

Signals in the Transcript: Comparing AI and Human Predictions of Worker Retention

Bernard Shen

University of Chicago • July 2026 • blshen@uchicago.edu

Senior Thesis

Worker turnover is a first-order cost for high-volume employers: do big data and advanced machine learning (ML) models offer a solution? As screening processes become increasingly automated, firms have access to more information than ever to assess candidate quality, and tools that can process it at scale. I ask whether the linguistic content of an interview transcript predicts retention, and whether ML models can identify who will stay more accurately than the human evaluators presently making the firm's hiring decisions. This document summarizes the experimental design, its rationale, and the project's qualitative findings. The full manuscript is withheld (see *Data availability*).

SETTING AND DATA

The analysis draws on roughly twelve thousand interview transcripts from a recruitment firm (most AI-conducted). The retention models are fit on the hired subsample, where a 90-day outcome is observed, using a three-fold expanding-window split by interview date that mirrors deployment constraints.

MODELING STRATEGY AND RATIONALE

Two-stage residual design. As in a single boosting round, the transcript models are trained not on the outcome directly but on a structural baseline's residuals: the variation in retention that the structural features cannot explain.¹ By first establishing a performance floor, this architecture cleanly isolates the predictive lift provided by just the transcript, and the two-stage approach prevents the pipeline from fitting spurious correlations between structural and transcript features.²

Structural baseline. A shallow, lightly boosted XGBoost model is fit on a small set of structural applicant features available at the point of hire (sourcing, site, interview modality, and pre-interview assessment scores). The restricted capacity balances two competing objectives: it must be tight enough to prevent overfitting and leave stable in-sample residual targets for the transcript models, yet loose enough that the baseline is not artificially weakened.³

Structured linguistic features (SLF). The first transcript model is an XGBoost fit on ten hand-engineered linguistic and content features. The goal is to establish that a transcript-derived signal *exists and is recoverable with lightweight, transparent tooling*, not to offer a production-optimal representation. All ten features are inherited from prior causal research on this dataset; none were screened, tuned, or added for predictive performance.⁴

Fine-tuned large language model. The second transcript model fine-tunes LLaMA-2-7B on the same residual target using a linear projection head, testing whether a learned representation recovers signal the hand-engineered features miss.⁵

Common protocol. Hyperparameters are selected from related literature and held fixed across folds so that performance differences reflect temporal variation rather than fold-specific optimization.

IDENTIFICATION CHALLENGES

Because retention is observed only for hired candidates, the unconditional screening question is unidentified: for applicants the model would have hired but the firm did not, no outcome was ever realized. The project therefore focuses on the conditional question, whether the model's choices would yield greater retention *among the candidates the firm actually hired*.

¹Training on in-sample rather than out-of-sample residuals is a concession to sample size: holding out data for this purpose would have sacrificed already-scarce negative outcomes in the training/test sets. The choice is validated by the tight correlation between the baseline's in- and out-of-sample predictions.

²A single model trained jointly on all features underperforms the two-stage specification in every fold, confirming the regularization benefit of this separation.

³Refitting with package defaults fails to improve out-of-sample discrimination while greatly widening the generalization gap, confirming that the extra capacity merely fits noise.

⁴The features were selected on the full sample rather than the training folds alone, so some risk of leakage remains. That risk is mitigated by how the inventory was constructed, which was primarily on the basis of established NLP research.

⁵The model is fine-tuned via LoRA (Low-Rank Adaptation), chosen for its data efficiency on small folds, and framed as an instruction-based task. Base weights are 4-bit quantized (QLoRA) to fit on a single NVIDIA T4.

A flawed instrument for the marginal hire. One approach to recovering the effect of hiring on the marginal applicant is a judge-leniency design, which leverages cross-reviewer variation in offer rates within a job code to instrument the hire decision. I implement and reject this design because candidate assignment to reviewers is not as-good-as-random.⁶ Consequently, the exercise serves as a blueprint rather than a usable design: to deploy this instrument successfully, future data collection must randomize candidate-to-reviewer assignment, or at minimum stratify it on observables.

Approximating human judgment via propensity scores. To determine whether human evaluators already use the retention signal present in the transcripts, I train an XGBoost model on the full feature set (structural baseline and SLF) to estimate each candidate's probability of receiving an offer. As offers prove highly predictable, this score closely reproduces the firm's hiring decisions, and feature importances reveal which inputs those decisions track. The propensity model places almost no weight on the SLF. The transcript models are therefore not re-deriving the firm's implicit judgment; they recover retention signal the pipeline leaves largely unused.⁷

FINDINGS AND EVIDENCE

Transcripts carry retention signal beyond structural features. The SLF model lifts discrimination over the structural baseline, with significance established by a DeLong test on AUC-ROC and a paired bootstrap on average precision.⁸

Interpretable features rival the fine-tuned language model. Pooled across folds, the SLF model and LLaMA-2-7B converge to statistically indistinguishable discrimination on both AUC-ROC and average precision. Thus, transparency does not appear to cost accuracy at this data scale.⁹

The signal is temporally fragile. Discrimination collapses in the final evaluation window and two diagnostics constrain the explanation. Reweighting training toward recent cohorts via a half-life sweep produces no structured improvement, so staleness does not appear to be the cause. A feature-by-time interaction test marginally fails to reject stationarity, leaving no clear evidence that the mapping drifted within the training window. What remains is a base-rate (prior-probability) shift: churn in the final window falls to less than half its training-period average, leaving too few events to distinguish a real loss of signal from insufficient power to detect one.¹⁰

DEPLOYMENT IMPLICATIONS

The models' outputs are validated as rankings, not as probabilities: calibration is never assessed, and a score fitted at one churn rate would misstate absolute risk at another. Aggregate churn forecasting is therefore out of scope, and because regime shifts occur in this setting with untested consequences for ranking, any deployment must monitor for them. Within a hiring cohort, the supported use case is recovering retention signal the pipeline leaves unused, which makes the transcript scores well suited to a retention-aware layer on top of existing screening. Whether they could replace that screening is a question the data cannot answer.

DATA AVAILABILITY

The transcripts and employment records underlying this project are proprietary and governed by a university data-use agreement that prohibits redistribution. To protect the individuals interviewed and the partner firm, the manuscript and data are not publicly released; this note reports methodology and qualitative findings only, with further detail available on reasonable request.

Keywords: worker retention, algorithmic hiring, interview transcripts, large language models, selective labels, distribution shift.

⁶This failure of conditional independence is evidenced by severe covariate imbalance and a wrong-signed reduced form (lenient reviewers' hires retain better). Separately, the design also suffers from a weak first stage on the full applicant pool.

⁷Restricting to the hired cohort, where retention is observed, I regress the 90-day outcome on each standardized score. The propensity score's null is consistent with Berkson-type selection on unobservables and does not necessarily imply the firm ignores retention entirely, only that it ignores the transcript's specific signal.

⁸The lift is robust across 30-, 60-, and 120-day retention horizons and to a "superstar" relabeling of candidates who received but declined offers, whose decision suggests stronger outside options.

⁹Fold-level performance crosses over, with the language model ahead in the smallest fold and the SLF model ahead as data accumulate; the difference is not statistically significant.

¹⁰A hyperparameter grid search returns essentially no gain, making a tuning failure unlikely. The stationarity test is run on the training window preceding the collapse, so it has limited power against an abrupt break located in the test window itself. Under a match-quality reading, the interview measures a candidate-side component distinct from the firm-side environment that sets the churn level; the data are consistent with this interpretation but cannot test it.