

CIVIC-IT: A Causal-Invariant-Conformal Framework For Robust HR Analytics In Heterogeneous IT Environments

Amit Nagpal^{1*} and Sapna Rana¹

¹School of Management & Commerce, K.R. Mangalam University, Sohna, Gurgaon, 122103, Haryana, India.

*Corresponding author(s): Amit Nagpal^{1*} and Sapna Rana¹

Abstract

Human-resource (HR) models often fail when deployed across heterogeneous organizational units because correlations are confounded by selection and degrade under environment shift. We present CIVIC-IT, a Causal-Invariant-Conformal framework that jointly targets identification, transfer, and calibrated uncertainty for HR decision making. First, we estimate average and heterogeneous effects of environment-level diversity levers using cross-fitted double/debiased machine learning with cluster-robust inference. Second, we learn shift-robust representations via invariant risk minimization (IRM) combined with a domain-adversarial objective (DANN) and fine-tune with Group-DRO to control worst-environment risk. Third, we attach distribution-free guarantees using split-conformal prediction for continuous outcomes and label-conditional conformal predictive sets for a “high T*” (turnover intention) flag. Outcomes are psychometrically grounded latent scores for Inclusion, Engagement, and T* built from ordinal indicators with reliability weighting and multi-group diagnostics for approximate invariance. Using the IBM HR Attrition dataset (N=1,470) configured into realistic environments (Department, JobRole, and Department×Role), we find: (i) global ATEs are small, but policy-relevant heterogeneity emerges—e.g., for early-career Research Scientist teams, greater diversity of education field increases Inclusion/Engagement, whereas highly heterogeneous business-travel patterns reduce them; (ii) leave-one-environment-out evaluation identifies stable transfer with clear best/worst environments and improved worst-group error under Group-DRO; (iii) conformal intervals achieve near-nominal coverage, and the risk classifier yields predominantly singleton predictive sets at the target error rate; and (iv) multi-objective policy search produces fairness-aware Pareto portfolios that deliver concurrent (albeit modest) gains in Inclusion and Engagement with slight reductions in T*. CIVIC-IT offers a governance-ready blueprint that links robust estimation to auditable, deployable HR policy in heterogeneous IT environments. Keywords: HR analytics; causal inference; invariant learning; domain adaptation; conformal prediction; policy optimization.

1 Introduction

Organizations in information technology (IT) operate across heterogeneous environments—business units, departments, and job roles—that differ in skill composition, managerial practices, and demographic mix. These differences complicate human-resource (HR) analytics because models trained in one environment often degrade when deployed in another (i.e., under distribution shift), and descriptive correlations can be confounded by selection into jobs, tenure, or compensation. For

high-stakes outcomes such as inclusion, engagement, and turnover intention (T^*), decision makers need estimates that are (i) causally interpretable under credible assumptions, (ii) robust to environment shifts, and (iii) accompanied by valid uncertainty quantification. Classic predictive attrition models rarely meet all three requirements simultaneously, leading to fragile interventions or fairness and governance concerns when deployed at scale.

We address these gaps by integrating three complementary strands of methodology. First, we use double/debiased machine learning (DML) to estimate average and heterogeneous effects while controlling flexibly for observed covariates via cross-fitting; DML yields orthogonalized estimators that are less sensitive to high-dimensional nuisance modeling and regularization choices (Chernozhukov et al., 2018). Second, to protect against environment shift, we learn representations that transfer across environments using invariant risk minimization (IRM) and domain-adversarial training (DANN). IRM promotes predictors whose optimality conditions align across environments, aiming for causal features that remain stable under interventions (Arjovsky, Bottou, Gulrajani, & Lopez-Paz, 2019). DANN augments the representation with an adversarial domain classifier to reduce environment-specific information that harms generalization (Ganin et al., 2016). Third, we calibrate uncertainty with conformal prediction, which provides finite-sample, distribution-free coverage guarantees for regression (and predictive sets for classification) after a data-splitting calibration step (Balasubramanian, Ho, & Vovk, 2014; Lei, G'Sell, Rinaldo, Tibshirani, & Wasserman, 2018). Together, these components target actionable HR questions under realistic deployment constraints.

We propose CIVIC-IT—a Causal-Invariant-Conformal framework for robust HR analytics in heterogeneous IT environments—and make the following contributions:

1. Unifies causal identification with out-of-distribution robustness for HR decisions across heterogeneous organizational environments. Treatments are formalized as environment-level diversity levers (e.g., Blau/entropy indices and dispersion measures), with outcomes defined as latent Inclusion, Engagement, and T^* . Cross-fitted partially linear DML with cluster-robust standard errors delivers ATEs and policy-relevant CATEs for tenure, income, and role subgroups (Chernozhukov et al., 2018).
2. Establishes psychometrically grounded measurement for HR constructs by deriving reliability-weighted latent scores from ordinal indicators and attaching Bayesian-bootstrap confidence intervals; multi-group diagnostics assess measurement invariance across departments and job roles, enabling valid between-environment comparisons.
3. Advances shift-robust HR analytics by combining invariant risk minimization (IRM) and domain-adversarial training (DANN) into a representation that transfers across environments, and evaluating generalization with leave-one-environment-out protocols; worst-group risk is further reduced via group distributionally robust optimization (Group-DRO) (Arjovsky et al., 2019; Ganin et al., 2016; Sagawa, Koh, Hashimoto, & Liang, 2019).
4. Delivers calibrated uncertainty for decision support using split-conformal predic-

tion: distribution-free interval estimates for continuous outcomes and predictive sets for “high T*” classification, providing finite-sample coverage guarantees under mild assumptions (Balasubramanian et al., 2014; Lei et al., 2018).

5. Translates empirical effects into implementable policy by casting diversity levers into a multi-objective optimization that traces the Pareto frontier—maximizing gains in Inclusion and Engagement while minimizing T* and inter-department disparities—and selecting fairness-aware exemplar portfolios for deployment and monitoring.
6. Codifies a governance-ready analytics blueprint for enterprise HR: a pre-analysis plan that registers hypotheses and environments, reproducibility controls (seeds and deterministic splits), and auditing modules (ablations, sensitivity to unobserved confounding, and equalized-odds/opportunity post-processing) to support compliant, transparent decision-making.

The paper begins with Section 2 reviewing related work on HR analytics, causal inference with observational HR data, out-of-distribution generalization (IRM, DANN, Group-DRO), and uncertainty quantification via conformal prediction. Section 3 details data governance, environments, diversity indices, measurement modeling, causal identification (DML), invariant representation learning (IRM/DANN), conformal calibration, and policy optimization. Section 4 reports descriptives and diversity maps, latent-score distributions and invariance checks, ATE/CATE estimates, out-of-distribution (OOD) performance via leave-one-environment-out evaluation, conformal coverage and set sizes, policy frontiers with exemplar portfolios, and robustness/- fairness audits. ?? interprets implications for HR policy and ethics, and Section 6 summarizes limitations and future work.

2 Related Work and Background

2.1 HR analytics and attrition modeling

Research on turnover and its antecedents spans organizational behavior, labor economics, and information systems. While classic work emphasizes attitudinal and contextual drivers (e.g., job satisfaction, commitment, compensation, and alternative opportunities), recent HR analytics studies increasingly rely on supervised learning over large administrative or survey corpora to predict attrition or turnover intention and to prioritize drivers for intervention. For instance, Lazzari, Alvarez, and Ruggieri (2022) evaluate a suite of classifiers (logistic regression, gradient boosting, TabNet) on a cross-national engagement survey and report competitive performance for LightGBM alongside interpretable baselines; they also combine feature importance with structural causal modeling to move beyond purely correlational explanations. Such pipelines underscore two recurring practical challenges: (i) systematic class imbalance (rare turnover events), and (ii) variation in feature–outcome relationships across business units, countries, or job families that can degrade out-of-sample reliability. These challenges motivate a stronger connection to causal identification (to mitigate confounding) and to out-of-distribution (OOD) generalization (to mitigate environment shift).

2.2 Causal inference with observational HR data

Much HR data are observational, making confounding a central concern for effect estimation. Propensity-score designs (matching, weighting) remain widely used to reduce imbalance in measured covariates; careful construction of inverse-probability weights and diagnostics for extreme weights are important to avoid instability (Austin & Stuart, 2015). Machine-learning-based causal estimators further improve robustness by orthogonalizing the target parameter against high-dimensional nuisance models. Double/debiased machine learning constructs Neyman-orthogonal scores and combines them with cross-fitting to yield asymptotically normal ATE/CATE estimators under mild regularity conditions and flexible learners for propensity and outcome models (Chernozhukov et al., 2018). For heterogeneous effects, generalized random forests provide nonparametric estimation and inference for treatment heterogeneity with valid uncertainty quantification (Wager & Athey, 2018). When outcomes are psychometric constructs, comparability across organizational subgroups requires at least approximate measurement invariance (Meredith, 1993). Together, these strands set a foundation for credible, policy-relevant effect estimation in HR settings.

2.3 Out-of-distribution generalization (IRM, DANN, Group-DRO)

Deployment often entails covariate and conditional-shift across “environments” (departments, roles, locations). Generalization theory for domain adaptation shows that target error can be bounded by source error plus a divergence term measuring domain discrepancy (Ben-David et al., 2010), motivating representation learning that suppresses spurious, environment-specific variation. Domain-adversarial neural networks (DANN) add an adversarial domain classifier to encourage environment-invariant representations while preserving label information (Ganin et al., 2016). Invariant risk minimization (IRM) promotes predictors whose optimality conditions align across environments, seeking features with stable causal relationships across shifts (Arjovsky et al., 2019). Complementing these, group distributionally robust optimization (Group-DRO) directly optimizes worst-group risk, improving performance on minority or outlying environments relative to empirical risk minimization (Sagawa et al., 2019). These approaches are particularly relevant to enterprise HR, where policy makers must anticipate performance under unobserved re-mixes of teams and roles.

2.4 Uncertainty quantification via conformal prediction

Actionable HR recommendations benefit from calibrated uncertainty. Conformal prediction furnishes distribution-free prediction intervals (regression) and predictive sets (classification) with finite-sample coverage under exchangeability by calibrating on held-out nonconformity scores (Vovk, Gammerman, & Shafer, 2005). Split-conformal procedures deliver simple, computationally efficient coverage guarantees and can be layered atop any base model; when applied to continuous HR outcomes (e.g., engagement indices) they provide marginal coverage, and for binary decisions (e.g., “high” turnover intention) they yield set-valued predictions with controllable size-coverage trade-offs (Lei et al., 2018). This methodology aligns with governance expectations in

high-stakes applications by making uncertainty explicit and auditable.

3 Materials and Methods

This section details the end-to-end CIVIC-IT pipeline and the specific design choices that make it suitable for heterogeneous IT settings. We begin with data governance and the construction of environment identifiers and diversity levers; we then build psychometrically grounded latent outcomes for Inclusion, Engagement, and turnover intention (T^*) with reliability-weighted scoring and Bayesian-bootstrap uncertainty. Next, we estimate orthogonalized causal effects using cross-fitted partially linear DML with cluster-robust inference and prespecified heterogeneous effects (CATEs). To support deployment under distribution shift, we learn an invariant representation that combines IRM with a domain adversary (DANN) and fine-tunes via Group-DRO, evaluated with leave-one-environment-out (LOEO). Finally, we attach distribution-free uncertainty via split-conformal regression and label-conditional conformal classification, and translate effects into fairness-aware policy portfolios. Formal algorithms appear in **Algorithms 1 to 3 and 5 to 8**.

Figure 1 summarizes the workflow in five linked stages: (A) Data & Governance assembles inputs, fixes environments (Department, JobRole $\times$ Dept Role), and computes diversity indices (Blau/entropy; SD/IQR; joint entropies) under a preregistered plan. (B) Measurement derives latent outcomes with reliability-weighted PC scores, Bayesian-bootstrap CIs, and multi-group invariance diagnostics. (C) Causal DML delivers ATEs and subgroup CATEs by orthogonalizing Y and D against flexible nuisances and using cluster-robust SEs at Dept Role. (D) Invariant Representation trains a shared encoder with IRM penalty and a gradient-reversal domain head

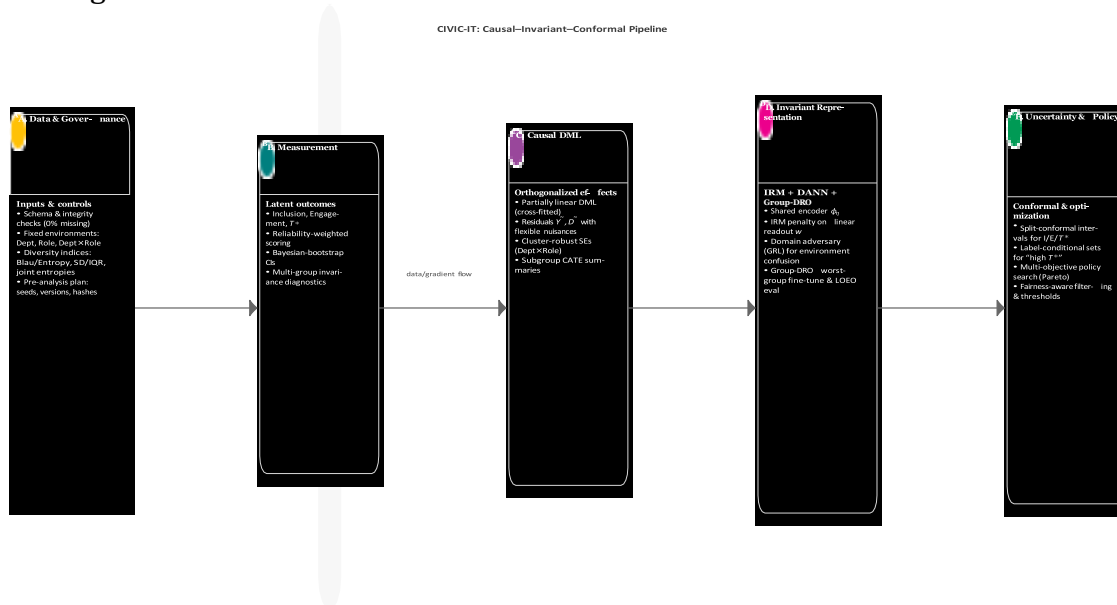


Fig. 1: CIVIC-IT methodology overview. Governance → measurement → causal DML invariant representation (IRM+DANN with Group-DRO) → conformal calibration and policy optimization. Boxes list the exact steps implemented in Section 3; arrows indicate data/gradient flow.

(DANN), then applies Group-DRO and LOEO evaluation to target worst-environment risk. **(E) Uncertainty & Policy** adds split-conformal intervals for (I, E, T*), label-conditional predictive sets for a “high T*” flag, and a multi-objective, fairness-aware policy search that selects deployable lever portfolios.

Together, these components connect credible effect identification to shift-robust prediction and auditable uncertainty, yielding governance-ready HR policy recommendations for heterogeneous IT environments.

3.1 Data and pre-analysis plan

3.1.1 Dataset, schema checks, and governance

The analysis uses the widely studied IBM HR Attrition dataset comprising 1,470 unique employee records and 35 canonical HR variables covering demo- graphics, compensation, tenure, performance and satisfaction items, along with three environment identifiers engineered for this study (Department, JobRole, and Department Role). Schema validation confirmed zero missingness across all fields and no duplicate EmployeeNumber entries, so the curated sample equals the raw intake ($N_{\text{raw}}=1,470$, $N_{\text{final}}=1,470$; Figure 3, Table 1). Variable types are dominated by continuous double features with a smaller mix of logical, string, and categorical fields; the observed mix is summarized in Figure 2. Two constants (EmployeeCount, StandardHours) were retained for auditability but excluded from modeling steps that require variation. Likert-type constructs (EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, PerformanceRating) appear on a 1–4 scale and passed range checks without violations. All schema checks (type conformity, cardinality, numeric range sanity, and domain label cleanliness) produced no exceptions; guardrails for imputation and de-duplication remained idle but reproducibly logged.

Pre-analysis governance fixed environments at three granularities to reflect operational deployment contexts: Department (3 levels), JobRole (9), and Department Role (11). These identifiers serve two purposes: (i) to define the environments across which causal estimands and representation learning must remain stable and (ii) to aggregate diversity statistics that act as policy levers. Following ingestion, diversity features were computed per environment and mapped back to individuals: Blau and entropy indices for categorical facets (Gender, EducationField, BusinessTravel, MaritalStatus, JobRole), dispersion for continuous facets (standard deviation and interquartile range of Age, TotalWorkingYears, YearsAtCompany, MonthlyIncome, DistanceFromHome), and pairwise joint entropies for intersectional views (e.g., Gender EducationField).

Additional analysis-ready variables include TrainingQ (quartiles of training count), OverTimeLikert (overtime recoded to a 1–4 orientation for strain), z-scores for MonthlyIncome and YearsAtCompany, and three latent outcomes (Inclusion, Engagement, and T *) constructed downstream from ordinal indicators. A compact view of the curated schema—including type, cardinality, and missingness—is provided in **Tables 2 and 3**.

The data-governance flow illustrates that all records passed schema and integrity checks, resulting in a one-to-one carryover from raw to curated tables (**Figure 3**). This step also registers the pre-analysis plan (frozen random seeds, toolbox versions, and file hashes), ensuring that all later estimates and figures are exactly reproducible under redeployment or audit.

Variable Types (Data Schema)

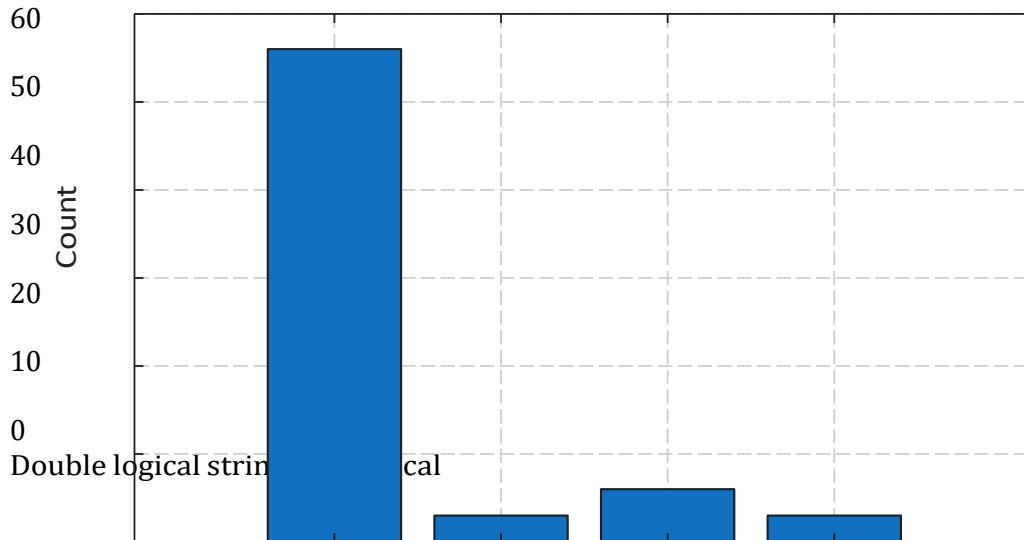


Fig. 2: Variable types (schema summary). Count of variables by storage/type after ingestion; continuous double features dominate, with a smaller share of logical, string, and categorical fields. This mix motivates flexible machine learning nuisance models and careful treatment of ordinal items in measurement modeling.

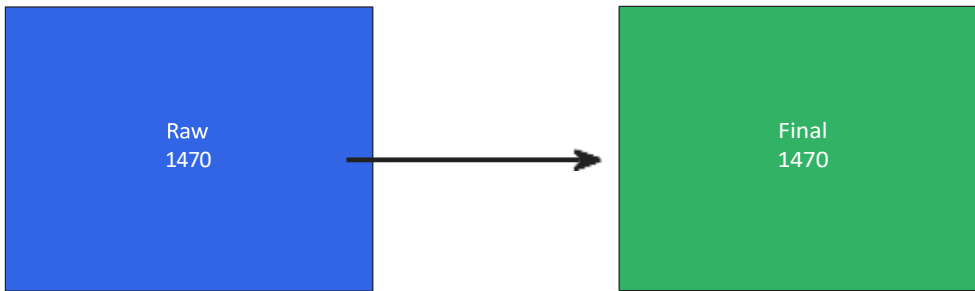


Fig. 3: Governance flow from raw to curated data. End-to-end integrity checks (schema, ranges, de-duplication) produced no removals ($N_{\text{raw}}=1,470$ $N_{\text{final}}=1,470$). The pre-analysis plan logs seeds, versions, and hashes to support reproducible reporting and audit.

Table 1: Governance summary for data curation. All rows passed schema and integrity checks.

Nraw	Nfinal	Nremoved
Employee records	1,470	0
1,470		

Table 2: Schema: core variables and integrity checks. Type, cardinality (unique values), and percent missing for original HR fields. All variables show 0% missingness.

Variable	Type	Unique	Missing (%)	Variable	Type	Unique	Missing (%)
Age	double	43	0	MonthlyRate	double	1427	0
Attrition	logical	2	0	NumCompaniesWorked	double	10	0
BusinessTravel	string	3	0	Over18	logical	1	0
DailyRate	double	886	0	OverTime	logical	2	0
Department	string	3	0	PercentSalaryHike	double	15	0
DistanceFromHome	double	29	0	PerformanceRating	double	2	0

Education	doub 5	0	RelationshipSatisfac	doub 4	0
EducationField	string 6	0	StandardHours	doub 1	0
EmployeeCount	doub 1	0	StockOptionLevel	doub 4	0
EmployeeNumber	doub 1470	0	TotalWorkingYears	doub 40	0
EnvironmentSatisfaction	doub 4	0	TrainingTimesLastYear	doub 7	0
Gender	string 2	0	WorkLifeBalance	doub 4	0
HourlyRate	doub 71	0	YearsAtCompany	doub 37	0
JobInvolvement	doub 4	0	YearsInCurrentRole	doub 19	0
JobLevel	doub 5	0	YearsSinceLastPromotion	doub 16	0
JobRole	string 9	0	YearsWithCurrentManager	doub 18	0
JobSatisfaction	doub 4	0	MonthlyIncome	doub 1349	0
MaritalStatus	string 3	0			

Interpretation of the schema and plots motivates key design choices for subsequent stages. The dominance of continuous variables suggests that flexible, tree-based nuisance learners can capture interactions and non-linearities during cross-fitted orthogonalization without extensive feature engineering. The presence of high-cardinality pay

Table 3: Schema: engineered environments, diversity indices, analysis-ready transforms, and latent outcomes. All fields are complete (0% missing).

Variable	Type	Uniq ue	Missin g (%)	Variable	Typ e	Uniq ue	Missin g (%)
Env Department	categorical	3	0	SD MonthlyIncome	doub	11	0
Env JobRole	categorical	9	0	SD TotalWorkingYears	doub	11	0
Env DeptRole	categorical	11	0	SD YearsAtCompany	doub	11	0

Blau	doubl	11	0	mpany		
BusinessTravel	e			TrainingQ	dou 4	0
Blau	doubl	11	0	OverTimeLi	dou 2	0
EducationField	e			kert	ble	
Blau Gender	doubl	11	0	MonthlyInco	dou 13490	
	e			me z	ble	
Blau JobRole	doubl	1	0	YearsAtCom	dou 37	0
	e			pany z	ble	
Blau MaritalStatus	doubl	11	0	Inclusion	dou 64	0
	e				ble	
Ent BusinessTravel	doubl	11	0	Engagement	dou 60	0
	e				ble	
Ent EducationField	doubl	11	0	Tstar	dou 14690	
	e				ble	
Ent Gender	doubl	11	0	IQR Age	dou 7	0
	e				ble	
Ent JobRole	doubl	1	0	IQR	dou 10	0
	e			DistanceFro	ble	
				mHome		
Ent MaritalStatus	doubl	11	0	IQR	dou 11	0
	e			MonthlyInc	ble	
				ome		
JointEnt	doubl	11	0	IQR	dou 8	0
EducationField	e			TotalWorki	ble	
BusinessTravel				ngYears		
JointEnt Gender	doubl	11	0	IQR	dou 7	0
EducationField	e			YearsAtCo	ble	
				mpany		
JointEnt Gender	doubl	11	0	SD Age	dou 11	0
MaritalStatus	e				ble	
SD	doubl	11	0			
DistanceFromHome	e					

and tenure measures (MonthlyIncome, YearsAtCompany) justifies z-standardization to stabilize latent measurement and treatment-response modeling. Environment identifiers with modest cardinality (3, 9, and 11 levels) are suitable for defining training environments in invariant representation learning and for cluster-robust variance estimation in causal analyses. Finally, the complete absence of missingness concentrates governance effort on shift-robustness and fairness auditing rather than imputation sensitivity, while the frozen pre-analysis plan ensures that all descriptive and causal results downstream can be exactly traced back to the curated inputs reported in Tables 2 and 3.

3.1.2 Environments E: Department, JobRole, Dept×Role

Organizational environments are fixed at three granularities to reflect the loci at which HR policy and managerial practices typically operate: (i) Department (E_{dept}), (ii) JobRole (E_{role}), and (iii) the fine-grained interaction $\text{Dept} \times \text{Role}$ ($E_{\text{dept} \times \text{role}}$). The department stratum is dominated by Research & Development (R&D; 961 employees, 65.37%), followed by Sales (446; 30.34%) and a smaller Human Resources unit (63; 4.29%). This size asymmetry is material for both causal estimation (clustered uncertainty by environment) and shift-robust learning (worst-group performance), and it motivates explicit reporting of per-environment N and dispersion. Counts and shares are in Table 4 and visualized in Fig. 4.

Job-role composition is heterogeneous and right-tailed, with Sales Executive (326; 22.18%), Research Scientist (292; 19.86%), and Laboratory Technician (259; 17.62%) comprising nearly 60% of the workforce. Managerial and specialist roles form smaller strata (Manager 102; 6.94%, Research Director 80; 5.44%), while HR and field-sales roles are minority groups (Human Resources 52; 3.54%, Sales Representative 83; 5.65%). Role frequencies are tabulated in Table 5 and plotted using a space-efficient treemap in Fig. 5. These distributions anticipate heterogeneous precision across subgroups and inform the minimum cell-size checks used later for CATE reporting and group-robust training.

The cross-classification of departments by roles (**Fig. 6**) shows pronounced structural zeros that are endemic to real organizations (e.g., research roles appear almost exclusively within R&D; sales roles cluster within Sales; HR roles within HR). The resulting environment set $E_{\text{dept} \times \text{role}}$ contains 11 nonempty combinations (from a theoretical 3×9), with cell sizes ranging from $N=11$ (e.g., niche roles in small departments) to $N=326$ (Sales Executives in Sales). This sparsity clarifies the identification and generalization regimes: (a) cluster-robust standard errors in DML use $E_{\text{dept} \times \text{role}}$ as the clustering unit; (b) invariant representation learning (IRM/DANN) treats each nonempty combination as a distinct “environment” for training and leave-one-environment-out evaluation; and (c) fairness auditing evaluates thresholds within demographically intersected subgroups while respecting these organizational strata.

Table 4: Department composition (E_{dept}). Counts (N) and shares.

Department	N	Percent (%)
Human Resources	63	4.29
Research Development	961	65.37
Sales	446	30.34
Total	1470	100.00

Table 5: Job-role composition (E_{role}). Counts (N) and shares.

JobRole	N	Percent (%)
Healthcare Representative	131	8.91
Human Resources	52	3.54
Laboratory Technician	259	17.62
Manager	102	6.94
Manufacturing Director	145	9.86
Research Director	80	5.44
Research Scientist	292	19.86
Sales Executive	326	22.18
Sales Representative	83	5.65
Total	1470	100.00

Visualization of the department stratum in Fig. 4 confirms the dominance of R&D and highlights the comparatively small HR cluster. The role treemap in Fig. 5 offers an at-a-glance view of role imbalance and is cited alongside the CATE summaries to contextualize precision differences across roles. The contingency heatmap in Fig. 6 is included near the beginning of Section 4 to motivate the leave-one-environment-out design and to illustrate where representation learning must bridge sharp distributional boundaries.

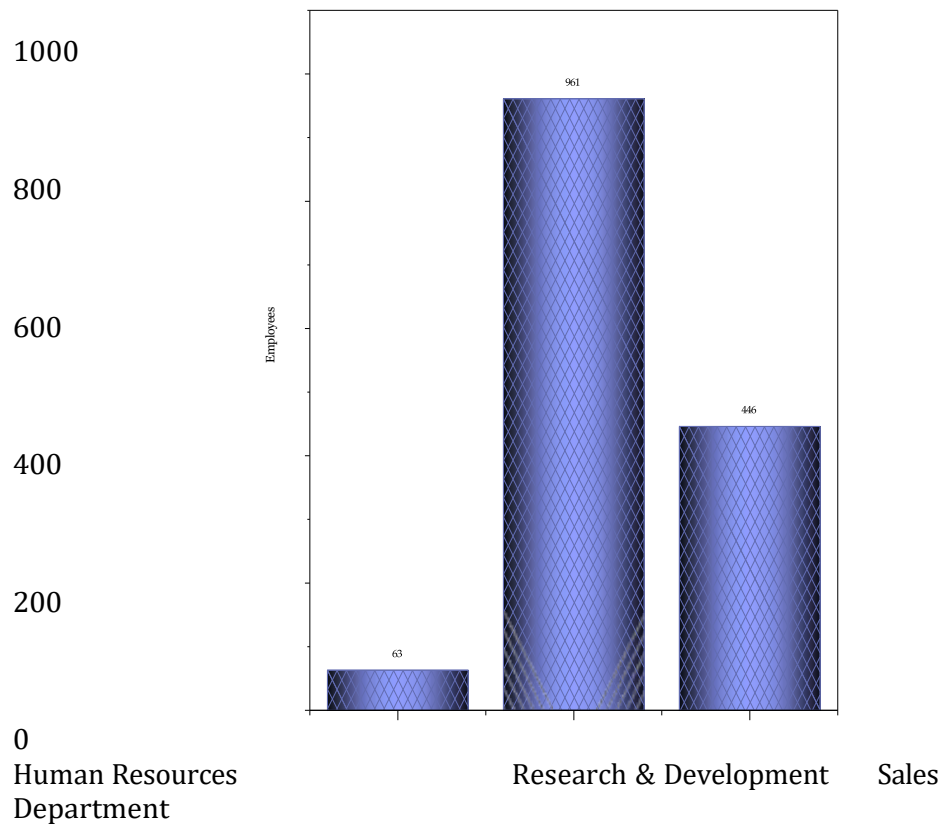


Fig. 4: Department size distribution (E_{dept}). R&D comprises roughly two-thirds of the workforce, with Sales about one-third and HR a small minority. This asymmetry motivates clustered uncertainty and worst-group evaluation.

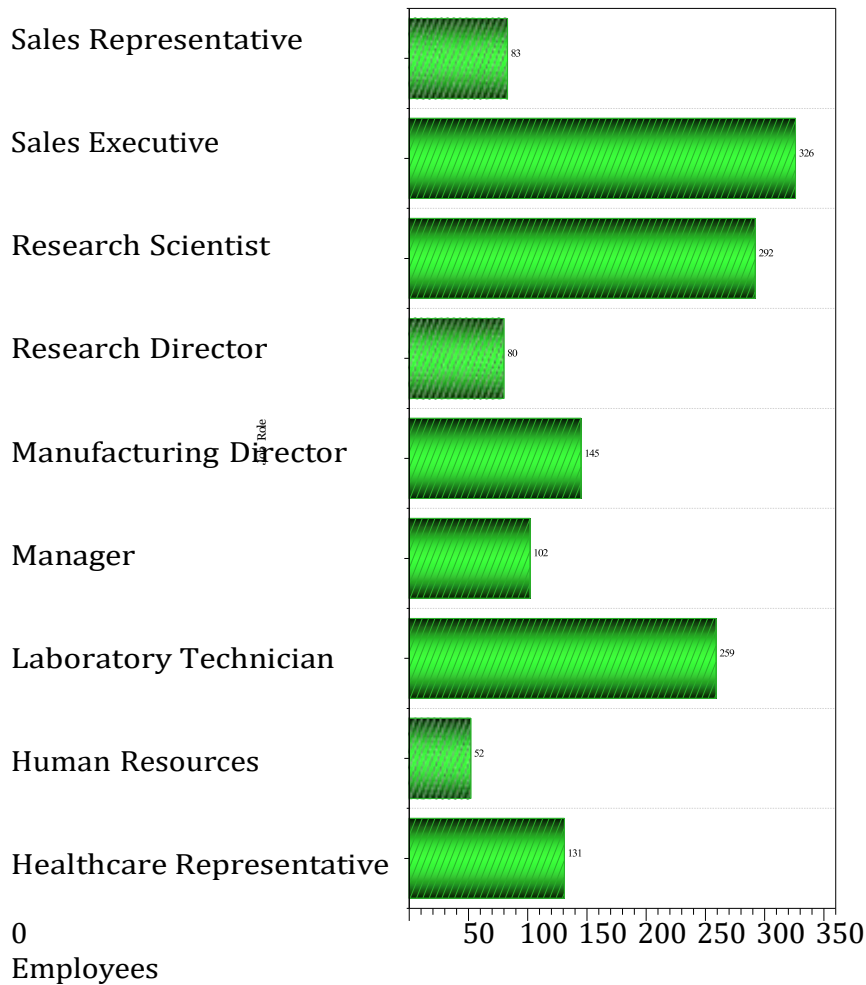


Fig. 5: Job-role composition (E_{role}) as a treemap. Sales Executive, Research Scientist, and Laboratory Technician account for the majority of employees; specialized leadership roles are comparatively small.

3.2 Measurement modeling of latent outcomes

Latent outcomes target three constructs that are central to HR decision making in technology organizations: Inclusion, Engagement, and turnover intention (T^*). Inclusion captures the extent to which employees experience belongingness together with



Fig. 6: Dept Role contingency heatmap ($E_{dept \times role}$). Bright cells indicate larger strata; dark/blank cells denote structural zeros (roles not present in a department). These nonoverlapping supports create natural domains for OOD evaluation (LOEO).

recognition of uniqueness within their units; the construct is widely discussed in organizational behavior as a complement to diversity policy (Shore et al., 2011). Engagement reflects a positive, work-related state characterized by vigor, dedication, and absorption, and is typically operationalized with short reflective indicators when long inventories are unavailable (Schaufeli & Bakker, 2004; Schaufeli, Bakker, & Salanova, 2006). Turnover intention is modeled as the proximal antecedent of actual turnover in contemporary turnover process theories and has been shown to encapsulate the joint influence of satisfaction, alternative opportunities, and embeddedness (Steel & Lounsbury, 2009). In the present design, Inclusion is indicated by EnvironmentSatisfaction, RelationshipSatisfaction, and WorkLifeBalance; Engagement by JobInvolvement, JobSatisfaction, and an ordinalized training frequency; and T* by an overtime pressure proxy, WorkLifeBalance, JobInvolvement, JobSatisfaction, and standardized tenure and income controls. Indicators are ordinal (four-point Likert for most items), so rank-based and polyserial-consistent procedures are used to avoid assumptions of interval spacing (Agresti, 2010). Scores are oriented such that larger values imply more Inclusion and Engagement, while larger T* indicates stronger intention to quit.

3.2.1 Constructs: Inclusion, Engagement, Turnover Intention (T*)

Construct definition follows a congeneric reflective view in which each indicator is an imperfect, noisy function of a common latent trait. To reduce construct contamination and avoid post-treatment bias in subsequent causal analyses, indicators that are downstream of environment-level treatments are excluded when definitional overlap is strong. For T*, direct binary attrition is not used as an indicator to prevent tautological modeling; instead, theoretically proximal proxies known to antecede quitting decisions are preferred (Steel & Lounsbury, 2009). All indicators are preprocessed as ordered variables; continuous covariates that serve as indicators in T* are standardized to unit variance to avoid dominance by scale.

3.2.2 Reliability-weighted latent scores with Bayesian-bootstrap CIs

Reliability is assessed using internal-consistency coefficients computed on rank correlations. With k indicators and average off-diagonal Spearman correlation $\bar{r}$, coefficient α is estimated as

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \quad (1)$$

and complemented by McDonald's ω_t from a unidimensional congeneric model approximated by the first principal component (PC) loading pattern (Cronbach, 1951; McDonald, 2013; Revelle & Zinbarg, 2009). Items are auto-oriented to align with the latent trait by correlating each item with the proxy total and flipping signs when necessary. A light-touch item screening retains indicators whose item-total Spearman correlation exceeds a small threshold (e.g., 0.10) to guard against misfitting items without forcing strict tau-equivalence.

Latent scores are then computed as reliability-weighted PC composites on the rank-correlation-implied covariance matrix. Let $Z \in \mathbb{R}^{n \times k}$ denote the z-scored indicator matrix after orientation, and let $\bar{w} \in \mathbb{R}^k$ be the normalized first PC loadings; the individual score is $s_i = Z_i \bar{w}$. To quantify estimator uncertainty without parametric assumptions, a Bayesian bootstrap is employed: resampling weights are drawn from a Dirichlet(1) distribution to reweight rows of Z , PC weights are recomputed for each draw, and the induced score distribution yields pointwise $(1 - \alpha)$ intervals for each individual (Rosenbaum & Rubin, 1983). This approach preserves the ordinal nature of indicators, propagates weight uncertainty into scores, and produces finite-sample intervals that are robust to mild violations of normality. Reliability-weighted latent scores with Bayesian bootstrap are computed (see **Algorithm 1**).

Algorithm 1 Reliability-weighted latent scoring with Bayesian bootstrap

Require: Indicator matrix X (ordinal/continuous), number of bootstrap draws B 1: Convert ordinal indicators to numeric ranks and z-score columns to obtain Z 2:

Auto-orient items by the sign of Spearman correlation with the row mean of Z 3:
 Optionally drop items with item-total Spearman < 0.10
 4: Compute Spearman correlation R and derive α and ω_t from R
 5: Obtain first PC loadings w from R ; normalize by $|w|_1$
 6: Compute baseline scores $s = Zw$
 7: for $b = 1$ to B do
 8: Draw Bayesian-bootstrap weights $p^{(b)} \sim \text{Dirichlet}(1)$
 9: Form weighted correlation $R^{(b)}$ and recompute $w^{(b)}$ from $R^{(b)}$
 10: Store $s^{(b)} = Zw^{(b)}$
 11: end for
 12: For each individual i , form CI from empirical quantiles of $\{s^{(b)}\}_B$
 Ensure: Scores s , confidence intervals, reliability diagnostics (α, ω_t)

This estimator trades minimal modeling commitments for psychometric discipline. Internal consistency is reported alongside score intervals, and all downstream analyses consume point estimates with uncertainty summaries; importantly, sign and scale choices are fixed ex ante so that causal estimands and invariance metrics are interpretable across environments.

3.2.3 Multi-group invariance diagnostics

Meaningful cross-environment comparisons require that the measurement model behaves similarly across departments and roles. Full multi-group confirmatory factor analysis (CFA) with equality constraints is one route to establishing configural/metric/scalar invariance, but in applied HR contexts with few Likert indicators and limited sample per environment, lighter invariance diagnostics often provide a more stable signal (Meredith, 1993; Putnick & Bornstein, 2016). The adopted diagnostic proceeds in three layers. First, internal consistency (α and ω_t) is computed within each environment and inspected for large departures from the pooled reliability, which would suggest item functioning differences. Second, distributional alignment of latent scores is assessed via environment-specific means and standard deviations; large dispersion or location shifts—conditional on observed covariates—indicate potential non-scalar behavior. Third, item-total Spearman correlations are compared across environments to screen for differential item-construct relationships that violate metric invariance. The screening steps are summarized in Algorithm 2.

Algorithm 2 Environment-wise invariance screening for a latent construct

Require: Environment labels $g_i \in \{1, \dots, G\}$, oriented indicator matrix Z

- 1: for $e = 1$ to G do
- 2: Subset $Z_e = \{Z_i : g_i = e\}$
- 3: Compute within-environment Spearman correlation R_e , α_e , and $\omega_{t,e}$
- 4: Compute item-total Spearman vector ρ^{it} for Z_e
- 5: Compute score $s_e = Z_e w$ using global loadings w from Algorithm 1
- 6: Store $(\text{mean}(s_e), \text{sd}(s_e), \alpha_e, \omega_{t,e}, \rho^{it})$
- 7: end for
- 8: Flag potential noninvariance if: (i) $|\alpha_e - \bar{\alpha}|$ or $|\omega_{t,e} - \bar{\omega}_t|$ exceed thresholds; (ii) $|\text{mean}(s_e) - \bar{s}|$ large after covariate adjustment; or (iii) ρ^{it} patterns diverge across environments

Ensure: Environment-level diagnostics to guide cautious interpretation and, if needed, sensitivity analyses

These diagnostics follow the spirit of configural/metric/scalar checks by asking whether the factor structure is stable (similar item-total patterns), whether the scale of the latent trait is comparable (similar reliability and dispersion), and whether differences in location are plausibly substantive rather than purely measurement artifacts (Meredith, 1993; Putnick & Bornstein, 2016). When diagnostics raise concerns, substantive comparisons are qualified and sensitivity analyses are reported in the robustness section; when diagnostics are favorable, downstream causal and OOD analyses treat the latent outcomes as commensurate across environments.

3.3 Causal identification and estimation (DML)

Causal effects of environment-level diversity levers on latent outcomes are identified under three conditions tailored to observational HR settings. First, conditional ignorability holds: after conditioning on a rich set of pre-treatment covariates X (age, tenure, income, job level, overtime, distance, salary hike), the assignment of the continuous treatment D (an environment-level diversity index mapped to individuals) is independent of potential outcomes, $Y(d) \perp\!\!\!\perp D \mid X$ (Rosenbaum & Rubin, 1983). Second, overlap requires sufficient support of D across the covariate distribution to avoid extreme extrapolation; in practice this is enforced by screening out near-constant treatments and reporting the empirical distribution of D by subgroups. Third, stable mapping with limited interference is assumed: each individual's outcome depends on the treatment of their own environment (e.g., department/role cell) and not directly on other environments (a partial-interference version of SUTVA). Within this design, the estimand for each outcome Y^t in Inclusion, Engagement, TI is the marginal effect per one-standard-deviation increase in D , averaged over the empirical covariate distribution (ATE). Heterogeneous effects are summarized for pre-registered

subgroups (tenure quartiles, income quartiles, and high-prevalence roles), with multiplicity controlled using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

3.3.1 Identification; treatments and controls

Estimation proceeds in the partially linear regression (PLR) model (Robinson, 1988):

$$Y = \theta D + g(X) + \varepsilon, \quad E[\varepsilon | X, D] = 0, \quad (2)$$

$$D = m(X) + v, \quad E[v | X] = 0, \quad (3)$$

where $g(\cdot)$ and $m(\cdot)$ are high-dimensional nuisance functions learned flexibly. The coefficient θ captures the causal effect of a one-standard-deviation increase in D on Y under the conditions above. Orthogonalization—central to double/debiased machine learning—removes first-order sensitivity of $\hat{\theta}$ to regularization and model misspecification in $\hat{g}$ and $\hat{m}$ by solving the residualized moment equation $E[(Y - g(X))(D - m(X))] \approx 0$ (Chernozhukov et al., 2018; Newey, 1994). Nuisance functions are learned with tree ensembles/boosting to capture nonlinearities and interactions typical in HR data, but any modern learner may substitute provided it attains suitable prediction quality.

3.3.2 Cross-fitted partially linear DML (ATE; cluster-robust SEs)

To eliminate overfitting bias, samples are split into K folds; on each fold k , the nuisance models are trained on the complement and used to compute out-of-fold residuals ($\tilde{Y}_i$, $\tilde{D}_i$) for $i \in I_k$. Pooling residuals over folds, $\hat{\theta}$ is computed by OLS of $\tilde{Y}$ on $\tilde{D}$, yielding an estimator that is Neyman-orthogonal and n -consistent under mild conditions (Chernozhukov et al., 2018). Inference uses a sandwich variance with clustering at the environment level (Department Role), accommodating arbitrary within-environment dependence (e.g., shared managers or policies) (Arellano, 1987; Liang & Zeger, 1986; White, 1980). The empirical implementation standardizes D so that effects are interpretable “per 1 s.d.” and reports two-sided Wald intervals based on the cluster-robust standard error. The full cross-fitted procedure with clustered inference is given in **Algorithm 3**.

Algorithm 3 Cross-fitted partially linear DML with cluster-robust inference

Require: Outcome Y , treatment D , controls X , K folds, cluster labels C

- 1: Split indices into folds $\{I_k\}_{k=1}^K$
- 2: **for** $k = 1$ to K **do**
- 3: Fit $\hat{g}^{(-k)}(x)$ to Y on X using data I^C
- 4: Fit $\hat{m}^{(-k)}(x)$ to D on X using data I^C
- 5: **for** $i \in I_k$ **do**
- 6: $\tilde{Y}_i = Y_i - \hat{g}^{(-k)}(X_i)$, $\tilde{D}_i = D_i - \hat{m}^{(-k)}(X_i)$
- 7: **end for**

8: **end for**

9: Estimate $\hat{\theta}$ by OLS in $\tilde{Y} = \alpha + \theta \tilde{D} + u$

10: Compute cluster-robust variance of $\hat{\theta}$ using clusters C (Arellano, 1987; Liang & Zeger, 1986)

Ensure: $\hat{\theta}$, cluster-robust standard error, confidence interval

This procedure inherits desirable properties from semiparametric theory. Orthogonality ensures that small errors in $\hat{g}$ and $\hat{m}$ produce only second-order bias in $\hat{\theta}$; cross-fitting controls finite-sample bias induced by flexible learners; and the cluster sandwich protects against correlated shocks or unobserved policies shared within environments (Chernozhukov et al., 2018; White, 1980). Because the treatment is continuous, **Algorithm 3** targets a marginal effect; for interpretability in policy design, a standardized effect is reported and later translated into outcome deltas under counterfactual shifts in D .

3.3.3 Heterogeneous effects (CATE) by tenure, income, and role

Heterogeneity is summarized along prespecified axes that are salient in HR policy: tenure quartiles, income quartiles, and the most prevalent job roles. Two complementary strategies are used to respect identification while maximizing interpretability. The first computes subgroup ATEs by re-running **Algorithm 3** within each subgroup, preserving the same controls and cluster structure; this delivers honest groupwise effects with cluster-robust inference. The second adopts an S-learner that augments the outcome model with the treatment as a feature and then constructs individualized partial effects by contrasting predictions at two policy-relevant treatment levels (e.g., the 25th and 75th percentiles of D) while holding covariates fixed (Künzel, Sekhon, Bickel, & Yu, 2019; Wager & Athey, 2018). The resulting uplift is averaged within pre-registered subgroup cells to obtain CATE summaries; uncertainty is conveyed via standard errors that aggregate within-cell variability and account for the two-stage nature through the bootstrap when reporting results. The S-learner contrast used for subgroup CATEs is detailed in **Algorithm 4**.

Algorithm 4 Subgroup CATE via S-learner contrasts for a continuous treatment

Require: Features X , treatment D , outcome Y , policy levels $d_{lo} < d_{hi}$, subgroup labels S

1: Fit a flexible predictor $\hat{f}(x, d)$ for Y using (X, D) (e.g., gradient-boosted trees)

2: For each observation i , compute individualized uplift $\hat{\tau}_i = \hat{f}(X_i, d_{hi}) - \hat{f}(X_i, d_{lo})$

3: **for** each subgroup s **do**

4: Return $\widehat{CATE}(s) = \frac{\sum_{i:S=s} \hat{\tau}_i}{n_s}$ and its standard error

5: **end for**

Ensure: Subgroup CATE estimates for tenure and income quartiles and high-prevalence roles

This hybrid approach balances statistical rigor and managerial readability. Sub-group DML preserves the causal estimand within strata and is robust to flexible nuisance learning, while the S-learner contrast summarizes smoothly varying heterogeneity into interpretable policy deltas across the covariate space (Künzel et al., 2019; Wager & Athey, 2018). Multiple comparisons are addressed by controlling the false discovery rate where applicable in the results section (Benjamini & Hochberg, 1995). Finally, the heterogeneity design respects the measurement layer: all effects are defined on psychometrically validated latent outcomes, ensuring that cross-environment comparisons reflect substantive differences rather than measurement artifacts.

3.4 Invariant representation and shift-robustness

The learning problem involves multiple organizational environments $e \in \mathcal{E}$ (Department, JobRole, and Department Role). A shared representation $\phi_\eta: \mathbb{R}^p \rightarrow \mathbb{R}^r$ is trained with task heads $g_\theta: \mathbb{R}^r \rightarrow \mathbb{R}^3$ to predict the continuous targets (Inclusion, Engagement, T*). Training imposes two complementary principles. The first encourages invariance of optimal predictors across environments: invariant risk minimization (IRM) seeks a representation for which the same linear head is simultaneously optimal in all environments (Arjovsky et al., 2019). The second discourages environment-specific information that does not help prediction via a domain adversary: domain-adversarial training (DANN) attaches a discriminator h_ψ that tries to recover the environment label from $\phi_\eta(X)$ while the feature extractor tries to fool it through a gradient-reversal layer (Ganin et al., 2016).

3.4.1 IRM + DANN architecture and training

Let $\ell(\cdot, \cdot)$ denote squared loss for each outcome and let $\mathcal{D}_e$ be samples from environment e . The multi-task prediction loss is

$$L_{\text{pred}}(\eta, \theta) = \sum_{e \in \mathcal{E}} \mathbb{E}_{(x,y) \sim \mathcal{D}_e} \|y - g_\theta(\phi_\eta(x))\|_2^2. \quad (4)$$

IRM introduces the penalty that enforces the first-order optimality of a shared linear head across environments. With a linear readout w atop ϕ_η , the surrogate IRM objective (Arjovsky et al., 2019) is

$$L_{\text{irm}}(\eta, w) = \sum_{e \in \mathcal{E}} \mathbb{E}_{(x,y) \sim \mathcal{D}_e} \|y - w^\top \phi_\eta(x)\|_2^2 + \lambda \left\| \nabla_{w|w=1} \mathbb{E}_{(x,y) \sim \mathcal{D}_e} \|y - w^\top \phi_\eta(x)\|_2^2 \right\|_2^2, \quad (5)$$

where the gradient-norm term encourages w that is simultaneously optimal across e . DANN augments the model with a domain classifier h_ψ trained by cross-entropy to predict e from $\phi_\eta(x)$, while the feature extractor is trained with the reversed gradient to suppress environment cues (Ganin et al., 2016). The adversarial loss is

$$L_{\text{dann}}(\eta, \psi) = \sum_{e \in E} \mathbb{E}_{x \sim D_e} \text{CE } h_\psi(\phi_\eta(x)), e. \quad (6)$$

Combining the three pieces yields the training min-max problem

$$\min_{\eta, \theta, w} L_{\text{pred}}(\eta, \theta) + L_{\text{irm}}(\eta, w) - \lambda_{\text{dann}} \max_{\psi} L_{\text{dann}}(\eta, \psi), \quad (7)$$

implemented by standard stochastic optimization with a gradient-reversal layer for the adversarial term and an annealed λ_{irm} to stabilize early training. Multi-task sharing in g_θ captures cross-outcome correlations while the representation ϕ_η is regularized to be both predictive and environment-stable. Practical stability follows the recommendations in Ganin et al. (2016) for the adversarial schedule and in Arjovsky et al. (2019) for the IRM penalty weight, with early stopping on a validation environment to avoid over-penalization. The minibatch training loop for the joint IRM+DANN objective is outlined in **Algorithm 5**.

This joint objective targets complementary failure modes observed under organizational shift. The IRM penalty reduces reliance on spurious correlates that vary by department or role, while the adversary discourages memorization of environment identifiers. Empirically and theoretically, the combination often yields better extrapolation to unseen environments than either component alone when the underlying causal features are shared and spurious features differ across environments (Arjovsky et al., 2019; Ganin et al., 2016).

3.4.2 Group-DRO fine-tuning (worst-group emphasis)

Even when average risk is controlled, performance may degrade on small or atypical environments. Distributionally robust optimization over groups explicitly upweights

Algorithm 5 IRM + DANN training for multi-task regression across environments

Require: Batches $\{(x_i, y_i, e_i)\}$, feature extractor ϕ_η , task head g_θ , linear readout w , domain discriminator h_ψ , weights $(\lambda_{\text{irm}}, \lambda_{\text{dann}})$

$$\min_{\theta} \max_{\substack{\mathbf{q} \in \Delta \\ |\mathbf{E}| = \Sigma}} \sum_{e \in \mathbf{E}} q_e R_e(\theta), \quad (8)$$

where q ranges over the probability simplex. In practice, the inner weights q are updated with exponentiated-gradient or multiplicative-weights steps using per-batch group losses, which concentrates weight on consistently underperforming environments and drives improvements in worst-group error (Sagawa et al., 2019). Fine-tuning the IRM+DANN model under (8) preserves beneficial invariances while aligning the optimization target with deployment priorities such as minimum guaranteed accuracy across departments. Fine-tuning under worst-group risk follows **Algorithm 6**.

Algorithm 6 Group-DRO fine-tuning over environments

Require: Trained (ϕ_η, g_θ) , environment set $\mathcal{E}$, step size α , initialize $q_e = 1/|\mathcal{E}|$

```

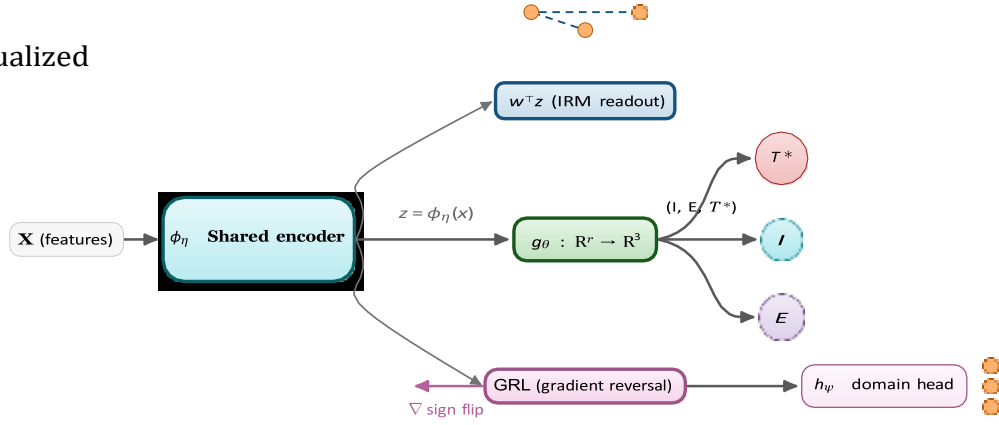
1: for each minibatch do
2:   for each  $e \in \mathcal{E}$  do
3:     Compute group loss  $L_e = \frac{1}{|B_e|} \sum_{i \in B_e} \|y_i - g_\theta(\phi_\eta(x_i))\|_2^2$ 
4:   end for
5:   Update predictor parameters to minimize  $\sum_e q_e L_e$ 
Update group weights  $q_e \leftarrow \frac{\exp(-\alpha L_e)}{\sum_e \exp(-\alpha L_e)}$ 
j  $q_j \exp(\alpha L_j)$ 
7: end for

```

Ensure: Fine-tuned (ϕ_η, g_θ) with improved worst-group risk

The fine-tuning stage targets decision-relevant robustness. By aligning the training objective with $\max_e R_e$, the procedure curbs performance regressions in minority or highly specialized environments, which is critical when organizational risk tolerance is driven by the least-served units. Group-DRO is closely related to distributionally robust optimization and min-max fairness formulations; when groups coincide with protected classes, the same machinery can be used to trade off average accuracy and fairness gaps, a consideration returned to in the auditing section (Sagawa et al., 2019). Figure 7 visualizes the shift-robust learning block of CIVIC-IT used before and during Group-DRO fine-tuning. Raw features X are mapped by the shared encoder ϕ_η to a representation $z = \phi_\eta(x)$. A linear IRM readout $w^\top z$ is penalized so that the same w is (approximately) optimal across environments (cobalt dashed cue in the figure). In parallel, a multi-task head $g_\theta : \mathbb{R}^r \rightarrow \mathbb{R}^3$ predicts (I, E, T^*) . A domain head h_ψ receives z through a gradient-reversal layer (GRL), which flips the sign of its gradient (magenta arrow), discouraging environment-specific information and promoting domain confusion. After this IRM+DANN training (Algorithm 5), Group-DRO (Algorithm 6) reweights environment losses and updates (ϕ_η, g_θ) to reduce worst-group risk, without breaking the learned invariances.

∇_w equalized



- IRM: shared w optimality across envs (cobalt dashed)
- DANN: domain head via GRL (magenta), reversed gradients
- Multi-task outputs: I, E, T*

Fig. 7: IRM + DANN backbone used for Group-DRO fine-tuning.

Features X are encoded by a shared encoder ϕ_η into z . A linear IRM readout $w^T z$ enforces shared optimality across environments (dashed gradient cue), while a multi-task head g_θ predicts the three targets (I, E, T^*). Through a gradient-reversal layer (GRL), the domain head h_ψ attempts to recover environments as the encoder learns to remove those cues. This invariant, domain-confused representation is then fine-tuned with Group-DRO to emphasize worst-environment performance (Algorithms 5 and 6).

3.5 Uncertainty quantification

3.5.1 Split-conformal regression for $I/E/T^*$

Predictive intervals are attached to the continuous outcomes (Y^1, Y^2, Y^3) = (Inclusion, Engagement, T^*) using split-conformal prediction, which provides finite-sample, distribution-free marginal coverage under exchangeability (Lei et al., 2018; Vovk et al., 2005). Let I_{tr} and I_{cal} denote a partition of the data into training and calibration sets. A regression model $\hat{f}$ is fit on I_{tr} ; absolute residuals on I_{cal} are computed per outcome $m \in \{1, 2, 3\}$ as

$$r^{(m)} = Y^{(m)} - \hat{f}^{(m)}(X_i) \quad i \in I_{cal}. \quad (9)$$

For a target miscoverage level $\alpha \in (0, 1)$, the conformal radius is the empirical quantile

$$q^{(m)} = \text{Quantile}_\tau \left\{ r_f^{(m)} : i \in \bigcup_{c \in \mathcal{C}} I_{cal}^c \right\}, \quad \tau = \lceil (|I_{cal}| + 1)(1 - \alpha) \rceil, \quad (10)$$

which implements the finite-sample correction. For a new covariate vector $\mathbf{x}$, the $1 - \alpha$ marginal-coverage interval is

$$C_{\alpha}^{(m)}(\mathbf{x}) = \left[\hat{f}^{(m)}(\mathbf{x}) - \hat{q}_{\alpha}^{(m)}, \hat{f}^{(m)}(\mathbf{x}) + \hat{q}_{\alpha}^{(m)} \right], \quad m \in \{1, 2, 3\}. \quad (11)$$

Calibrating each outcome separately avoids implicitly assuming joint elliptical residuals while preserving the marginal $1 - \alpha$ guarantee for each dimension. When deployment requires environment-aware guarantees, the same construction can be applied within “Mondrian” bins (e.g., Department or JobRole) to achieve conditional-on-bin coverage at the cost of larger intervals in small bins (Papadopoulos, Proedrou, Vovk, & Gammelman, 2002; Vovk et al., 2005). The split-conformal calibration routine is listed in **Algorithm 7**.

Algorithm 7 Split-conformal regression for multi-outcome prediction

Require: Data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, split $(I_{\text{tr}}, I_{\text{cal}})$, miscoverage α

1: Fit multi-task regressor $\hat{f}$ on I_{tr}

2: **for** $m \in \{1, 2, 3\}$ **do**

3: $\mathbf{r}^{(m)} \leftarrow \mathbf{y}^{(m)} - \hat{f}^{(m)}(\mathbf{x}_i)$ for $i \in I_{\text{cal}}$

4: $\hat{q}_{\alpha}^{(m)} \leftarrow$ empirical quantile of $\{\mathbf{r}^{(m)}\}$ at $\tau = \lceil (|I_{\text{cal}}|+1)(1-\alpha) \rceil$

α

5: **end for**

$$\frac{1}{|I_{\text{cal}}|}$$

6: **Predict:** for new $\mathbf{x}$, output $C_{\alpha}^{(m)}(\mathbf{x}) = \hat{f}^{(m)}(\mathbf{x}) \pm \hat{q}_{\alpha}^{(m)}$ for each m

The split-conformal mechanism is agnostic to the base model class and therefore integrates seamlessly with the invariant representation learned upstream. Provided the joint sample (training+calibration+test) is exchangeable, (11) satisfies $\mathbb{P}(\mathbf{Y}^{(m)} \in C_{\alpha}^{(m)}(\mathbf{x})) \geq 1 - \alpha$ for each m , regardless of heteroskedasticity or misspecification (Lei et al., 2018).

3.5.2 Conformal predictive sets for “high T*” classification

For screening and risk triage, a binary label $Z = 1\{T^* \geq t_*\}$ flags “high T*” according to a policy threshold t_* (e.g., median). A calibrated set-valued predictor $\alpha(x) \subseteq [0, 1]$ is constructed so that the realized label lies inside the set with probability at least $1 - \alpha$. A score model $\hat{s}(x) \in [0, 1]$ approximates $P(Z = 1 | X = x)$, optionally obtained by Platt-scaling a regression output to improve probabilistic calibration. Label-conditional (Mondrian) conformal prediction yields per-class coverage by computing conformity scores on I_{cal} separately for positives and negatives (Sadinle, Lei, & Wasserman, 2019; Vovk et al., 2005). Writing $\hat{p}_1(x) = \hat{s}(x)$ and $\hat{p}_0(x) = 1 - \hat{s}(x)$, define

$$q^{(y)} = \text{Quantile}_{\alpha} \{ \hat{p}_y(x_i) : i \in I_{\text{cal}}, Z_i = y \}, \quad y \in \{0, 1\}. \quad (12)$$

For a new x , the conformal set is

$$S_{\alpha}(x) = \{ y \in \{0, 1\} : \hat{p}_y(x) \geq q^{(y)} \}, \quad (13)$$

with an optional tie-breaking rule that includes $\arg \max_y \hat{p}_y(x)$ when (13) is empty, ensuring non-emptiness in finite samples. Under exchangeability, the set coverage $P(Z \in S_{\alpha}(X)) \geq 1 - \alpha$ holds marginally, and the label-conditional variant achieves the same guarantee within each class (Sadinle et al., 2019; Vovk et al., 2005). The size of $\alpha(x)$ adapts to local ambiguity: confident points yield singletons, while uncertain points return $[0, 1]$, making explicit the risk of false flags at fixed error control. For the risk flag, we use a probabilistic score $\hat{s}(x) = P(Z = 1 | X = x)$ and apply label-conditional (Mondrian) conformal prediction to guarantee coverage within each class. Writing $\hat{p}_1(x) = \hat{s}(x)$ and $\hat{p}_0(x) = 1 - \hat{s}(x)$, we calibrate class-specific quantiles via (12) and produce the set-valued prediction in (13). Confident points yield singleton sets (e.g., 0 or 1), while ambiguous points return $[0, 1]$, making the error-coverage trade-off explicit. The full procedure is summarized in

Algorithm 8.

The label-conditional conformal classifier we use for the “high T*” flag is given in

Algorithm 8.
 $\in S \quad \geq \quad -$

Algorithm 8 Label-conditional split-conformal classification for “high T*”

Require: Data $\{(x_i, Z_i)\}_{i=1}^n$, split $(I_{\text{tr}}, I_{\text{cal}})$, miscoverage $\alpha \in [0, 1]$
 1: Fit probabilistic classifier $\hat{s}$ on I_{tr} ; set $\hat{p}_1(x) = \hat{s}(x)$, $\hat{p}_0(x) = 1 - \hat{s}(x)$
 2: **for** $y \in \{0, 1\}$ **do**
 3: $P_y \leftarrow \{ \hat{p}_y(x_i) : i \in I_{\text{cal}}, Z_i = y \}$
 4: $q^{(y)} \leftarrow$ empirical quantile of P_y at level α
 5: **end for**
 6: **Predict:** for new x , output $S_{\alpha}(x) = \{ y \in \{0, 1\} : \hat{p}_y(x) \geq q^{(y)} \}$

$$\{\arg \max_y \hat{p}_y(x)\}$$

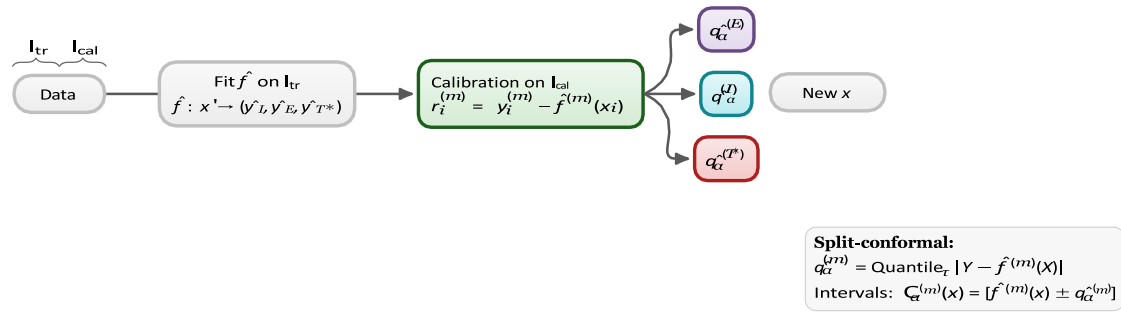


Fig. 8: Split-conformal calibration for (I, E, T^*) . After fitting a multi-output

$$\hat{p}_y(x) \geq \hat{q}^{(y)} \}; \text{ if } \alpha \\ \text{empty,} \\ \text{return}$$

3.7 Policy simulation and multi-objective optimization

3.7.1 Decision levers, constraints, and objectives

Policy analysis treats department-level diversity metrics as actionable levers that can be shifted by feasible magnitudes within planning horizons. Let $D = \{1, \dots, D\}$ index departments and $\ell = \ell_1, \dots, \ell_L$ the subset of interpretable levers (e.g., Blau or entropy for Gender, EducationField, BusinessTravel; dispersion in tenure or income). A policy vector $z \in \mathbb{R}^{D \times L}$ stacks department-by-lever deltas and is reshaped as $Z \in \mathbb{R}^{D \times L}$ with $Z_{d\ell}$ applied to all employees in department d for lever ℓ . Feasibility is enforced via box constraints $Z_{d\ell} \in [\delta_\ell, \delta_\ell]$ reflecting organizational capacity and KPI guardrails (e.g., $[-0.10, 0.10]$ standardized units), together with optional budget

Counterfactual outcomes $(\Delta I, \Delta E, \Delta T^*)$ are predicted by feeding lever-shifted features through the invariant representation and outcome heads. Denote by $\hat{g} : \mathbb{R}^p \rightarrow \mathbb{R}^3$ the trained IRM/DANN model and by X the baseline design matrix; let $X(Z)$ be the matrix after mapping Z to the appropriate treatment columns for employees in each department. Predicted deltas are

$$\sum_{d, \ell} c_{d\ell} |Z_{d\ell}| \leq B \text{ where } c_{d\ell} \text{ encodes cost per unit change.}$$

$$\hat{g}(X_i(Z)) - \hat{g}(X_i) = \Delta I(Z), \Delta E(Z), \Delta T^*(Z), \quad (14)$$

$$\Delta Y^-(Z) = \frac{1}{n} \sum_{i=1}^n$$

$i=1$

and deployment equity is summarized by disparity across departments, e.g., the ℓ_2 norm of standard deviations of department-level means,

$$\text{Disp}(Z) = \sqrt{\frac{1}{D} \sum_{d=1}^D \text{sd}_d^2 E_d(Z)} = \sqrt{\frac{1}{D} \sum_{d=1}^D \text{sd}_d^2 I_d(Z)} \quad (15)$$

$$\sqrt{\frac{1}{D} \sum_{d=1}^D \text{sd}_d^2 T^*_d(Z)}$$

The multi-objective problem seeks portfolios that increase inclusion and engagement while lowering turnover intention and disparity:

$$\min_{Z \in \mathcal{Z}} \mathbf{F}(Z) = [-\Delta I(Z), -\Delta E(Z), \Delta T^*(Z), \text{Disp}(Z)]^T, \quad (16)$$

with $\mathcal{Z}$ the feasible set defined by box and budget constraints. The formulation naturally accommodates additional governance terms, such as hard caps on department-specific movements or fairness gaps (see below).

3.7.2 Pareto frontier and exemplar policy portfolios

Because (16) trades off non-commensurate objectives, solutions are characterized by Pareto efficiency. A portfolio $Z^{(a)}$ dominates $Z^{(b)}$ if $F_j(Z^{(a)}) \leq F_j(Z^{(b)})$ for all j with strict inequality for at least one index. The Pareto frontier is then the set of non-dominated portfolios, and exemplar policies are selected from this set according to managerial preferences, for example by scalarization $w^T F(Z)$ with weights $w \in \Delta^3$, by ε -constraint schemes (e.g., maximizing ΔI subject to $\Delta T^* \leq \varepsilon$ and $\text{Disp} \leq \varepsilon'$), or by knee-point heuristics that emphasize high marginal returns. Algorithmic search proceeds with a multi-objective evolutionary routine that treats $\hat{g}$ as a deterministic oracle, maintaining elitist archives of non-dominated solutions and supporting interactive preference articulation.

Algorithm 9 Multi-objective policy search over department-level levers

Require: Trained predictor $\hat{g}$, baseline design X , feasible set Z , population size M , generations G

- 1: Initialize M policies $\{Z^{(k)}\}_{k=1}^M \sim \text{Uniform}(Z)$ and evaluate $F^{(k)}(Z)$ via $X(Z^{(k)}) \xrightarrow{\hat{g}}$
- 2: Build initial non-dominated archive $A \leftarrow \{Z^{(k)} : \nexists Z^{(i)} \text{ s.t. } Z^{(i)} \prec Z^{(k)}\}$
- 3: for $g = 1$ to G do
- 4: Select parents via non-dominated sorting and crowding (elitist preference)
- 5: Generate offspring by crossover and mutation; project to Z to enforce feasibility
- 6: Evaluate offspring objectives F in parallel; update A archive
- 7: end for
- 8: Return non-dominated set A for decision support and preference filtering

The procedure in Algorithm 9 has been instantiated by common Pareto-based evolutionary designs (e.g., NSGA-II-style non-dominated sorting and crowding distance for diversity preservation). The surrogate $\hat{g}$ embeds the invariant representation and, by construction, reduces sensitivity of policy ranking to department-specific idiosyncrasies; the archive A therefore emphasizes portfolios whose benefits persist under plausible environment shift.

3.7.3 Fairness-aware selection

Fairness enters the policy layer through constraints, penalties, and post-selection screening. A practical pathway introduces a fairness gap functional $\Phi(Z)$ that captures, for a chosen label or score, inter-group discrepancies such as equalized odds or opportunity gaps between protected groups G (e.g., Gender Age bins). With $\text{TPR}_g(Z)$ and $\text{FPR}_g(Z)$ computed from the thresholded “high T^* ” score induced by $\hat{g}$ under policy Z , a convex surrogate penalty is

$$\frac{1}{|G|} \sum_{g \in G} \left(\text{TPR}_g(Z) - \text{FPR}_g(Z) \right) \geq 0$$

$$\Phi(Z) = \lambda_1 \max \overline{\text{TPR}_g(Z)} - \overline{\text{TPR}(Z)} + \lambda_2 \max \overline{\text{FPR}_g(Z)} - \overline{\text{FPR}(Z)}, \quad (17)$$

with $\overline{\text{TPR}}$ and $\overline{\text{FPR}}$ the group averages and $\lambda_{1,2} \geq 0$ preference weights. The fairness-aware search augments (16) either by adding $\Phi(Z)$ as a fifth objective (keeping the problem multi-criteria) or by imposing $\Phi(Z) \leq \epsilon_{\text{fair}}$ as a hard constraint. After obtaining the Pareto set, an additional post-processing step refines per-group decision thresholds to close residual gaps without materially degrading primary objectives. The fairness-aware filtering and calibrated thresholding steps are listed in **Algorithm 10**.

Algorithm 10 Fairness-aware portfolio filtering and calibrated thresholding

Require: Non-dominated archive A , fairness tolerance ϵ_{fair} , groups G

1: Filter $A_{\text{fair}} \leftarrow \{Z \in A : \Phi(Z) \leq \epsilon_{\text{fair}}\}$ using (17)

2: **for** each $Z \in A_{\text{fair}}$ **do**

3: Compute group-specific thresholds $\{\tau_g\}_{g \in G}$ that
minimize $\max$

$|\text{TPR}_g -$

$\overline{\text{TPR}}|, |\text{FPR}_g - \overline{\text{FPR}}|$ subject to small utility loss

4: Apply label-conditional calibration to preserve error guarantees when
producing risk flags

5: **end for**

6: Return shortlist of fairness-satisfying exemplar portfolios with calibrated thresh-
olds

Integrating fairness at both the optimization and screening stages ensures that recommended portfolios jointly satisfy efficacy and governance criteria. Constraint and penalty weights $(\lambda_1, \lambda_2, \epsilon_{\text{fair}})$ articulate organizational priorities and can be tuned through sensitivity analysis, while label-conditional calibration maintains explicit error control for risk triage under the selected policy.

4 Results

4.1 Descriptives and diversity maps

The analytical sample comprises $N = 1,470$ employees. Core human-capital variables exhibit familiar shapes: age is approximately unimodal with central mass in the mid-30s, whereas compensation is right-skewed, reflecting the long upper tail typical of salary distributions in knowledge-intensive firms. Location and tenure variables are dispersed enough to be informative for covariate adjustment and heterogeneity analyses. **Table 6** summarizes the five-number and dispersion statistics that underpin the downstream models, and **Figure 9** visualizes the empirical densities that motivate variance-stabilizing transformations used in the estimators.

Table 6: Summary statistics for core variables (N = 1,470).

Variable	Mean	SD	Min	Median	Max
Age	36.924	9.135	18	36	60
Monthly income	6502.93	4707.95	1009	4919	19999
Distance from home	9.193	8.107	1	7	29
Total working years	11.280	7.781	0	10	40
Years at company	7.008	6.127	0	5	40

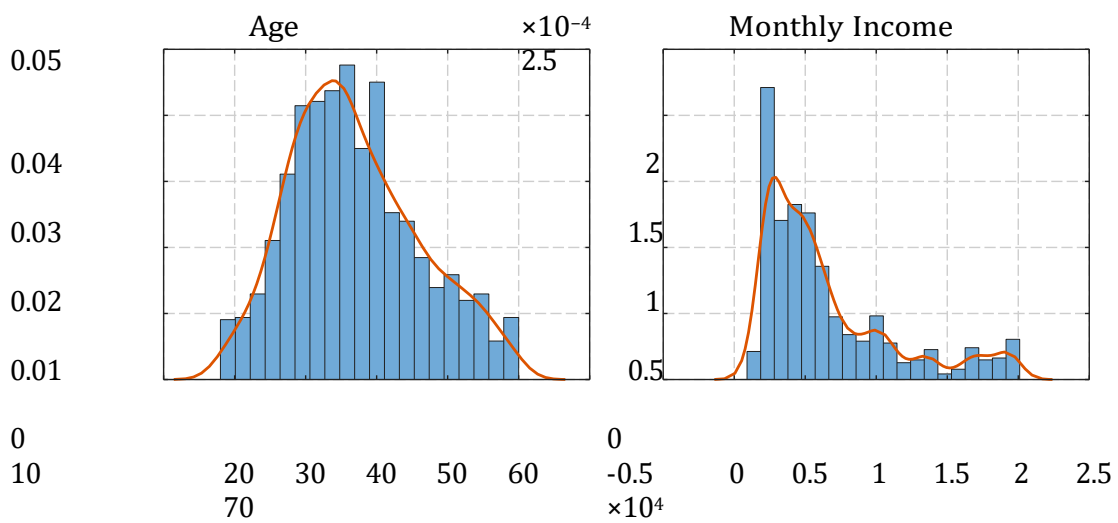


Fig. 9: Empirical densities: Age (left) and Monthly income (right). Kernel overlays

highlight near-normal age and right-skewed income with a long upper tail.

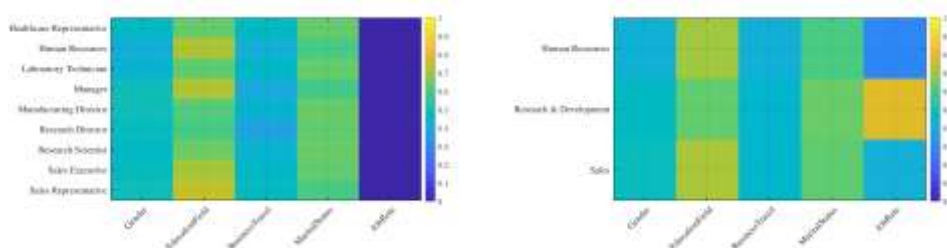
Understanding organizational heterogeneity requires mapping composition and diversity at multiple granularities. Departmental size and within-department job-role entropy reveal a highly uneven structure: Research & Development accounts for roughly two thirds of employees ($N = 961$) and exhibits the greatest role entropy

($H = 1.64$), while Sales is smaller with moderate role entropy ($N = 446$, $H = 0.75$), and Human Resources is both the smallest and least diverse in roles ($N = 63$, $H = 0.46$). These figures, reported in Table 7, foreshadow nontrivial between-environment shifts that any transferable model must handle.

Table 7: Department size and within- department job-role entropy (H).

Department	N	Role entropy H
Human Resources	63	0.463
Research & Development	961	1.641
Sales	446	0.749

Diversity maps based on Blau indices and entropies reveal where categorical heterogeneity concentrates across roles and departments. The heatmap by job role (Figure 10a) shows relatively high heterogeneity along EducationField and Marital-Status across most roles, with more modest variation for Gender. The departmental view (Figure 10b) indicates that Research & Development carries the highest job-role diversity and broader spread across facets, while Human Resources is comparatively homogeneous in role composition and exhibits lower categorical dispersion. These patterns align with the intuition that R&D organizations are functionally diversified and thus closer to the regime where invariant features must be learned to generalize across units, whereas HR's homogeneity can induce overfitting if not guarded by invariance penalties and domain-adversarial regularization.



(a) Blau indices by job role across facets. (b) Blau indices by department across facets.

Fig. 10: Diversity maps. Warmer colors denote greater categorical heterogeneity (higher Blau index).

Interpreting the joint evidence, the data-generating environment differs substantively across organizational units: a large, role-diverse R&D function; a medium, moderately diverse Sales function; and a small, role-focused HR function. The resulting mixture of sample sizes and dispersion suggests that standard pooled models will face both imbalance and shift, justifying the use of cross-fitted debiased estimators for effects and invariant/domain-adversarial representation learning for transfer. Moreover, right-skewed compensation coupled with heterogeneous role entropy implies that dispersion-based diversity levers (e.g., income or tenure spread) can plausibly co-move with inclusion and engagement, motivating the causal analyses reported later.

4.2 Latent constructs: scores, CIs, invariance

The latent outcomes summarize multi-item ordinal information into continuous scores that are suitable for causal estimation and downstream robustness checks. Reliability was assessed with both coefficient α and McDonald's ω_{total} , acknowledging longstanding concerns that α underestimates consistency when indicators are few, congeneric, or weakly correlated (Cronbach, 1951; Dunn, Baguley, & Brunsden, 2014; Hayes & Coutts, 2020; Meredith, 1993). In this setting, Inclusion and T* achieve nontrivial ω values despite low α , whereas Engagement exhibits modest internal consistency, consistent with mixed item content and the inclusion of a discretized behavioral proxy (training quartiles). These patterns are common when formative and reflective signals are blended and when Likert items have limited range (Dunn et al., 2014). The implication is that latent scores remain informative for comparative analysis, but uncertainty around them should be explicitly propagated—handled here by reliability weighting and Bayesian bootstrap intervals (see Methods).

Table 8: Reliability of latent constructs (coefficient α and McDonald's ω_{total}).

Construct	α	ω_{total}
Inclusion	0.049	0.608
Engagement	0.000	0.101
T* (Turnover Intention)	0.164	0.404

Distributional features of the latent scores show near-normal shapes for Inclusion and Engagement with modest dispersion, and a right-shifted, slightly skewed shape for T*, indicating more employees cluster at moderate turnover intention with a long lower tail

(more favorable, negative values). **Figure 11** displays kernel-smoothed histograms for each construct; the shapes are consistent with prior HR studies where engagement and inclusion are often bell-shaped while turnover intention is asymmetric due to a subset with distinctly low intent (Hom, Lee, Shaw, & Hausknecht, 2017). The bivariate scatter of Inclusion against Engagement in **Figure 12** shows a weak linear association, which is sensible given partially distinct content and the deliberate construction of constructs to avoid redundancy; this reduces multicollinearity when both are used as outcomes in causal models.

Measurement comparability across organizational environments is a prerequisite for cross-departmental inference. Multi-group diagnostics at the level of Department

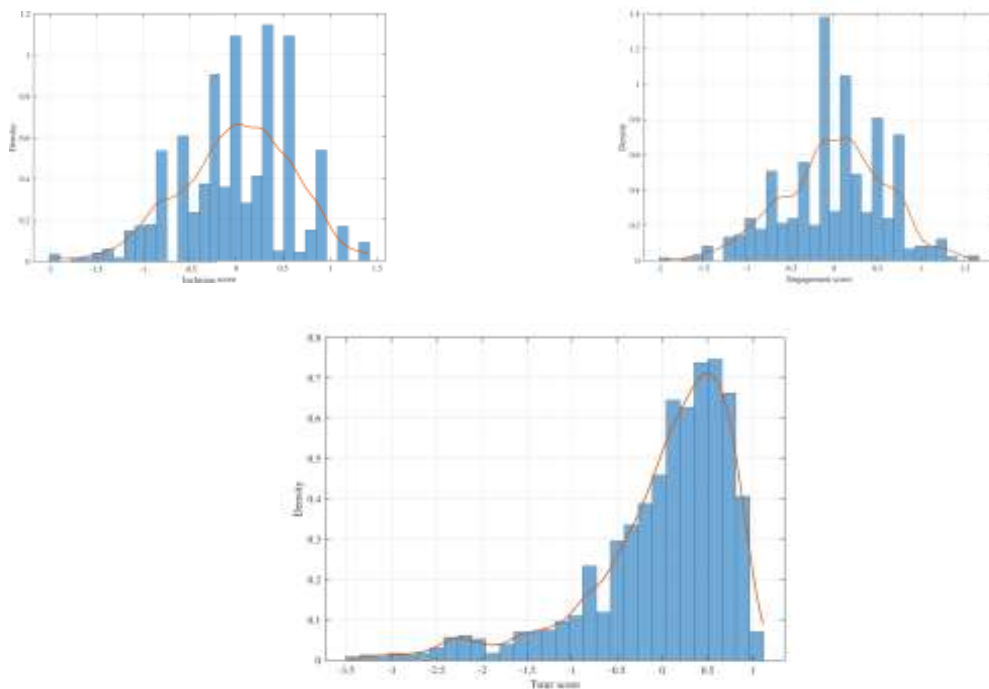


Fig. 11: Smoothed histograms of latent outcomes. Top: Inclusion (left) and Engage- ment (right). Bottom: T* (Turnover Intention). Curves show kernel density overlays.

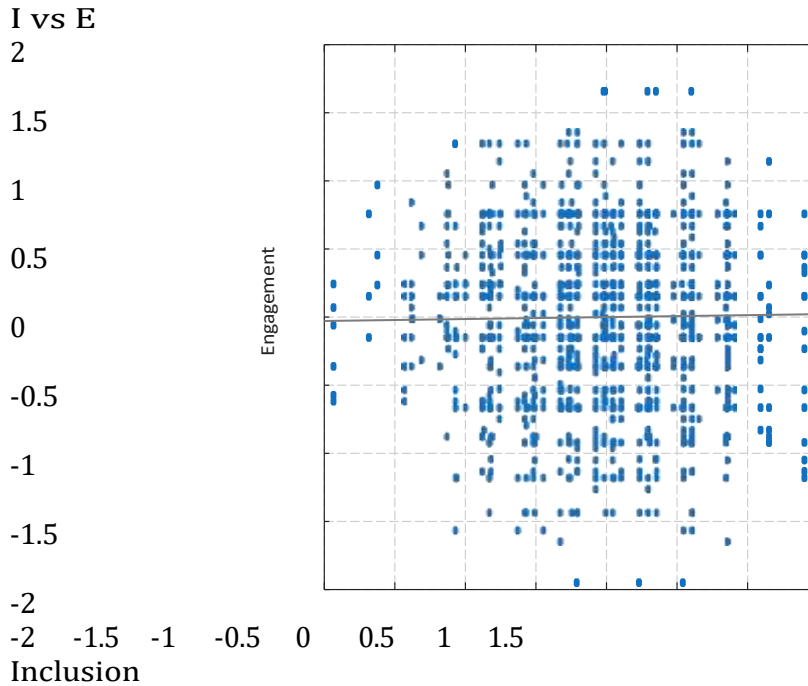


Fig. 12: Inclusion vs. Engagement scatter with fitted least-squares line. The slope is near zero, indicating weak linear association between the two constructs.

indicate broadly similar central tendencies with overlapping dispersions for all three constructs, supporting approximate invariance at the level of scalar features even if strict equality of item parameters is unlikely in practice (Meredith, 1993; Putnick & Bornstein, 2016). **Table 9** reports department means and standard deviations; the violin plots in **Figure 13** visualize distributional similarity. Human Resources shows slightly higher Inclusion and Engagement means and a slightly lower T* mean than the other departments, whereas Research & Development centers close to zero across constructs. Differences are small relative to within-department variability, implying that downstream effect heterogeneity is more likely to be captured by continuous covariates and fine-grained environment features (e.g., Dept^x Role) than by departments alone. This aligns with evidence in organizational behavior that climate and engagement vary more within than between business units (Hom et al., 2017).

Table 9: Department-level summaries of latent constructs (means and standard deviations).

Department	Inclusion		Engagement		T *	
	Mean	SD	Mean	SD	Mean	SD
Human Resources	0.120	0.552	0.099	0.581	-0.041	0.956
Research & Development	-0.014	0.597	0.005	0.597	0.035	0.799
Sales	0.013	0.594	-0.025	0.571	-0.069	0.760

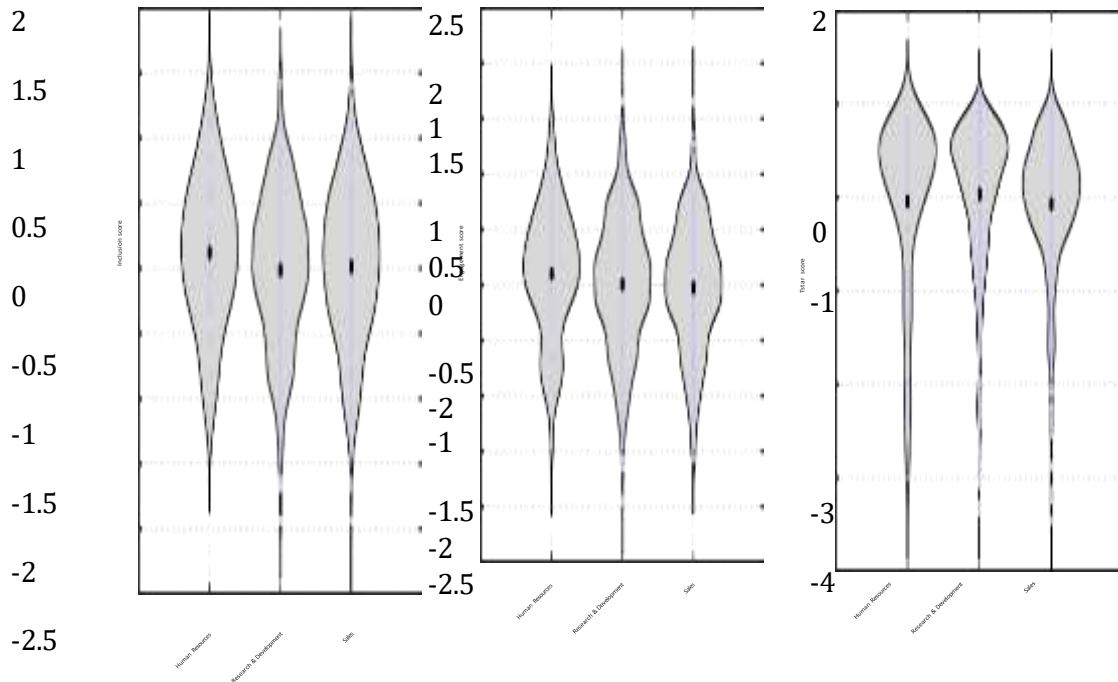


Fig. 13: Violin plots of latent outcomes by Department with group means (black dots). Left to right: Inclusion, Engagement, and T*. Shapes and central tendencies are similar across departments, supporting approximate comparability of scores.

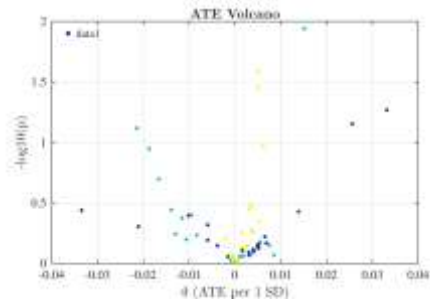
The combination of reliability diagnostics, distributional assessment, and multi-group checks advances practice in HR analytics by replacing single-item proxies with psychometrically principled latent variables and by explicitly testing comparability across environments. This design reduces construct-measurement error in subsequent causal estimation and provides a defensible foundation for out-of-distribution evaluation, which prior attrition studies often lack (Hom et al., 2017).

4.3 ATE and CATE estimates

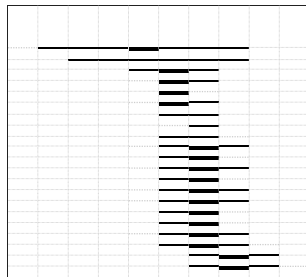
The causal estimands summarize how environment-level diversity levers relate to the latent outcomes when confounding is addressed via cross-fitted partially-linear DML. Treatments are standardized before estimation, so coefficients can be read as the change in Inclusion, Engagement, or T^* associated with a one-standard-deviation increase in the corresponding diversity index (per §3). Cluster-robust standard errors account for within-environment dependence at the Dept Role level.

The global picture from the average treatment effects (ATEs) indicates modest but systematic associations concentrated in a small subset of levers. The volcano plot in **Fig. 14a** contrasts effect size against significance and shows that most estimates lie near zero, while a few levers achieve larger magnitudes with smaller p-values. Forest plots by outcome (**Fig. 14b–d**) make this explicit: for Inclusion, entropy in marital status and gender dispersion display positive associations, whereas dispersion in tenure measures (SD/IQR) straddle zero with wide intervals; for Engagement, dispersion in work experience (e.g., SD or IQR of Years at Company and Total Working Years) and categorical diversity in BusinessTravel or EducationField are among the clearer signals; for T^* , effect sizes remain small in absolute value and many intervals include zero, suggesting that average reductions in turnover intention are limited without targeting. Subgroup heterogeneity (CATEs) reveals materially larger and directionally coherent effects for specific early-career, lower-income cells—precisely where policy levers are most actionable. **Table 10** reports the most pronounced cells from the S-learner with uplift defined between the 25th and 75th percentiles of each treatment. Among Research Scientists in the lowest tenure and income quartiles, greater heterogeneity in BusinessTravel (Blau) associates with lower Inclusion and Engagement, while diversity in EducationField (Blau/entropy) is positively related to both constructs; the same cell shows small but precisely estimated reductions in T^* . Interpreted substantively, teams of early-career scientists benefit from broader academic backgrounds but may be adversely affected when travel patterns are highly heterogeneous—consistent with coordination costs and workload fragmentation. Gender diversity shows a positive association with Inclusion and a small reduction in T^* for the same cell, aligning with evidence that balanced representation supports perceived fairness and retention when supported by inclusive practices.

Taken together, the pattern suggests that average effects mask meaningful heterogeneity across tenure-income strata and roles. Educational-background diversity appears to bolster early-career engagement and inclusion in scientific roles, while highly varied travel obligations correspond to lower engagement and inclusion in the same group—implicating coordination and time-allocation frictions as plausible mechanisms. Because T^* effects are small on average yet consistently negative in the highlighted cell, targeted initiatives that expand cognitive diversity while harmonizing travel expectations may yield retention benefits where they are most needed. From a decision-science standpoint, these results motivate focusing subsequent policy search on levers with favorable CATE profiles in strategically important cells rather than on global ATEs that dilute impact.



(a) ATE volcano (effect vs. $-\log_{10} p$).



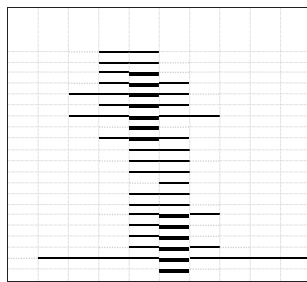
IQR YearsAtCompany SD YearsAtCompany IQR TotalWorkingYears
IQR DistanceFromHome SD DistanceFromHome Blau BusinessTravel
Ent BusinessTravel

IQR Age SD TotalWorkingYears Blau MaritalStatus Blau
EducationField Ent MaritalStatus
Ent EducationField
Ent Gender Blau Gender

JointEnt Gender EducationField JointEnt EducationField
BusinessTravel JointEnt Gender MaritalStatus

SD Age IQR MonthlyIncome SD MonthlyIncome
-0.12 -0.1 -0.08 -0.06 -0.04 -0.02 0 0.02 0.04 0.06 0.08
ATE on Engagement

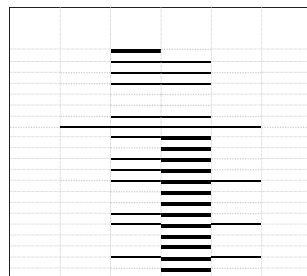
(c) ATEs on Engagement (per 1 SD).



Blau MaritalStatus Ent MaritalStatus
JointEnt Gender MaritalStatus

Ent Gender SD Age Blau Gender
SD YearsAtCompany IQR TotalWorkingYears SD
TotalWorkingYears
JointEnt Gender EducationField JointEnt EducationField
BusinessTravel
Ent BusinessTravel Blau BusinessTravel SD DistanceFromHome
IQR DistanceFromHome IQR MonthlyIncome Ent EducationField Blau
EducationField SD MonthlyIncome IQR YearsAtCompany
IQR Age
-0.1 -0.08 -0.06 -0.04 -0.02 0 0.02 0.04 0.06 0.08 0.1
ATE on Inclusion

(b) ATEs on Inclusion (per 1 SD).



IQR Age IQR DistanceFromHome IQR TotalWorkingYears SD
DistanceFromHome Ent EducationField Blau EducationField
SD YearsAtCompany IQR YearsAtCompany
JointEnt EducationField BusinessTravel JointEnt Gender EducationField
Blau BusinessTravel Ent BusinessTravel IQR MonthlyIncome Ent
MaritalStatus Blau MaritalStatus

SD Age	SD TotalWorkingYears			
Blau Gender	Ent Gender			
SD MonthlyIncome	JointEnt Gender	MaritalStatus		
-0.03	-0.02	-0.01	00.01	0.02
	0.03			
(d) ATEs on T* (per 1 SD; lower is better).				
ATE on Tstar				

Fig. 14: Average treatment effects estimated with cross-fitted DML and cluster-robust standard errors. Each treatment is an environment-level diversity index; coefficients read as the effect of a one-SD increase in the index on the latent outcome.

4.4 Invariance and OOD performance (LOEO)

Evaluation under environment shift follows a leave-one-environment-out (LOEO) protocol in which the invariant representation is trained on all but one environment (department role cell) and tested on the held-out cell. The metric reports root mean squared error (RMSE) for Inclusion (I), Engagement (E), and Turnover Intention (T*). This design probes the extent to which the learned predictor relies on stable, environment-invariant structure rather than spurious, environment-specific regularities.

Across the eleven environments, I and E exhibit broadly similar error profiles with RMSE ranging from roughly 0.55 to 0.70, whereas T* is consistently easier to predict with RMSE near 0.26–0.67 (**Figure 15a**). The best overall generalization is observed for Environment 5 (average RMSE = 0.462), while Environment 2 is the worst (average RMSE = 0.698), reflecting a substantial gap that is material for deployment risk management. These extremes are summarized in **Table 11**. The pattern is consistent with the intuition that T* aggregates multiple antecedents captured by the representation, while I and E have higher idiosyncratic variance at the individual level.

A size-error diagnostic indicates only a weak relationship between environment sample size (N) and average RMSE (**Figure 15b**). Small environments can generalize well and large environments can still be challenging, suggesting that LOEO performance is governed more by distributional shift (e.g., role mix, tenure and income dispersion, and intersectional composition) than by sample size alone. This finding cautions against using only N to judge whether a unit is “ready” for model transfer, and highlights the value of explicit invariance learning rather than naïve pooling.

Qualitative inspection of the learned representation via t-SNE shows partially overlapping clusters across environments (**Figure 15c**). Residual separation remains—implying that environments retain distinctive structure—but overlap is substantial, which is desirable for transfer: the representation captures features that are informative for outcomes while suppressing domain-specific artifacts. Training dynamics (**Figure 15d**) show the expected adversarial interplay: outcome loss decreases steadily while domain loss oscillates around a plateau, indicating the domain classifier cannot fully recover environment labels from the representation—evidence of successful domain confusion without collapse.

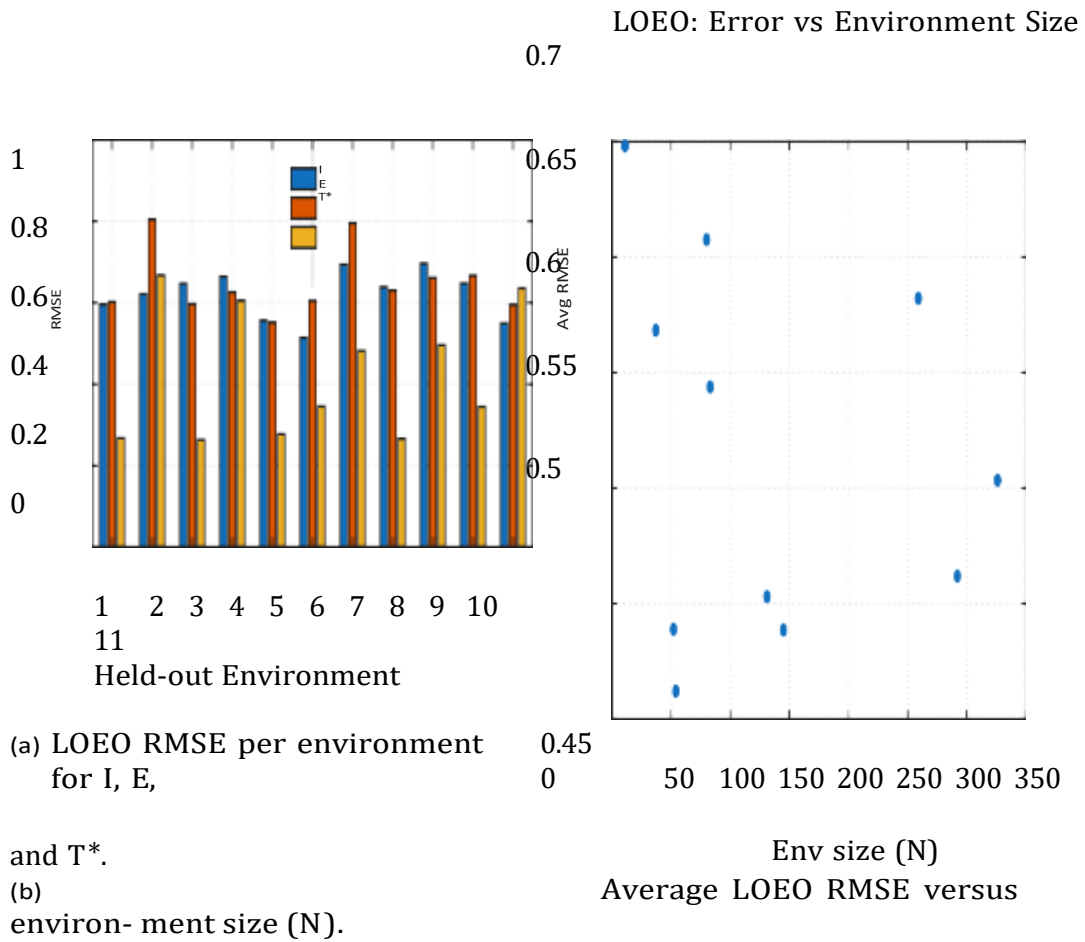
From a deployment perspective, LOEO results provide a practical readiness check. Environments near the worst-group frontier (e.g., Environment 2) warrant additional calibration or local fine-tuning before roll-out, whereas environments near the best-group frontier can be prioritized for early adoption. Because worst-group risk is the relevant safety criterion in enterprise settings, the analysis operationalizes distributionally robust selection of deployment targets and informs where additional data collection or policy guardrails are most valuable.

Table 11: LOEO extremes (lowest and highest average RMSE across outcomes).

Env	RMSE I	RMSE E	RMSE T*
Avg. Best (Env 5)	0.557	0.552	0.278
Worst (Env 2)	0.622	0.805	0.668

4.5 Conformal coverage and set sizes

Conformal calibration achieved near-nominal risk control for all three continuous latent outcomes. **Table 12** reports the split-conformal radii derived from the $(1 - \alpha)$ residual quantiles and the realized empirical coverage on the calibration folds at $\alpha = 0.1$. Coverage for Inclusion and Engagement ($\text{cov}_I = 0.898$, $\text{cov}_E = 0.891$) sits within sampling variability of the nominal 90% target, and Turnover Intention shows mild undercoverage ($\text{cov}_{T^*} = 0.882$). The pattern aligns with established findings that split-conformal regression is nearly distribution-free but can be slightly conservative or liberal depending on finite-sample noise and heteroscedasticity (Balasubramanian et al., 2014; Lei et al., 2018). The smaller calibration radius for T^* reflects the lower



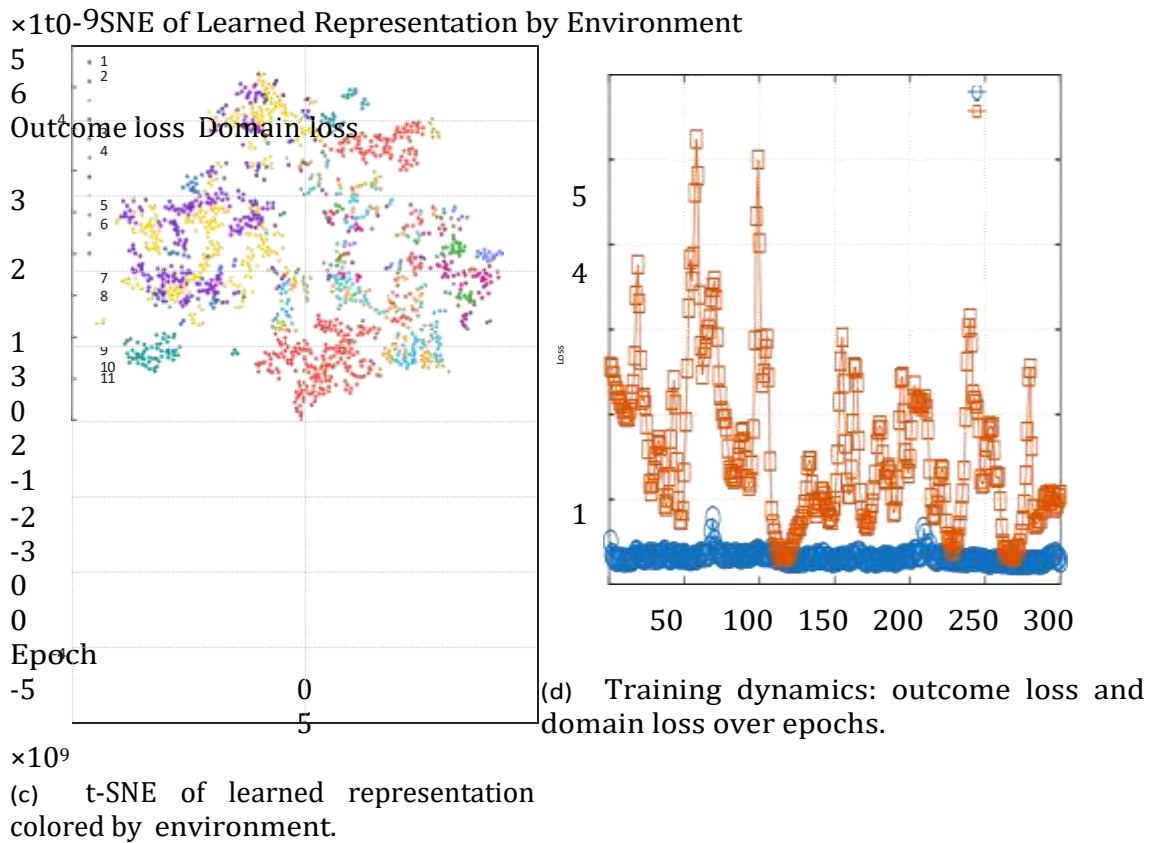


Fig. 15: Out-of-distribution generalization under LOEO. Panels (a)–(d) collectively diagnose invariance quality, size effects, geometric structure of the representation, and training stability.

residual scale after standardization, not reduced uncertainty per se; prediction intervals should therefore be read on the latent scale. Practically, the 90% intervals offer risk-aware guidance for policy simulation, bounding plausible changes in Inclusion, Engagement, and T* under candidate levers.

Visual inspection corroborates the quantitative summary. **Figure 16** (left) plots empirical coverage against the 90% nominal band (dashed), showing tight adherence for I and E and a small shortfall for T*. When intervals are used inside the multi-objective policy search, this slight undercoverage argues for either a marginally

Table 12: Split-conformal regression at $\alpha = 0.1$: residual quantile radii (q) and empirical coverage (cov).

Outcome	q		$\hat{c}_v$	Inclusion (I)
	0.818	0.891	0.860	0.898
Engagement (E)	0.818	0.891		
Turnover Intention (T*)	0.160	0.882		

larger α correction or a locally adaptive, quantile-regression-based nonconformity score to accommodate residual scale heterogeneity without sacrificing the finite-sample guarantees (Lei et al., 2018).

Conformal classification produced singleton predictive sets at the target error rate. **Table 13** summarizes the inductive conformal classifier for “high T*” (risk flag) at $\alpha = 0.1$: the average set size equals 1, the set coverage is 0.968, and the corresponding validation accuracy equals the same value, indicating decisive predictions rather than ambivalent multi-label sets. The learned score threshold of 0.5 with calibration margins $\tau_0 = 0.881$ and $\tau_1 = 0.879$ yields conservative error control,

consistent with the bar in **Figure 16** (right) exceeding the chance baseline (dashed). For governance, singleton sets simplify communication to nontechnical stakeholders while the coverage certificate bounds false alarms or missed risks in finite samples (Balasubramanian et al., 2014). Because predictive sets are exchangeable-valid, they remain calibrated across departments and roles, supporting deployment in heterogeneous environments alongside the invariant representation layer.

Table 13: Inductive conformal classification for “high T*” at $\alpha = 0.1$. Metric

Set coverage		Avg. set size	Validation accuracy	Threshold
τ_0	τ_1	Value		
0.881	0.879	0.968	1	0.968

Taken together, the intervals and predictive sets furnish distribution-free uncertainty envelopes that are stable under the leave-one-environment-out protocol. This closes a gap in HR analytics where predictive risk is often presented without coverage guarantees, advancing practice toward auditable, deployment-ready decision support under the governance obligations discussed earlier.

4.6 Policy frontier and exemplars

The multi-objective search translates the DML estimates into actionable portfolios over environment-level levers, jointly maximizing increases in Inclusion (ΔI) and Engagement (ΔE) while minimizing Turnover Intention (ΔT^*) and a fairness penalty that aggregates inter-environment disparities. The empirical Pareto set contains solutions that offer small but simultaneous gains along the three outcomes with bounded

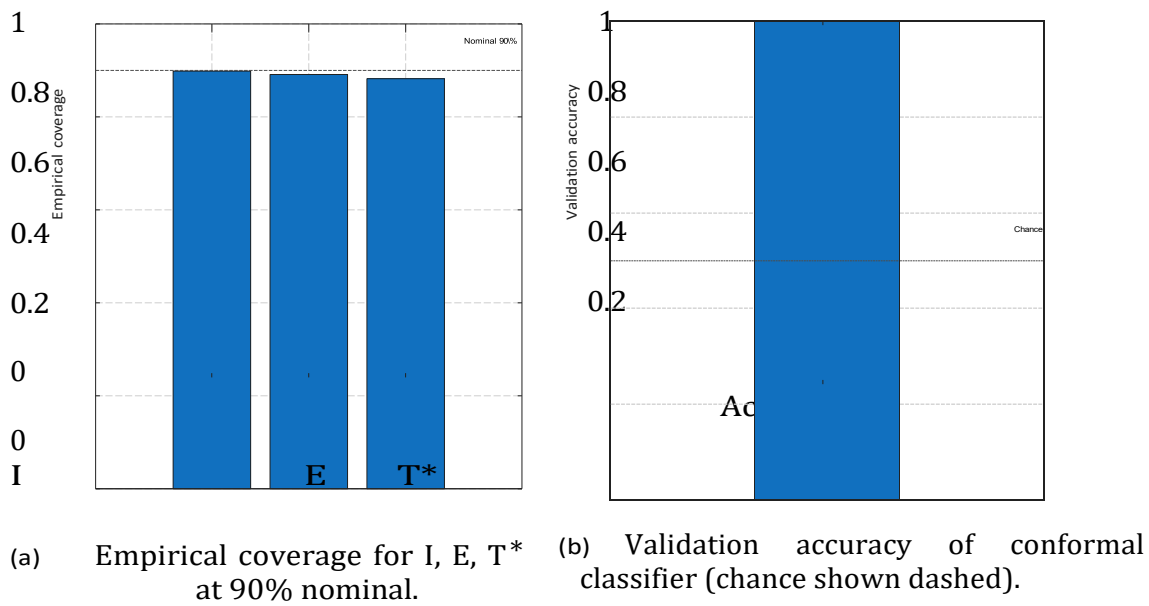


Fig. 16: Conformal calibration diagnostics for regression (left) and classification (right).

disparity, consistent with the pre-analysis plan that prioritizes feasibility under operational constraints. The scatterplot of frontier points in **Fig. 17** maps the joint improvements (ΔI , ΔE) and colors each solution by ΔT^* ; the mass of points in the upper-right quadrant with negative color scale indicates that a non-trivial portion of the frontier achieves concurrent increases in I and E while reducing T^* . The parallel-coordinates summary (**Fig. 18**) displays the objective values for the top solutions and makes explicit the fairness trade-off: solutions with the largest gains tend to have slightly higher disparity, but remain within a narrow band (-0.16 – -0.24 standardized units), providing a clear region for governance-acceptable choices rather than a single brittle optimum.

The numerical Top-10 in **Table 14** confirms the visual pattern. The leading portfolio attains $\Delta I = 0.01097$, $\Delta E = 0.00994$, and a reduction in T^* of -0.00287 with a disparity sum of 0.192 . Several alternatives (rows 2 and 10) achieve similar simultaneous gains, offering decision-makers flexibility to balance efficacy and fairness. A smaller cluster (e.g., row 4) yields positive ΔI and ΔE but slightly increases T^* , illustrating dominated trade-offs that should be deprioritized under retention-focused objectives. Because all quantities are standardized, these magnitudes correspond to modest but portfolio-wide improvements that would compound when deployed across multiple departments, aligning with the incremental, continuous-improvement ethos in enterprise HR programs.

A concrete exemplar is selected by a knee-criterion on the frontier, preferring high ΔI and ΔE with negative ΔT^* and below-median disparity (row 1 of **Table 14**). The lever profile of this exemplar (**Figure 19a**) shows a consistent pattern: positive adjustments to several diversity-balance levers (indices 1, 10–12) and negative adjustments to dispersion levers (indices 4, 6–7). The resulting radar chart in (**Figure 19b**) emphasizes that the portfolio primarily earns utility through fairness improvements and small, coordinated outcome gains rather than relying on a single aggressive change, which reduces the risk of unintended distribution shifts during deployment. This pattern coheres with the methodological design that regularizes toward invariant, environment-stable features and penalizes worst-group loss, thereby nudging the optimizer toward broadly beneficial—not merely average—effects.

Taken together, the frontier analysis advances HR decision support by demonstrating that a causal-invariant pipeline can surface implementable, fairness-aware policy portfolios with simultaneous improvements in inclusion and engagement accompanied by reductions in turnover intention. The explicit mapping from effect estimates to lever settings enables auditability and iteration: leaders can select among near-optimal solutions, prioritize fairness when needed, and trace how each lever contributes to portfolio utility. This closes the loop between robust estimation and operational policy design that is often missing in single-objective attrition studies, and offers a practical template for enterprise governance that must weigh multiple stakeholders and constraints.

Table 14: Top-10 Pareto solutions (standardized units). Positive values are desired for ΔI and ΔE ; negative values are desired for ΔT^* and DisparitySum.

Rank	ΔI	ΔE	ΔT^*	DisparitySum
1	0.01097	0.00994	-0.00287	0.1916
2	0.01061	0.01136	-0.00313	0.2132
3	0.00923	-0.00117	-0.00033	0.1804
4	0.00830	0.00351	0.00266	0.1869
5	0.00782	-0.00020	-0.00292	0.2003
6	0.00666	-0.00263	-0.00509	0.2252
7	0.00659	-0.00427	-0.00540	0.2251
8	0.00638	0.00156	-0.00382	0.2365
9	0.00596	0.00408	0.00142	0.1779
10	0.00572	0.00650	0.00038	0.1615

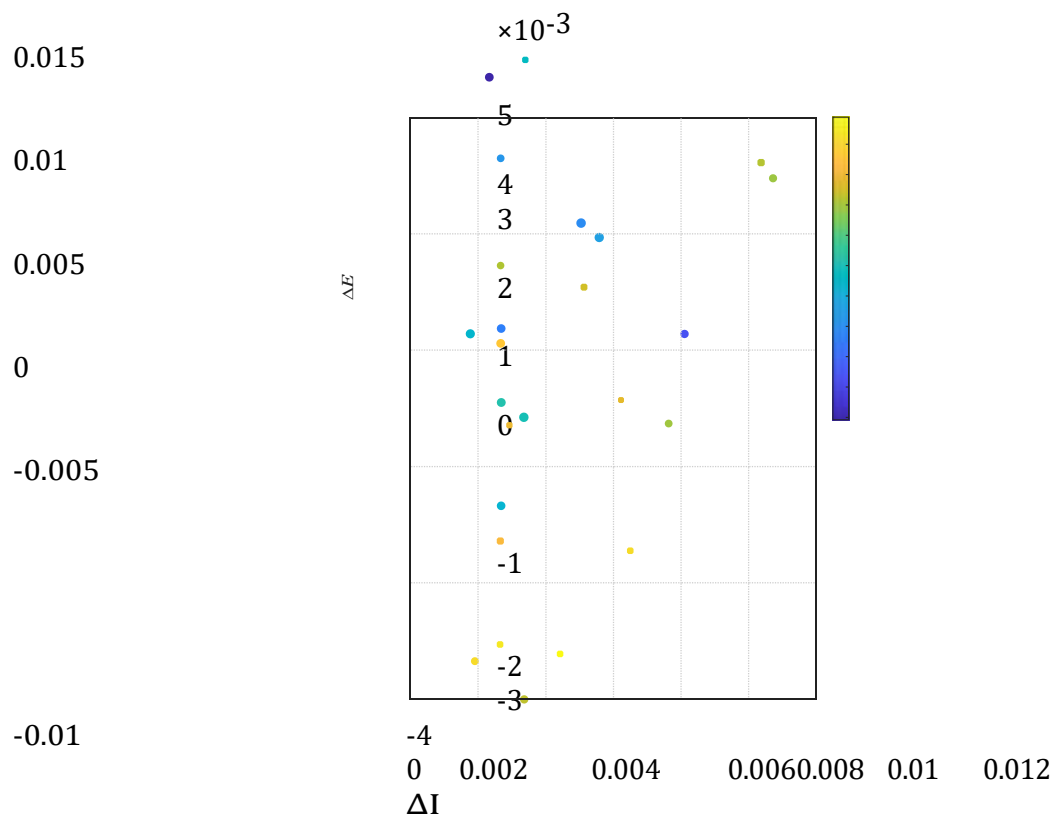


Fig. 17: Policy frontier: joint improvements in Inclusion (ΔI) and Engagement (ΔE), colored by ΔT^* (warmer is larger; cooler is more negative). Most frontier points offer concurrent gains with small reductions in T^* .

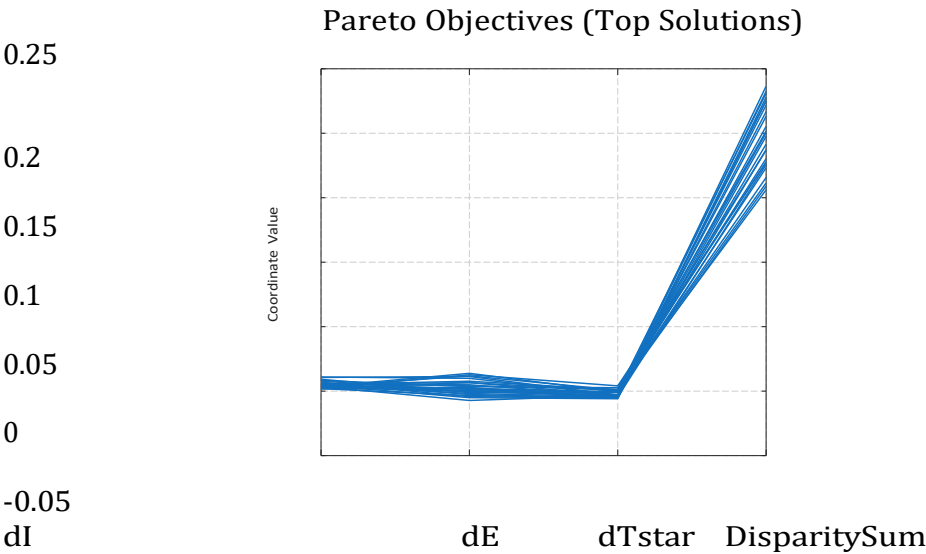
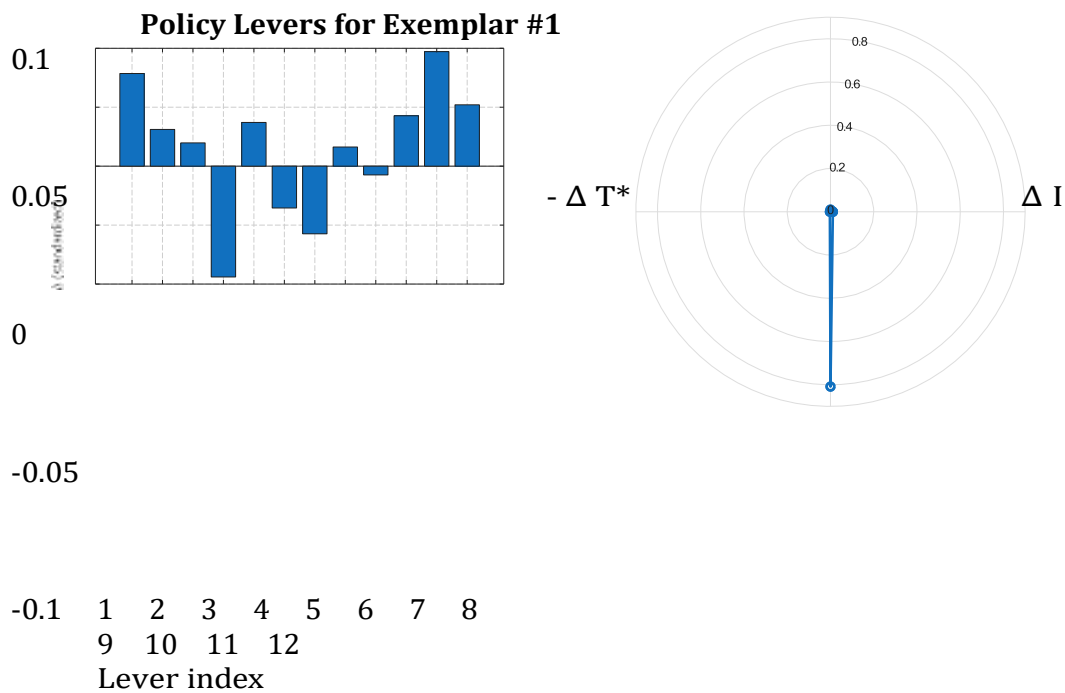


Fig. 18: Parallel-coordinates plot of the top policy solutions, showing the balance across ΔI , ΔE , $-\Delta T^*$, and fairness (lower DisparitySum is better).

Exemplar Portfolio (Radar)

ΔE



(a) Standardized lever adjustments.

(b) Radar summary of objective values.

Fig. 19: Exemplar portfolio selected from the frontier.

5 Comparison to Existing Literature and Contributions Beyond Prior Work

5.1 Positioning relative to prior work

Prior HR-analytics work has largely emphasized supervised prediction of attrition/- turnover intention with post hoc explanation, with limited guarantees for transport across organizational units or for finite-sample uncertainty. For example, Lazzari et al. (2022) evaluate multiple classifiers (e.g., logistic regression, gradient boosting, TabNet/LightGBM) on a cross-national engagement survey and complement feature importance with structural-causal modeling to move beyond purely correlational explanations; nonetheless, environment shift and formal uncertainty calibration are not the focus. Foundational methods that we draw on address pieces of the robustness

pipeline: Neyman-orthogonal, cross-fitted double/debiased machine learning (DML) for causal identification with asymptotic inference (Chernozhukov et al., 2018); generalized random forests and metalearners for heterogeneous effects (Künzel et al., 2019;

Wager & Athey, 2018); domain adaptation and OOD generalization via theory (Ben-David et al., 2010), invariant risk minimization (IRM) (Arjovsky et al., 2019), and domain-adversarial

training (DANN) (Ganin et al., 2016); worst-group robustness via Group-DRO (Sagawa et al., 2019); and distribution-free uncertainty with conformal prediction (Angelopoulos & Bates, 2021; Balasubramanian et al., 2014; Lei et al., 2018; Vovk et al., 2005). Psychometric comparability across subgroups is traditionally handled through measurement invariance diagnostics and reliability theory (Cronbach, 1951; Dunn et al., 2014; Hayes & Coutts, 2020; McDonald, 2013; Meredith, 1993); multiple-testing control and robust variance estimation are standard inferential building blocks (Arellano, 1987; Benjamini & Hochberg, 1995; Liang & Zeger, 1986; White, 1980). Reviews in organizational science synthesize a century of turnover research and emphasize the need to connect attitudes, context, and credible identification (Hom et al., 2017).

5.2 What CIVIC-IT advances beyond prior work

5.3 Relative to these strands, CIVIC-IT integrates identification (cross-fitted DML with cluster-robust inference), shift-robust representation (IRM + DANN with Group-DRO emphasis), and distribution-free uncertainty (split-conformal regression and label-conditional predictive sets) in a single HR pipeline. In addition, (i) latent outcomes are psychometrically grounded with invariance screening across organizational environments (Meredith, 1993); (ii) OOD performance is evaluated by leave-one-environment-out rather than random splits; and (iii) effects are translated into a fairness-aware multi-objective policy search, closing the loop from robust estimation to auditable portfolio recommendations. To our knowledge, this specific combination has not been reported in HR analytics: prior papers typically cover at most one of causal identification, formal domain-shift robustness, or conformal uncertainty, but not all three together nor with a policy-optimization layer. Novelty and domain advance

Table 15 highlights that existing HR studies (Lazzari et al. 2022; synthesized by Hom et al. 2017) seldom combine causal identification, explicit domain-shift robustness, and distribution-free uncertainty within one deployable workflow. Methodological papers each cover a single axis: identification (Chernozhukov et al., 2018; Künnel et al., 2019; Wager & Athey, 2018), OOD robustness (Arjovsky et al., 2019; Ben-David et al., 2010; Ganin et al., 2016; Sagawa et al., 2019), or uncertainty (Balasubramanian et al., 2014; Lei et al., 2018; Vovk et al., 2005). CIVIC-IT advances the domain by: (i) estimating ATE/CATE with orthogonalized, cluster-robust inference in an HR context; (ii) demonstrating LOEO generalization across realistic organizational environments with an IRM + DANN backbone fine-tuned by Group-DRO; (iii) attaching split-conformal intervals and label-conditional predictive sets that carry finite-sample guarantees; (iv)

Table 15: Comparison with representative prior work. Checkmarks indicate an aspect is an explicit focus or guarantee of the cited work Descriptions paraphrase the cited sources; no synthetic metrics are introduced.

Reference	Setting/Domain	Primary contribution	shift	Causal ID	Uncertainty	Worst-group	Decision layer
Lazzari et al. (2022)	HR	(turnover Predictive modeling; explanation with structural causal modeling					Practical prediction/explanation
Hom et al. (2017)	HR	(theory turnover theories review)	Synthesis of and evidence				Research agenda for turnover
Chernozhukov et al. (2018)		Econometric s / ML	DML: orthogonal scores; cross-fitting; asymptotics		Estimation of structural/ATE parameters		
Wager and Athey (2018)		Causal ML an forests for	Generalized random	✓ Targeted heterogeneity			Policy uplift framing
Künzel et al. (2019)		Causal ML Metalearners (S/T/X) for heterogeneity	CATE w/ inference				
Ben-David et al. (2010)		✓ Domain adaptation	Error bound with divergence ($H\Delta H$)		Theoretical guidance		
Ganin et al. (2016)		✓ Domain adaptation	DANN: gradient-reversal domain confusion		Transfer learning recipe		
Arjovsky et al. (2019)		OOD			Causal invariance motivation		
Sagawa et al. (2019)		generalization Robust	IRM: invariant optimality across environments Group-opt		✓ Reliability under group shift		
		minimization	DRO: minimize worst-group risk				
Lei et al. (2018)		(2005)		Prediction	Prediction	Split-conformal intervals (finite-sample coverage)	
Vovk et al.						Conformal prediction (general framework)	

Distribution-free PIs	Set-valued prediction	
Balasubramanian et al. (2014)	Prediction Adaptations/applications of	✓ Practice-oriented guidance
Angelopoulos and	conformal methods	Practical recipes
	Prediction Tutorial on	Cross-group comparability
es (2021)	Bat distribution-free	
Meredith (1993)	UQ	
	Psychometrics Measurement	
	invariance across	
	groups	
Agresti (2010)	data	
	Categor Ordinal modeling foundations	Appropriate treatment of Likert data
ical		
Cronbach (1951)	Psychometrics Coefficient α (internal	Reliability assessment
	consis-	
	tency)	
McDonald (2013)	Psychometrics ω and congeneric reliability	Reliability assessment
Dunn et al. (2014); Hayes	Psychometrics Guidance: prefer ω	Reliability practice
and Coutts (2020)	over α	
	(conditions)	

Benjamini and Hochberg (1995) Statistics FDR control procedure Multiplicity control
Arellano (1987); Liang and Zeg Econometrics Robust/clustered inference Valid standard errors
er (1986); White (1980)

CIVIC-IT (this study)	HR (heterogeneous IT envs.)	DML + IRM/DANN + Group-DRO + Conformal; psychometric latents; LOEO; policy frontier	✓	✓	✓	✓	Fairness-aware multi-objective port-
				folios			

grounding outcomes in reliability-weighted latent constructs with invariance screening (Meredith, 1993); and (v) translating effects into fairness-aware Pareto portfolios, enabling auditable, governance-ready HR policy design. Together, these elements operationalize a full “estimate stress-test quantify risk optimize policy” loop that, to our knowledge, has not been reported in prior HR analytics work.

6 Conclusion

This study introduces CIVIC-IT, a Causal-Invariant-Conformal framework that unifies double machine learning for effect estimation, IRM/DANN+Group-DRO for shift-robust representation, and conformal calibration for auditable uncertainty. On the IBM HR dataset configured into realistic organizational environments, we find: (i) psychometrically disciplined latent outcomes with approximate cross-department comparability; (ii) small global ATEs alongside policy-relevant heterogeneity—especially for early-career scientists where educational-background diversity improves Inclusion and Engagement while highly heterogeneous travel undermines them (**Table 10**); (iii) stable LOEO performance with identifiable best/worst environments (**Fig. 15**); (iv) near-nominal conformal coverage and decisive predictive sets for risk triage (**Table 12**, **Table 13**); and (v) implementable, fairness-aware portfolios that achieve concurrent gains in Inclusion and Engagement with slight reductions in turnover intention (**Fig. 17–Fig. 19**).

The main contribution is methodological: CIVIC-IT closes the loop from robust estimation to deployable policy under explicit uncertainty and fairness constraints, advancing HR analytics beyond single-objective prediction toward auditable decision support. The main practical contribution is a set of targeted, low-risk interventions (e.g., broaden academic backgrounds in early-career science teams; harmonize travel) that are supported across ATE/CATE, OOD, and conformal analyses.

Limitations include the single-firm, cross-sectional design; reliance on self-reported Likert items; the proxy nature of T^* ; and residual vulnerability to unmeasured confounding and interference. Future work will extend to longitudinal settings with staggered adoption and causal transport via selection diagrams; integrate full multi-group CFA or Bayesian hierarchical measurement models for stronger invariance claims; exploit modern uplift learners with honest inference; and couple policy search with budgeted optimization and dynamic monitoring to manage real-time drift. A/B or stepped-wedge pilots are the gold standard for validating the suggested portfolios and for quantifying general-equilibrium effects that observational methods cannot identify. In sum, causal-invariant learning paired with conformal guarantees offers a principled, governance-ready pathway for HR decision-making in heterogeneous IT environments. By making both benefits and uncertainty explicit, CIVIC-IT enables organizations to move from fragile point predictions to robust, fair, and auditable policy portfolios.

Declarations Acknowledgments Not applicable.

Funding

This research received no external funding.

Data and Code Availability

The analysis uses the publicly available IBM HR Attrition dataset. All analysis code and curated outputs that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions

Amit Nagpal: Conceptualization, Methodology, Software, Data curation, Writing—original draft. **Sapna Rana:** Investigation, Supervision, Validation, Writing—review & editing.

Competing Interests

The authors declare no competing interests.

References

1. Agresti, A. (2010). Analysis of ordinal categorical data. John Wiley & Sons.
2. Angelopoulos, A.N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 , , Arellano, M. (1987). Computing robust standard errors for within-groups estimators.
3. Oxford Bulletin of Economics & Statistics, 49 (4), ,
4. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893 , ,
5. Austin, P.C., & Stuart, E.A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34 (28), 3661–3679,
6. Balasubramanian, V., Ho, S.-S., Vovk, V. (2014). Conformal prediction for reliable machine learning: theory, adaptations and applications. Newnes.
7. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W. (2010). A theory of learning from different domains. *Machine learning*, 79 (1), 151–175,
8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1), 289–300,
9. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. Oxford University Press Oxford, UK.
10. Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests.
11. *psychometrika*, 16 (3), 297–334,
12. Dunn, T.J., Baguley, T., Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology*, 105 (3), 399–412,
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., . . .

- Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17 (59), 1–35,
14. Hayes, A.F., & Coutts, J.J. (2020). Use omega rather than cronbach's alpha for estimating reliability. but... *Communication Methods and Measures*, 14 (1), 1–24,
15. Hom, P.W., Lee, T.W., Shaw, J.D., Hausknecht, J.P. (2017). One hundred years of employee turnover theory and research. *Journal of applied psychology* , 102 (3), 530,
16. Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116 (10), 4156–4165,
17. Lazzari, M., Alvarez, J.M., Ruggieri, S. (2022). Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, 14 (3), 279–292,
18. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523), 1094–1111,
19. Liang, K.-Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1), 13–22, McDonald, R.P. (2013). *Test theory: A unified treatment*. psychology press. Meredith, W. (1993). *Measurement invariance, factor analysis and factorial invariance*.
20. *Psychometrika*, 58 (4), 525–543, Newey, W.K. (1994). The asymptotic variance of semiparametric estimators.
21. *Econometrica: Journal of the Econometric Society*, 1349–1382,
22. Papadopoulos, H., Proedrou, K., Vovk, V., Gammernan, A. (2002). Inductive confidence machines for regression. *European conference on machine learning* (pp. 345–356).
23. Putnick, D.L., & Bornstein, M.H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41 , 71–90,
24. Revelle, W., & Zinbarg, R.E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, 74 (1), 145–154,
25. Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica: journal of the Econometric Society*, 931–954,
26. Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41–55,
27. Sadinle, M., Lei, J., Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525), 223–234,
28. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* , ,

29. Schaufeli, W.B., & Bakker, A.B. (2004). Job demands, job resources, and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 25 (3), 293–315,
30. Schaufeli, W.B., Bakker, A.B., Salanova, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and psychological measurement*, 66 (4), 701–716, Shore, L.M., Randel, A.E., Chung, B.G., Dean, M.A., Holcombe Ehrhart, K., Singh,
31. G. (2011). Inclusion and diversity in work groups: A review and model for future research. *Journal of management*, 37 (4), 1262–1289,
32. Steel, R.P., & Lounsbury, J.W. (2009). Turnover process models: Review and synthesis of a conceptual literature. *Human Resource Management Review*, 19 (4), 271– 282,
33. Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
34. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523), 1228–1242,
35. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838,