A. Filippas, B. S. Manning, “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?,” *NBER Working Paper No. 31122*, National Bureau of Economic Research, 2023, doi: 10.3386/w31122.
- [24] N. Li, C. Gao, M. Li, Y. Li, Q. Liao, “EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities,” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, pp. 15523-15536, 2024.
- [25] China Internet Network Information Center, “The 55th statistical report on China’s internet development,” *CNNIC, Beijing*, Jan. 2025. [Online]. Available: <https://www.cnnic.net.cn/n4/2025/0117/c88-11229.html>
- [26] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, “ERASER: A Benchmark to Evaluate Rationalized NLP Models,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4443-4458, 2020.
- [27] A. Ignatiev, N. Narodytska, J. Marques-Silva, “Abduction-based explanations for machine learning models,” in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 1511-1519, doi: 10.1609/aaai.v33i01.33011511.
- [28] Z. Liu, H. Wu, I. H. Badar, “Impact of Rural Digital Economy Development on Rural Revitalisation — Evidence from China,” *Sustainability*, vol. 16, no. 11, pp. 4573, 2024, doi: 10.3390/su16114573.
- [29] Y. Liu, M. Zhou, “Can rural e-commerce narrow the urban-rural income gap? Evidence from coverage of Taobao villages in China,” *China Agricultural Economic Review*, vol. 15, no. 3, pp. 580-603, 2023, doi: 10.1108/CAER-09-2022-0221.
- [30] E. Wiles, Z. Munyikwa, J. J. Horton, “Algorithmic Writing Assistance on Jobseekers’ Resumes Increases Hires,” *Management Science*, vol. 71, no. 12, pp. 10144-10164, 2025, doi: 10.1287/mnsc.2024.04528.
- [31] X. Yi, J. Yang, L. Hong, D. Z. Cheng, L. Heldt, A. Kumthekar, Z. Zhao, L. Wei, E. H. Chi, “Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations,” *Proc. 13th ACM Conference on Recommender Systems (RecSys)*, pp. 269-277, 2019, doi: 10.1145/3298689.3346996.
- [32] Y. Du, D. Luo, R. Yan, H. Liu, Y. Song, H. Zhu, J. Zhang, “Enhancing Job Recommendation through LLM-Based Generative Adversarial Networks,” *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-24)*, vol. 38, no. 8, 2024, doi: 10.1609/aaai.v38i8.28678.
- [33] K. Wilson, A. Caliskan, “Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval,” *Proc. 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, vol. 7, no. 1, pp. 1578-1590, 2024.
- [34] M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20)*, Barcelona, Spain, 2020, doi: 10.1145/3351095.3372828.
- [35] A. E. Roth, E. Peranson, “The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design,” *American Economic Review*, vol. 89, no. 4, pp. 748-780, 1999, doi: 10.1257/aer.89.4.748.
- [36] A. Abdulkadiroglu, T. Sonmez, “School Choice: A Mechanism Design Approach,” *American Economic Review*, vol. 93, no. 3, pp. 729-747, 2003, doi: 10.1257/00028280322157061.
- [37] M. Jagadeesan, A. Wei, Y. Wang, M. I. Jordan, J. Steinhardt, “Learning Equilibria in Matching Markets with Bandit Feedback,” *Journal of the ACM*, vol. 70, no. 3, 2023, doi: 10.1145/3583681.
- [38] S. Devic, D. Kempe, V. Sharan, A. Korolova, “Fairness in Matching under Uncertainty,” *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, PMLR, vol. 202, pp. 7775-7794, 2023.

- [39] M. Cuturi, “Sinkhorn Distances: Lightspeed Computation of Optimal Transport,” *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*, vol. 26, pp. 2292-2300, 2013.
- [40] G. Peyré and M. Cuturi, “Computational Optimal Transport: With Applications to Data Science,” *Foundations and Trends in Machine Learning*, vol. 11, no. 5-6, pp. 355-607, 2019, doi: 10.1561/22000000073.
- [41] R. Li, X. Ye, H. Zhou, and H. Zha, “Learning to Match via Inverse Optimal Transport,” *Journal of Machine Learning Research*, vol. 20, no. 80, pp. 1-37, 2019.
- [42] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019, doi: 10.1038/s42256-019-0048-x.
- [43] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, “Accurate intelligible models with pairwise interactions,” *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, pp. 623-631, 2013, doi: 10.1145/2487575.2487579.
- [44] G. I. Allen, L. Gan, L. Zheng, “Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities,” *Annual Review of Statistics and Its Application*, vol. 11, no. 1, pp. 97-121, 2024, doi: 10.1146/annurev-statistics-040120-030919.
- [45] S. Jain, B. C. Wallace, “Attention is not Explanation,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1*, pp. 3543-3556, 2019, doi: 10.18653/v1/N19-1357.
- [46] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, I. Augenstein, “Faithfulness Tests for Natural Language Explanations,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, pp. 283-294, 2023.
- [47] Q. Lyu, M. Apidianaki, C. Callison-Burch, “Towards Faithful Model Explanation in NLP: A Survey,” *Computational Linguistics*, vol. 50, no. 2, pp. 657-723, 2024, doi: 10.1162/coli_a_00511.
- [48] S. Wachter, B. Mittelstadt, C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841-887, 2018.
- [49] A. Singh, T. Joachims, “Fairness of Exposure in Rankings,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*, pp. 2219-2228, 2018, doi: 10.1145/3219819.3220088.
- [50] M. Zehlike, K. Yang, J. Stoyanovich, “Fairness in Ranking, Part I: Score-Based Ranking,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 118:1-118:36, 2022, doi: 10.1145/3533379.
- [51] C. Xu, X. Ye, W. Wang, L. Pang, J. Xu, and T.-S. Chua, “A Taxation Perspective for Fair Re-ranking,” *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, 2024, doi: 10.1145/3626772.3657766.
- [52] F. Hu, P. Ratz, and A. Charpentier, “A Sequentially Fair Mechanism for Multiple Sensitive Attributes,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, pp. 12502-12510, 2024, doi: 10.1609/aaai.v38i11.29143.
- [53] L. Groves, J. Metcalf, A. Kennedy, B. Vecchione, A. Strait, “Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime,” *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024, doi: 10.1145/3630106.3658959.
- [54] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187-202, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.
- [55] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Medical Research Methodology*, vol. 18, pp. article 24, 2018, doi: 10.1186/s12874-018-0482-1.
- [56] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, A. Bender, “Deep learning for survival analysis: a review,” *Artificial Intelligence Review*, vol. 57, pp. article 65, 2024, doi: 10.1007/s10462-023-10681-3.
- [57] M. Al Akasheh, E. F. Malik, O. Hujran, N. Zaki, “A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review,” *Expert Systems with Applications*, vol. 238, pp. article 121794, 2024, doi: 10.1016/j.eswa.2023.121794.
- [58] M. Dudík, D. Erhan, J. Langford, L. Li, “Doubly Robust Policy Evaluation and Optimization,” *Statistical Science*, vol. 29, no. 4, pp. 485-511, 2014, doi: 10.1214/14-STS500.
- [59] S. Athey, S. Wager, “Policy Learning With Observational Data,” *Econometrica*, vol. 89, no. 1, pp. 133-161, 2021, doi: 10.3982/ECTA15732.
- [60] Y. Saito, H. Abdollahpouri, J. Anderton, B. Carterette, M. Lalmas, “Long-term Off-Policy Evaluation and Learning,” *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pp. 3432-3443, 2024, doi: 10.1145/3589334.3645446.
- [61] K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, J. Davis, “Machine learning with a reject option: A survey,” *Machine Learning*, vol. 113, no. 5, pp. 3073-3110, 2024, doi: 10.1007/s10994-024-06534-x.
- [62] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, “On Calibration of Modern Neural Networks,” *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, PMLR, vol. 70, pp. 1321-1330, 2017.
- [63] A. N. Angelopoulos, S. Bates, “Conformal Prediction: A Gentle Introduction,” *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494-591, 2023, doi: 10.1561/22000000101.