Large Language Models for Data Extraction Among 4,639 California Lifer Parole Hearings.

Cameron Golbahary* Lowell Justice†

Project advisor: Deanna Needell‡ Mentor: Chi-Hao Wu Consultant: Mike Semanchik

November 2026

Abstract

The California Board of Parole Hearings decides the outcomes of thousands of people serving life sentences each year, yet the factors that drive its decisions are hard to study at scale. We assemble a corpus of 5,265 California parole suitability hearing transcripts (August 2024–December 2025) and use a large language model (Claude Sonnet) to extract roughly thirty structured case characteristics per hearing, validated against hand-coding at approximately 99% agreement. We find differential reliability among demographic inferences with confidence levels by race ranging from 23.6% to 77.8%. A calibrated random-forest model predicts outcomes at an AUC of about 0.86, and adding demographic or commissioner information changes accuracy negligibly.

1 Introduction

The California Board of Parole Hearings holds thousands of parole suitability hearings per year, many of which are for incarcerated people serving life sentences. At each hearing, a panel of commissioners decides whether to grant parole based on a number of factors, including criminal history, institutional behavior, release plans, and remorse for the offense [1]. The "remorse" requirement creates a dilemma for incarcerated people who claim innocence. To gain parole they may feel pressure to express remorse for an offense they say they did not commit, but if they continue to claim innocence, the panel may interpret this as "lack of insight into the life crime" and deny parole on those grounds.

As of 2024, around 37,000 people are serving life sentences in California (39% of total prison population), of whom roughly 30,000 are eligible for parole [11]. A 2019 study of incarcerated individuals in Pennsylvania found that 6% reported having been wrongfully convicted of their controlling offense [9].

Our paper uses Claude Sonnet to extract JSON-structured data from 17 months of California parole hearings for incarcerated people serving life sentences. We use propensity-matching analysis to estimate how individual case factors are associated with parole outcomes in California, holding other factors constant. We found that claiming innocence, vocational training, comprehensive risk assessment score, conduct, sobriety, gang affiliation, victim involvement, and documented support

*University of California – San Diego, Department of Mathematics, cgolbahary@ucsd.edu

†Wesleyan University, Department of Mathematics, ljustice@wesleyan.edu

‡University of California, Los Angeles, Department of Mathematics

are all associated with grant rate after matching. These results provide insight into the parole hearing process through a quantitative lens.

2 Data

2.1 Extraction

Our dataset comprises 17 months of California parole hearing transcripts for incarcerated people serving life sentences, covering cases held from August 2024 to December 2025. This consisted of 5,265 hearings, 626 of which were postponed or otherwise undecided, leaving an analysis sample of 4,639 cases, 1,012 (21.8%) of which resulted in a grant. We used the Claude Sonnet 5 API to extract structured information from the transcripts in JSON format. This consisted of 27 case factors along with LLM reasoning and confidence levels, advised by our community partner. An abbreviated example record can be seen below (see the supplemental appendix for the complete list of fields):

```
{
    "commissioner_names": ["Presiding Commissioner", "Deputy Commissioner"],
    "outcome": "Granted",
    "victim_involvement": ["present", "statement"],
    "innocence_claimed": "Yes",
    "innocence_claim_text": "I did not commit the crime I was convicted of. I am innocent.",
    "innocence_confidence": 99,
    "race": "middle eastern",
    "lack_of_insight": "Yes",
    ...
}
```

We hand-coded an initial set of 30 transcripts, ran our extraction script, checked for agreement, and refined the prompt. We iterated until output was consistent at an agreement rate of ∼99%.

2.2 Unrecorded data

Any field where the desired information is not discussed in the transcript is coded "null" and treated as unrecorded in our analysis. For example, if an applicant's criminal history is not discussed, it will be recorded as `"criminal_history": null` and will be treated separately from an applicant for whom the panel establishes there is no record prior to the life crime. The innocence claim and lack of insight fields default to "No" if there is no mention, and the age field defaults to "adult" if there is no mention of youth or elderly parole considerations. Among hearings longer than 30 minutes, those under an hour have substantially more null fields. Section 4.1.11 reports a correspondingly low grant rate among hearings under an hour. We note that applicants stipulate unsuitability in many of these hearings, which may deflate the raw grant rate associated with null value count and short hearings.

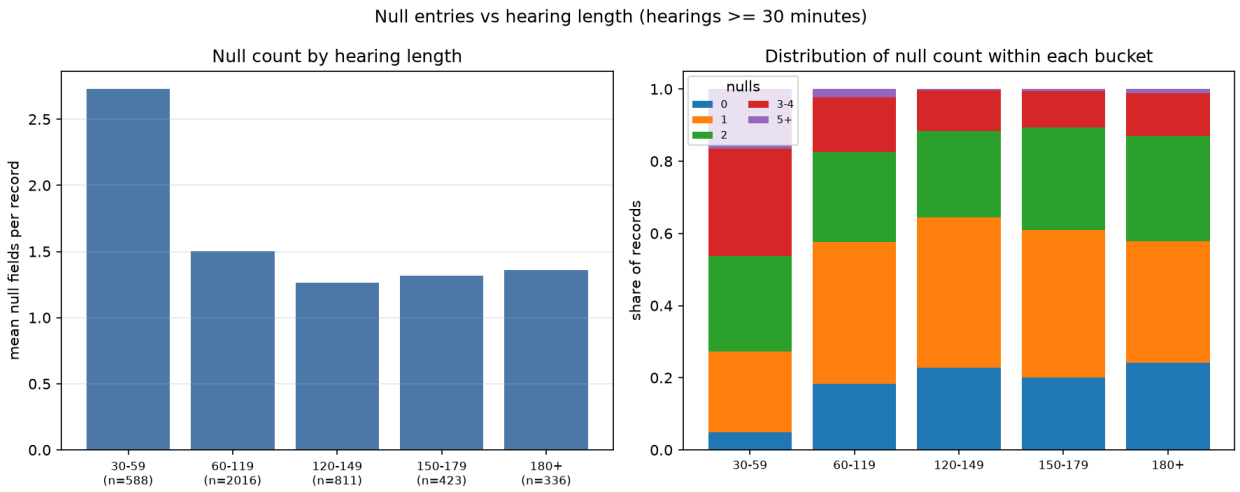


Figure 1: Mean null entry count by hearing length (left) and distribution of null count by hearing length (right).

A chi-square test (see Section 3.5) rejects the hypothesis that unrecorded fields are distributed evenly across hearing lengths with Cramér's $V = 0.18$ and $p < 0.0001$.

2.3 Demographic data

During data extraction, the LLM was prompted to report its confidence in its race and gender output on the following integer scale:

```
95-100 = stated directly in the transcript
70-94 = inferred from multiple strong signals
40-69 = inferred from a single weak signal
0-39 = little to no evidence
```

We found that gender data was generally highly confident with 99.9% of output at a confidence of at least 70, while race confidence varied greatly. Applying a confidence threshold of 70 retains race labels at markedly different rates across racial groups:

Recorded race	Confidence ≥ 70	Total	Rate
Asian	119	153	77.8%
Middle Eastern	10	13	76.9%
Native American	27	41	65.9%
Hispanic/Latino	1,075	1,831	58.7%
Other	25	43	58.1%
Black/African American	167	634	26.3%
White	85	360	23.6%
Total	1,508	3,075	49.0%

Table 1: Hearings retained by the race confidence threshold.
Retention differs across groups in our sample: Cramér's $V = 0.34$, $p < 0.0001$ (see Section 3.5).

Figure 2 shows which signals the model used when inferring race, split by the race it recorded (left) and by its own confidence (right). We note that name/surname is the most common reasoning, and gang affiliation is used disproportionately to infer race for Black/African American applicants.

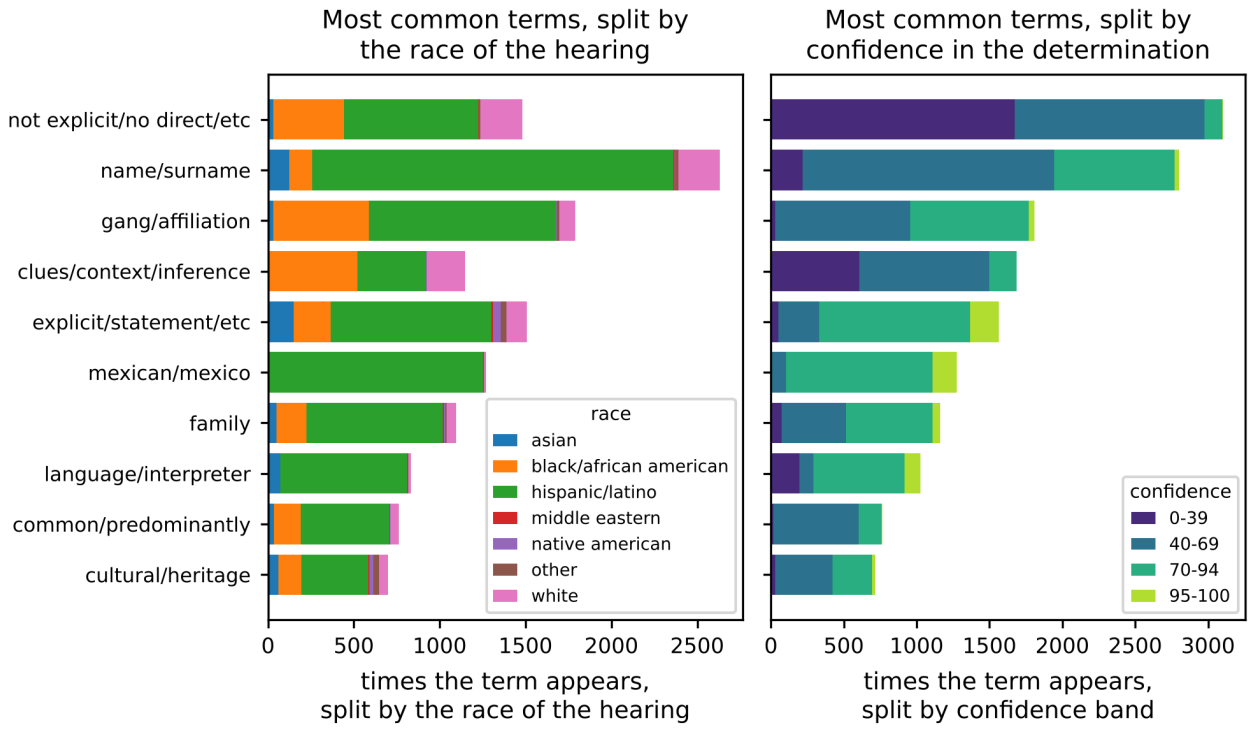


Figure 2: Most common phrases used in race reasoning, by race and confidence level.

In our race analysis, we report results using data at confidence of at least 70. Our predictive model (Section 3.6) treats demographic data with confidence under 70 as unrecorded.

3 Methods

We first compare raw grant rates (Section 3.1) with propensity-matched estimates that control for other case factors (Section 3.2), correcting for multiple comparisons (Section 3.3) and measuring uncertainty by bootstrapping (Section 3.4). We then test whether case factors differ across commissioners (Section 3.5) and finally build models that predict hearing outcomes (Section 3.6).

3.1 Grant rates

Grant rates are reported with a 95% Wilson score interval. Confidence intervals indicate how precisely our rate is measured given how many hearings it summarizes, where smaller samples produce larger intervals. The Wilson interval is preferred to the ordinary normal approximation because it remains within $[0, 1]$ and retains approximately nominal coverage for small n and for proportiona near 0 or 1 [17].

3.2 Propensity-score matching

We used propensity score matching to determine the associations of case factors with parole outcomes. Propensity matching uses the following procedure:

1. Identify a factor X that we want to observe.
2. Identify the covariates:

- Whether innocence was claimed.
- Risk assessment score.
- Criminal history (Yes/No).
- Gang affiliation (Yes/No/Former/not recorded).
- Time since last drug/alcohol use.
- Time since last misconduct.
- Time served.
- Sentence length.
- Number of past hearings.
- Age consideration (Youth/Elderly/ Adult/Youth and Elderly).
- Outside support.
- Applicant education level.
- Victim involvement.

Note that gender and race are left out of the covariates.

3. Fit a propensity score e_i for each hearing i: the probability that the hearing has factor X given its other case factors. This is estimated by logistic regression. Writing $T_i = 1$ when hearing i has the factor and $x_{i1}, \ldots, x_{ip}$ for its other case factors, the model is

$$\log\left(\frac{e_i}{1-e_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \qquad e_i = P(T_i = 1 \mid x_i). \tag{1}$$

4. Match each hearing with X to its nearest-score hearing that doesn't, without replacement. Matches are only valid if propensity scores are within 0.05.
   - When comparing greedy and optimal matching, we found that optimal found more matches than greedy in 63 of 75 factors, and the two methods tied in the other 12. We consequently decided to use optimal matching in our analysis.
5. Compare grant rates only within matched pairs, leaving out unmatched cases.
6. Report mean grant rate difference between the two groups.

This process controls for other factors in a hearing, so we can reduce confounding from measurable covariates [16, 14].

For example, applicants who completed vocational training may see an elevated grant rate, but vocational training could also be correlated with good institutional behavior, which could be artificially boosting our views of vocational training's association with outcomes. Matching lets us test whether that apparent association is explained by correlated covariates such as good behavior.

For some fields, we use baseline reference levels for our comparison rather than comparing each entry as its own factor:

1. Risk assessment scores are compared against "moderate".
2. Gang affiliation statuses are compared against an explicit "No Affiliation".
3. Education levels are compared against holding a high school diploma or GED.
4. Support types are compared against cases with no support.
5. Victim involvement is compared against having no victim involvement.
6. Time since last drug use/misconduct are compared against > 10 years.
7. Number of past hearings is compared against cases with one prior hearing.
8. Hearing length is compared against the modal entry.
9. Age considerations are compared against "adult" (no youth or elderly parole consideration).

When estimating the association of sobriety with grant rates, we also report an estimate that does not control for disciplinary write-ups, since consuming substances will almost always result in a write-up. This allows us to see the grant rate association of substance use aside from the write-up caused by it.

For each covariate c we measure how far apart the two groups are using the standardized mean difference,

$$\mathrm{SMD}_c = \frac{\bar{x}_{c,1} - \bar{x}_{c,0}}{\sqrt{\left(s_{c,1}^2 + s_{c,0}^2\right)/2}}, \tag{2}$$

where $\bar{x}_{c,1}$ and $\bar{x}_{c,0}$ are the mean values of covariate c among hearings with and without the factor, and $s_{c,1}^2$, $s_{c,0}^2$ are the corresponding variances. We compute Equation (2) twice for every factor, once between all hearings with and without the factor, and once within the matched pairs alone. If matching succeeds, the second is substantially smaller than the first. We report the number of covariates with $|\mathrm{SMD}| > 0.1$ after matching [2, 12].

We additionally use propensity-match analysis to estimate the association of innocence claims with mentions of lack of insight or minimization of the life crime. We conducted this analysis across all hearings and on hearings that ended in denial.

3.3 False discovery correction

We use the Benjamini-Hochberg correction at $\alpha = 0.05$ to control our false-discovery rate. After running our tests, we sort all n p-values in ascending order and each result $i = 1, \ldots, n$ is assigned a rank R_i. We calculate $t_i = \frac{R_i \times 0.05}{n}$ for each p-value and only reject the null hypothesis where $p_i < t_i$. This accounts for testing many case factors at once, which would otherwise inflate the false-positive rate [3].

3.4 Bootstrapping

We used bootstrapping (resampling) to measure the variability of our findings. For each sample size N of hearings we draw 15,000 new samples with replacement from the original sample, also of size N. We then report the mean and variance of these samples. The spread across resamples indicates how sensitive our estimates are to the particular hearings observed. We compute bootstrap percentile intervals at the 90%, 95% and 99% levels and treat a difference as significant at a given level when its interval excludes zero [7]. Paired resampling is used for the matched pairs to preserve paired dependence structure, while independent two-group resampling is used for raw differences.

3.5 Chi-square test for independence

We tested case factor independence between commissioners, the number of unrecorded fields across hearing lengths, and model confidence across races using a chi-square test of independence. In our commissioner analysis, the chi-square test measures the plausibility of the discrepancy between observed and expected counts of different case factors to determine if case characteristics are independent of the presiding commissioner hearing the case.

For a table of r commissioners by c levels of a case factor, with O_{ij} hearings observed in row i and column j, for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, row totals R_i, column totals C_j, and N hearings in total, the expected counts and the resulting statistic are:

$$E_{ij} = \frac{R_i C_j}{N}, \qquad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

which is compared to a chi-square distribution with $(r-1)(c-1)$ degrees of freedom. A large χ^2 signals that observed counts are far from what independence would produce [10].

Alongside each test we report Cramér's V, a measure of how strongly the two variables are associated which is invariant to sample size. We consider $V = 0.1$ as our benchmark for a detectable yet practically small imbalance and $V = 0.3$ as our benchmark for a moderate imbalance [6].

Our test for factor independence between commissioners only uses commissioners with at least 20 decided cases. We run two tests, the first checking independence in the following pre-determined case factors with FDR correction (Section 3.3):

- Risk assessment score.
- Past hearings.
- Education.
- Gang affiliation.
- Race.
- Criminal history.
- Support.
- Innocence claims.
- Conduct while incarcerated.
- Time served.
- Gender.
- Sentence length.
- Age.
- Victim involvement.
- Substance use.

and the second checking for independence in hearing length and mentions of lacking insight.

3.6 Predictive model

We fit a Platt-scaled random forest built to predict hearing outcomes based on case characteristics. Unlike the logistic models used elsewhere in this study, a forest can represent interactions and non-linear thresholds. This model gives each case a "score" based on its characteristics, then calibrates against other cases to output a probability that the applicant will be granted parole.

We assess calibration with expected calibration error (ECE), the average gap between predicted probabilities and observed grant rates. We use ECE to determine whether cases are scored in the correct "probability range," meaning the model does not over-or-under-score cases. We report accuracy in quantile bins alongside area under the ROC curve (AUC) and Brier score. AUC measures the probability that a randomly chosen granted hearing receives a higher score than a randomly chosen denied hearing, with 1.0 corresponding to perfect ordering and 0.5 corresponding to random ordering. The Brier score measures the mean squared difference between the predicted probability and the binary outcome, with lower scores being more reliable when using binary classification [8, 4].

We grew 500 trees with a maximum depth of 12, requiring at least 10 cases in each terminal node, and considering $\log_2(p)$ features at each split for p features in total. Because grants are the minority outcome, class weights inversely proportional to class frequency were applied when growing the trees. Of our 4,639 hearings we use 3,000 for training, 1,000 for calibration, and 639 for testing.

We use Platt-scaling calibration, where after hearings are scored a function is fitted from the score to the observed outcome,

$$P(\text{grant} \mid x) \ = \left(1 + e^{-(af(x)+b)}\right)^{-1}, \tag{4}$$

where a and b are scalars learned by the algorithm [13].

These settings were chosen via cross-validation, comparing 160 combinations of the number of features considered per split, the calibration method, the minimum terminal-node size, and the maximum tree depth. Cross-validated area under the curve (AUC) ranged from 0.835 to 0.866

across the 160 candidates. The number of features per split and the minimum terminal-node size accounted for most of this variation (0.010 and 0.008 in AUC, respectively).

We report the accuracy of models using varying factors in their analysis:

1. Base model: Uses the same features as our propensity matching covariates:

    - Whether innocence was claimed.
    - Risk assessment score.
    - Criminal history (Yes/No).
    - Gang affiliation (Yes/No/Former/not recorded).
    - Time since last drug/alcohol use.
    - Time since last misconduct.
    - Time served.
    - Sentence length.
    - Number of past hearings.
    - Age consideration (Youth/Elderly/ Adult/Youth and Elderly).
    - Outside support.
    - Applicant education level.
    - Victim involvement.

2. Base model + demographic information (race and gender).
3. Base model + presiding commissioner names.
4. Base model + demographic information + presiding commissioner names.

We created a simple calibrated additive model that uses the matched mean difference from propensity matching, considering the same factors as the base model. This model runs propensity matching (Section 3.2) on our training data, then adds the matched mean differences for chosen factors in a case (based on its respective configuration, as outlined above), where higher scores indicate a higher predicted likelihood of a grant. Writing $F(i)$ for the set of factors present in hearing i and $\hat{\tau}_f$ for the matched mean grant rate difference of factor f, the raw score is

$$s_i = \sum_{f \in F(i)} \hat{\tau}_f. \tag{5}$$

We check hybrid models which run the additive models on all cases then keeps the additive score within certain thresholds of 0 or 1 and runs random forest on the rest, with calibration applied at the end. We compared 36 basic hybrid models (using all combinations of 0.1, 0.15, 0.2, 0.25, 0.3, and 0.35 from 0 and 1). AUC and Brier scores were highly insensitive to threshold with ranges of 0.011 and 0.005 respectively. We chose to use a lower threshold of 0.1 and upper threshold of 0.35, since this will result in only the confident scores kept in the additive model, with neutral/mixed factor cases being assessed by the more robust random forest model. In our test set, about 40.7% of cases were kept by the additive model. We compare our forest model configurations against their hybrid counterparts using this threshold.

3.7 Reproducibility

All analyses were run in Python 3.14.6 using numpy 2.5.1, pandas 3.0.3, scipy 1.18.0, statsmodels 0.14.6, scikit-learn 1.9.0 and matplotlib 3.11.0. Random seeds are fixed for all resampling, matching, cross-validation, and model fitting. LLM output is not deterministic, so exact reproduction of the extraction step is not guaranteed.

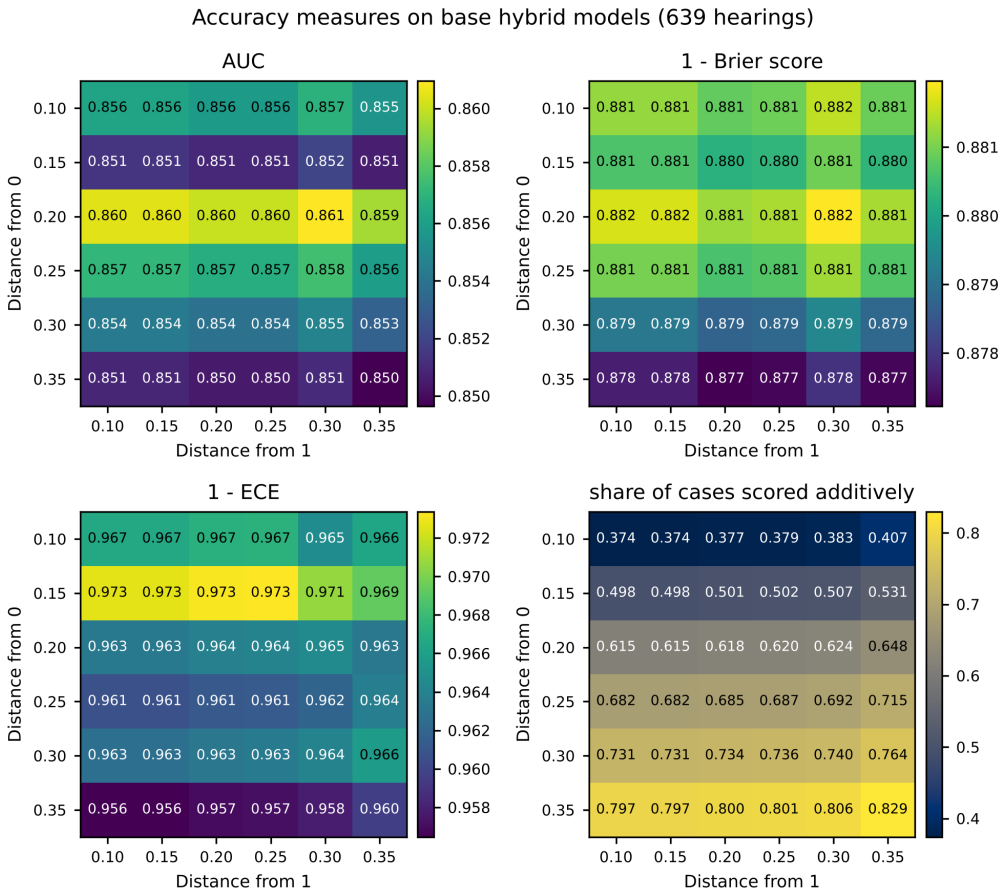


Figure 3: Graphs displaying AUC, 1−Brier score, and 1−Expected calibration error for all 36 hybrid models, where higher scores are better. "Share of cases scored additively" measures the share of cases that fell into the specified threshold when scored with the additive model. The remaining cases were scored using the base random forest model.

4 Results

4.1 Individual raw and matched associations

4.1.1 Innocence claims

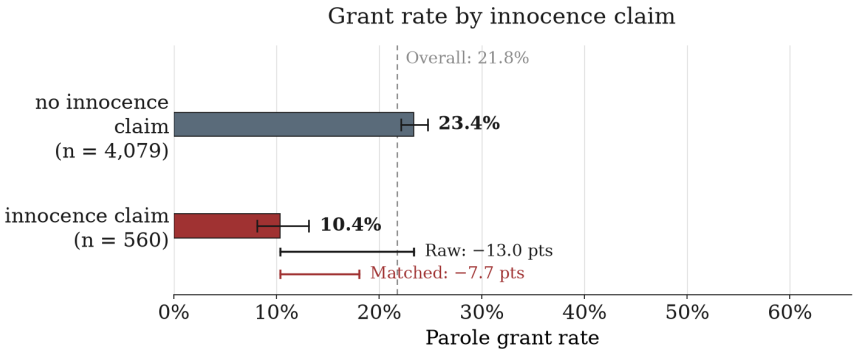


Figure 4: Raw and matched grant rate differences between innocence and non-innocence claimants, with non-claimants held as the reference points.

Innocence claims are associated with lower grant rates before and after matching. The grant rate difference shifts from -13.0 to -7.7 percentage points, meaning less than half of the raw association can be attributed to correlated case characteristics.

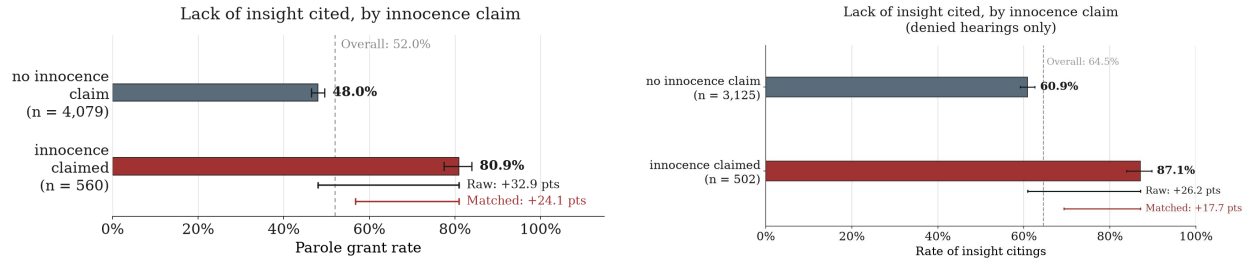


Figure 5: Raw and matched differences in rates of commissioners mentioning current lack of insight into the life crime among innocence claiming and non-claiming applicants.
Left: Analysis across all hearings. Right: Analysis is done only among denied hearings.

Among all hearings (parole-denied and parole-admitted), commissioners cited lack of insight in 52% of hearings, regardless of innocence claims. Commissioners mentioned lack of insight in roughly 81% of hearings where the applicant claimed innocence, compared to less than half of cases where the applicant did not claim innocence. The matched gap shrinks by around 25% from roughly 33 percentage points to roughly 24 percentage points, but there is still a clear matched gap in insight mentions between innocence and non-innocence claimants. Both raw and matched gaps shrink when considering only hearings which ended in denial. Non-innocence claimers see a roughly 61% rate of insight mentions, while innocence claimers see a 26.2 percentage point higher rate (when compared to non-innocence claimers). The matched gap is roughly 18 percentage points, still displaying a difference in mentions of lack of insight. Our findings are consistent with commissioners viewing innocence claims as a signal of lacking insight, regardless of other behavioral factors.

4.1.2 Criminal History

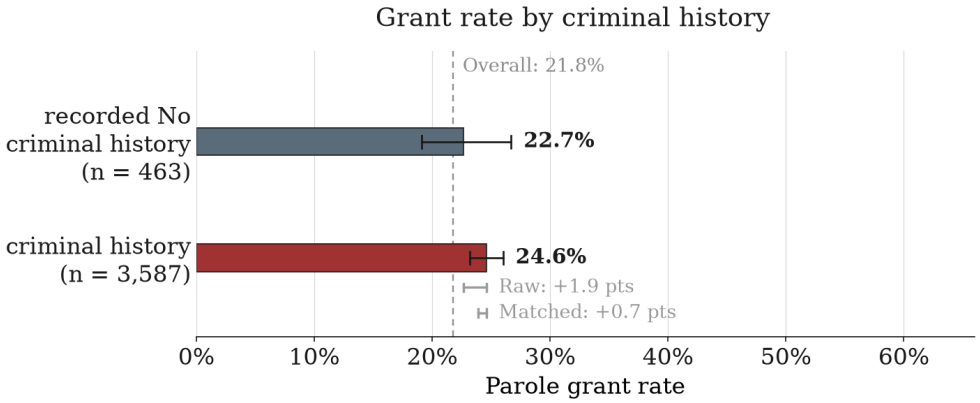


Figure 6: Raw and matched grant rate differences between applicants with and without a criminal history. Grey brackets represent insignificant results at $\alpha = 0.05$.

Criminal history is not associated with grant rates in our raw data or after matching. While those with criminal history saw a roughly 2 percentage point higher grant rate in our raw data, this is

statistically negligible. The matched estimated grant rate difference is marginal and statistically insignificant.

4.1.3 Gender and race

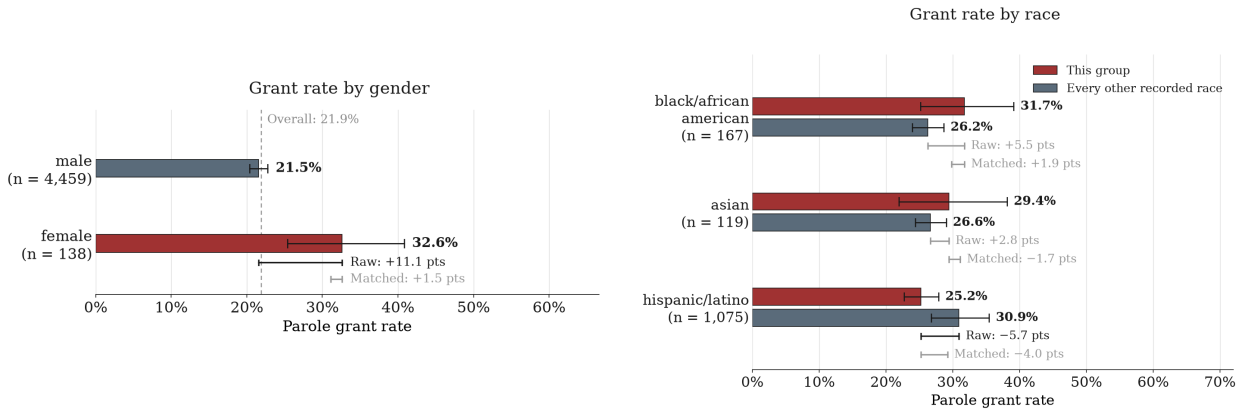


Figure 7: Left: Raw and matched grant rate differences between male and female applicants. Right: Raw and matched grant rate differences among Black/African American, Asian, and Hispanic/Latino applicants when compared against all other cases. Grey brackets represent insignificant results at $\alpha = 0.05$.

Raw grant rates differ between male and female applicants by roughly 11 percentage points, with this gap becoming insignificant after matching. Due to the small number of female applicants, these results should be read with caution. Applicants recorded as transgender male, transgender female, or other had sample sizes under 30 and were left out of our analysis. We consequently draw no conclusion about the association of gender with grant rates.

The only racial group whose raw grant rate differed significantly from other races was hispanic/latino. No group's estimate remains distinguishable from zero after matching. For Native American applicants, Middle Eastern applicants, and applicants marked as race: "other", matching was unsuccessful with at least 20 covariates above a standardized mean difference of 0.1.

4.1.4 Age

Neither elderly nor youth consideration displays a significant association with grant rates after matching, although both considerations and a combination of the two are correlated with higher grants. We note that commissioners are instructed to give extra consideration to elderly and youth applicants, but nearly all of this association can be explained by other correlated case characteristics.

During the comprehensive risk assessment, the psychologist is instructed to give extra mitigating effect to youth parole considerations [5]. In Figure 8 (right), we run propensity matching analysis to estimate the association of youth status with psychological risk score. Consider the "low moderate or better" category. We find that 26.4% of youth offenders hold a low or low moderate score at the time of hearing, while 27.9% of adult offenders hold one of these scores. This raw difference is statistically insignificant, and the difference remains insignificant after matching. These findings are consistent with no mitigating effect being given to youth or elderly parole applicants, either during the psychological risk assessment or at the time of hearing.

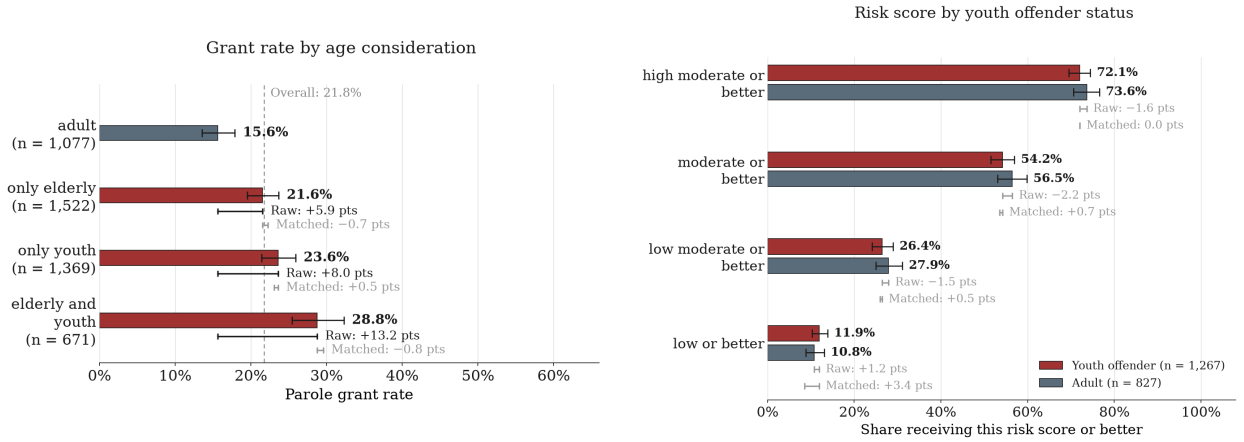


Figure 8: Left: Raw and matched grant rate differences among elderly and youth-eligible parole applicants when compared against adult applicants.
Right: Raw and matched differences in comprehensive risk assessment score between youth and adult applicants. Grey brackets represent insignificant results at $\alpha = 0.05$.

4.1.5 Education and vocational training

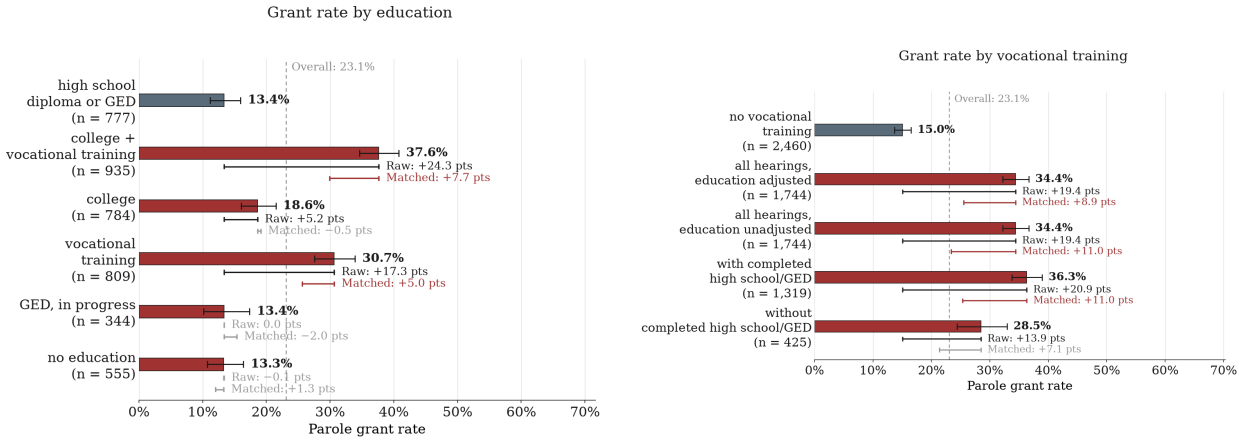


Figure 9: Left: Raw and matched grant rate differences among different highest education levels with a high school level of education used as reference.
Right: Raw and matched grant rate differences based on whether the applicant completed vocational training. "Education adjusted" uses other educational milestones as covariates in matching, while "education unadjusted" does not consider other educational milestones in propensity matching. "With" and "without completed high school/GED." Grey brackets represent insignificant results at $\alpha = 0.05$.

Education levels at or below high school have similar grant rates, both before and after adjusting. College-educated applicants without vocational training hold a higher raw grant rate, but most of this difference seems to be attributed to other case characteristics. Applicants with vocational training (with or without college) have higher grant rates both in our raw data and after matching. Further analysis shows applicants with vocational training hold higher raw grant rates regardless of other academic milestones, and those without a high school diploma or GED are the only group with insignificant matched association when compared to those without vocational training. Adjusting

for other educational factors changes the matched association by roughly 20%, showing that other educational factors hold some association with grant rates, even when the applicant has vocational training.

4.1.6 Risk assessment score

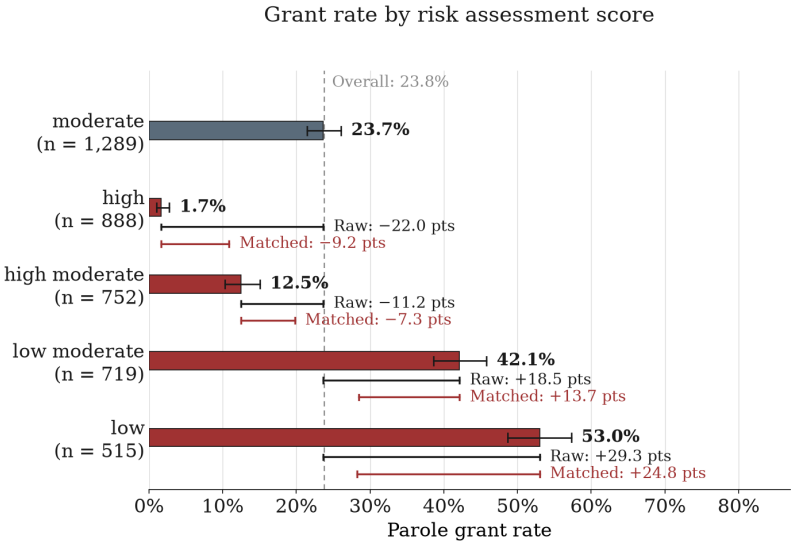


Figure 10: Raw and matched grant rate differences among different psychological risk score levels, with moderate scores used as a reference.

Risk assessment score shows the largest and most consistent associations in our data, with all matched estimates shrinking toward zero relative to the raw differences. A low score is associated with a 24.8-percentage-point higher grant rate, and a high score with a 9.2-percentage-point lower rate.

4.1.7 Past hearings

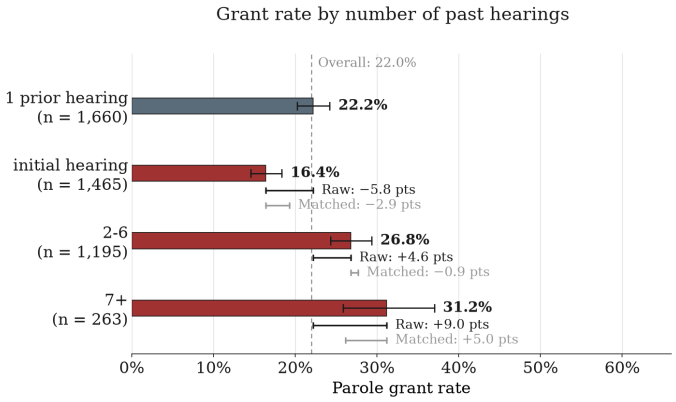


Figure 11: Raw and matched grant rate differences based on number of past parole suitability hearings, with 1 prior hearing used as the reference. Grey brackets represent insignificant results at $\alpha = 0.05$.

While there is a steady increase in raw grant rate by number of past hearings, the matched association is insignificant. It seems that most of this association can be explained by confounding case characteristics.

4.1.8 Conduct and sobriety while incarcerated

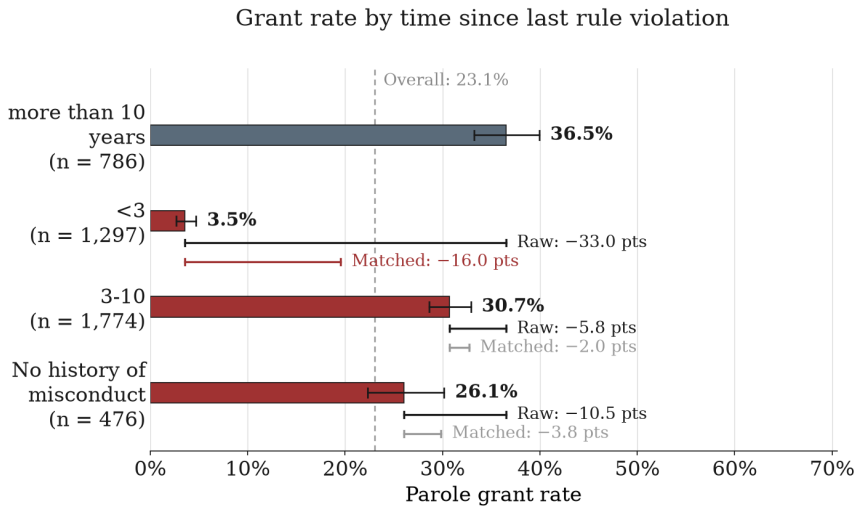


Figure 12: Raw and matched grant rate differences based on time since last recorded misconduct while incarcerated, with 10+ years used as the reference. Grey brackets represent insignificant results at $\alpha = 0.05$.

While applicants with no documented misconduct in the last 10 years have a higher grant rate, adjusting for other factors suggests that recent misconduct (less than three years) corresponds to a 16-percentage-point lower grant rate, while other records (3+ years, no history of misconduct) show no adjusted difference from those with 10+ years of clean records.

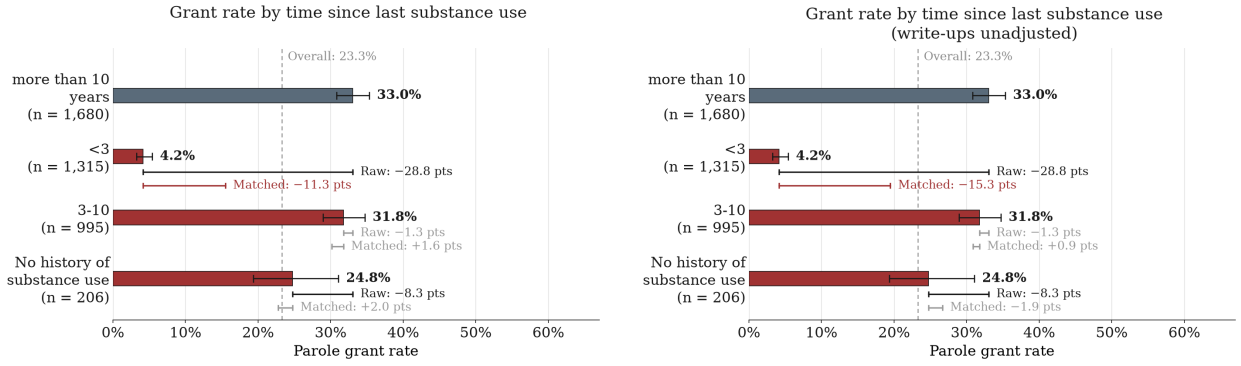


Figure 13: Raw and matched grant rate differences based on time since last substance use, with 10+ years used as the reference.
Left: Time since last misconduct is included in the covariates. Right: Time since last misconduct is excluded from the covariates. Grey brackets represent insignificant results at $\alpha = 0.05$.

Sobriety follows a similar pattern to disciplinary record. Use within the last three years is associated with an 11.3 percentage point lower grant rate than the reference, while longer intervals show no

adjusted difference. Leaving misconduct out of our covariates increases the estimated association of recent substance use, indicating that panels may weigh substance use beyond the write-up it may cause.

4.1.9 Gang affiliation

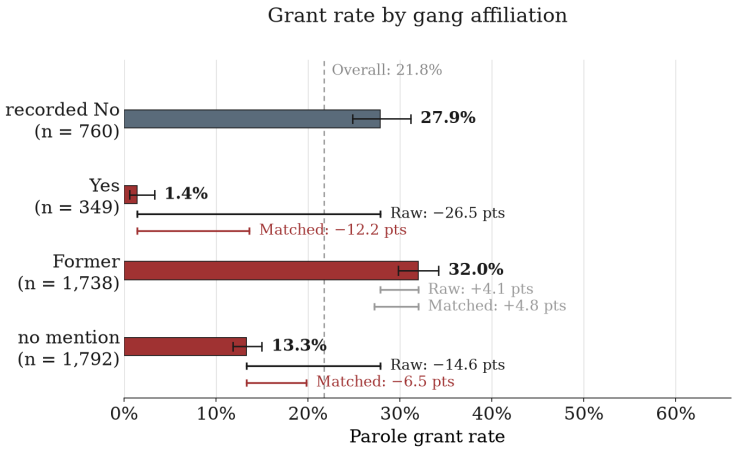


Figure 14: Raw and matched grant rate differences based on gang affiliation status at the time of hearing, with no current or former affiliation used as reference. Grey brackets represent insignificant results at $\alpha = 0.05$.

Current gang-affiliated applicants have a roughly 27 percentage point lower grant rate than applicants with no history of gang affiliation, with this association shrinking to roughly 12 percentage points after matching. Formerly gang affiliated applicants see an insignificant grant rate difference, both before and after matching.

Cases where gang affiliation is not mentioned see a lower grant rate, both in our raw data and after matching. We note that this relationship is likely not causal (see Section 2.2).

4.1.10 Victim involvement and support

We find no evidence that a victim's presence alone, or a statement alone, are associated with grant outcomes after adjusting. Cases where a victim is both present and gives a statement are associated with a raw grant rate roughly 3 percentage points lower than those with no victim involvement, and this association strengthens to 12 percentage points after matching, suggesting negative confounding; victims and their families may be more likely to attend and give statements at hearings where the applicant's profile is otherwise favorable (for example, lower risk scores, longer clean disciplinary records, or more prior hearings), so the raw comparison partly offsets the lower grant rate associated with victim participation, and matching on these characteristics removes this offset.

Documented support from a family member, community member, or sponsor is associated with a higher raw and matched grant rate than cases with no support. We found that internal support (such as laudatory chronos) shows no adjusted difference in grant rate, but having both internal and external support is associated with a 12.3 percentage point higher adjusted grant rate. Applicants in the "3+ types" category had internal and external support with some other type of support such

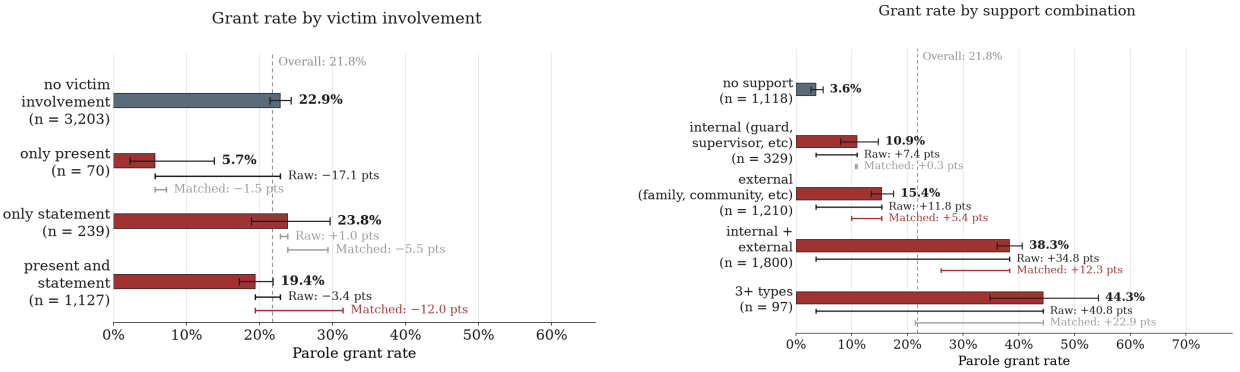


Figure 15: Left: Raw and matched grant rate differences based on victim involvement in the hearing, with no involvement used as the reference.
Right: Raw and matched grant rate differences based on documented applicant support at the time of hearing, with no support used as the reference. Grey brackets represent insignificant results at $\alpha = 0.05$.

as from the District Attorney or the victim. This appears to have the strongest raw association, but the low n in our data indicate these results should be read with caution.

4.1.11 Length of hearing

Short hearing lengths (0-59 minutes) are associated with lower grant rates, while longer hearings (120+ minutes) are associated with higher raw grant rates. When propensity matching with a 60-119 minute baseline, we found that no hearing lengths have a significant association with outcome. We note that hearing length is partly determined during the hearing since a panel leaning toward denial due to many aggravating case factors may spend less time.

4.2 Commissioner association

4.2.1 Commissioner-factor independence

Of fifteen pre-assignment characteristics, risk assessment, past hearings, and education depart from independence after correction at $\alpha = 0.05$. We note that the latter two factors may be institutional, as different commissioners generally sit at certain facilities.

We found that hearing lengths and mentions of lack of insight or minimization of the life crime vary by commissioner, so some commissioners may tend to hold longer hearings and some cite lack of insight more often. No factors tested have an exceptional Cramér's V.

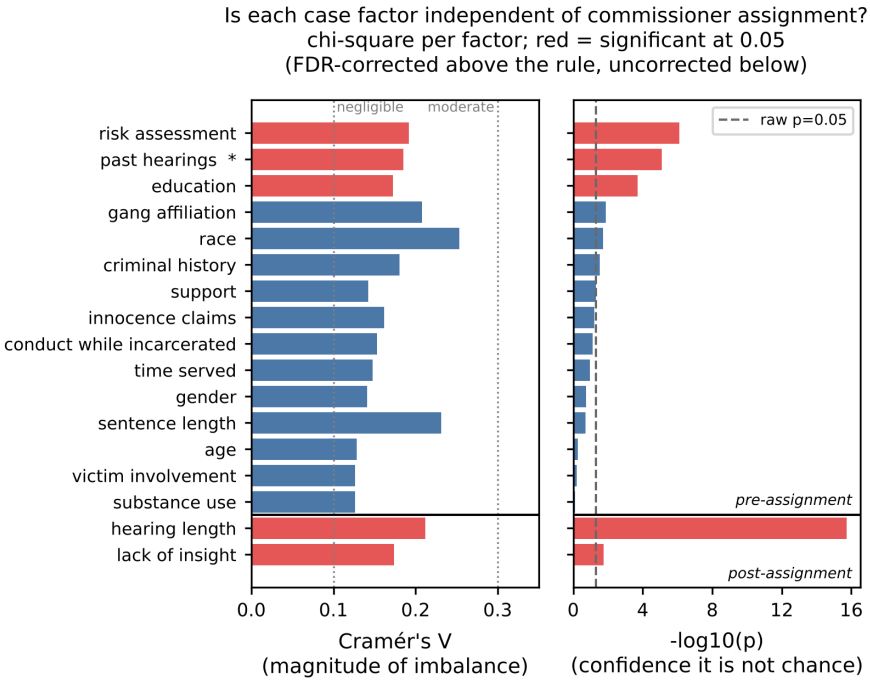


Figure 16: Cramér's V and $-\log_{10}(p)$ for all 17 factors tested. Factors with red bars remain significant after FDR correction.
*Past hearing count follows earlier denials, so an imbalance may be institutional.

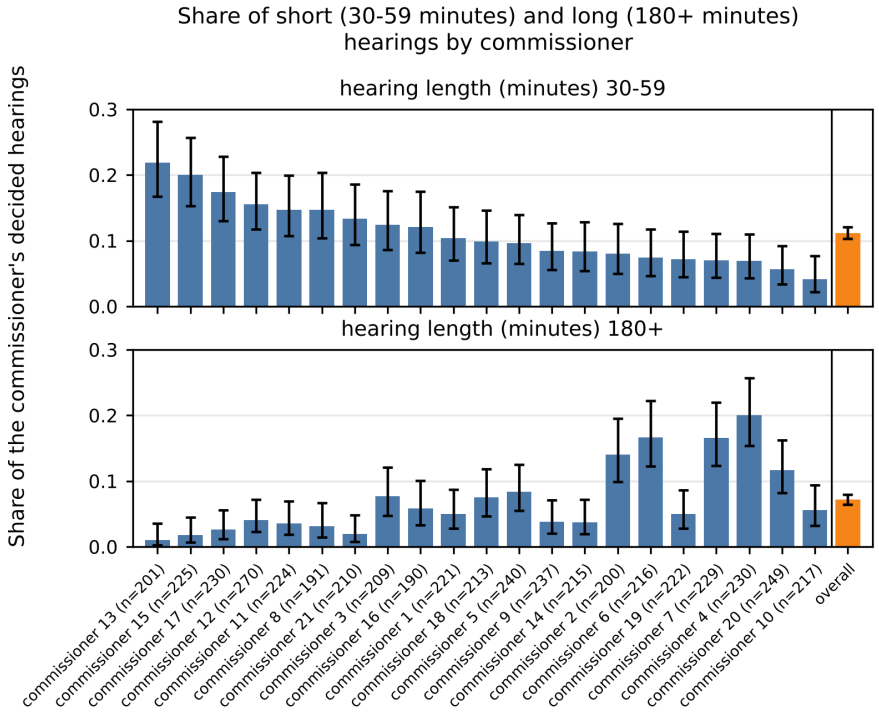


Figure 17: Share of long (180+ min) and short (30–59 min) hearings by commissioner.

In Figure 17, we see widely varying rates of long (180+ minutes) and short (30-59 minutes) hearings between commissioners, with ranges of 17.7 and 19.0 percentage points, respectively.

4.2.2 Innocence claims

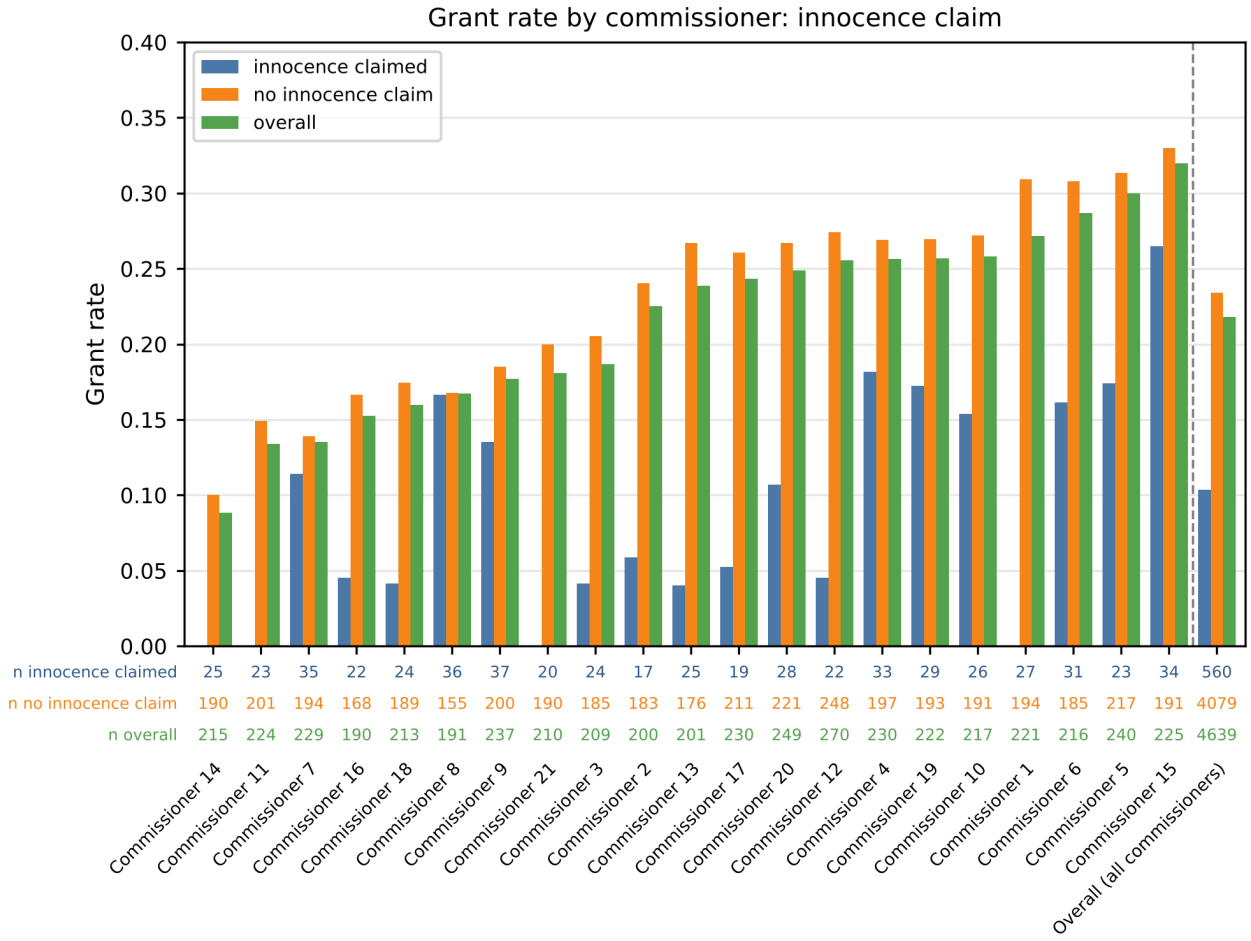


Figure 18: Grant rate by commissioner for innocence claimed cases, non-innocence claimed cases, and overall.

When unadjusted, grant rate gaps in innocence-claimed cases vary widely between commissioners. In Figure 18, Commissioner 1 has a raw grant rate gap of 31 percentage points with 0% of innocence-claimed cases granted, while Commissioner 8's gap is under 1 percentage point.

4.3 Predictive models

All eight models have similar binary-classification accuracy spanning a range of 0.004. The hybrid models have lower expected calibration error than random forest models, outperforming random forest on reliability in the low and high score ranges. Our basic random forest had an error rate of $\sim 1.57\%$ in its bottom 64 scored cases, while our basic hybrid model has an error rate of $\sim 0.3\%$ in its bottom 64 cases. The forest models outperform the hybrid models in AUC, meaning the forest models are more likely to score granted hearings above denied hearings. This difference is negligible at an AUC range of 0.011.

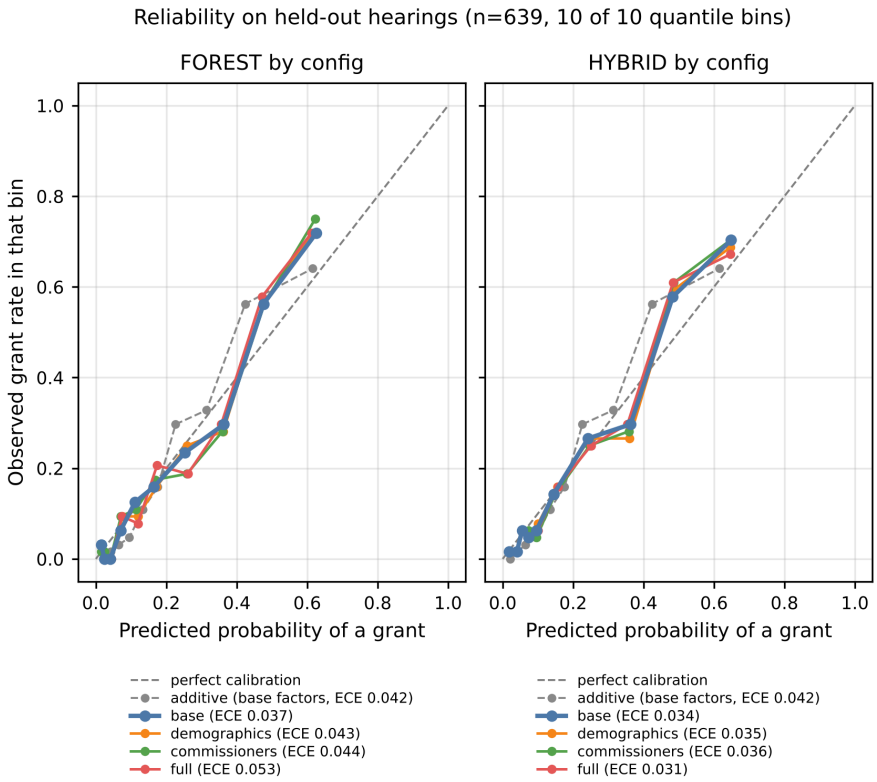


Figure 19: Reliability of random forest and hybrid models. Observed grant rates above the perfect calibration line signal under-scoring and observed grant rates below the line signal over-scoring.

5 Limitations and future work

Large language models have proven highly effective for parsing large volumes of legal text. Transcript corpora that would take a human weeks to read can have important data extracted in only a few hours. Even so, the model misses much of the nuance of a hearing. Applicant attitudes, body language, and attorney skill, among other things, go unrecorded in our analysis and likely affect parole outcomes. During data extraction, the model struggled with interpreting vague wording and needed strict guidance in its prompt. While these errors may seem insignificant in large-scale analysis, model data extraction may not paint the full picture of individual cases.

Our main limitation is in the reliability of inferred data, especially applicant race. The differential reasoning and confidence by the LLM when inferring race may raise concerns about racial biases in the LLM. Additional work can be done to improve our analysis and our model, including adding factors like the applicant's relationship to the victim, which could be an especially important consideration when analyzing victim support's association with grants.

Acknowledgments

This work was completed as part of the 2026 UCLA Computational and Applied Mathematics (CAM) REU: AI For Justice, with community partner The Innocence Center. We thank our project advisor, Deanna Needell, our mentor, Chi-Hao Wu, and our consultant, Mike Semanchik [15] for their guidance.

References

[1] *Information considered at a parole suitability hearing*, California Department of Corrections and Rehabilitation, (2025).

[2] P. C. Austin, *Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples*, Statistics in Medicine, 28 (2009), pp. 3083–3107.

[3] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1995), pp. 289–300.

[4] G. W. Brier, *Verification of forecasts expressed in terms of probability*, Monthly Weather Review, 78 (1950), pp. 1–3.

[5] *Cal. Code Regs. tit. 15, § 2240: Comprehensive Risk Assessments; Subdivision (b)*. California Code of Regulations, 2021. https://www.law.cornell.edu/regulations/california/15-CCR-2240.

[6] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[7] B. Efron, *Bootstrap methods: Another look at the jackknife*, The Annals of Statistics, 7 (1979), pp. 1–26.

[8] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (roc) curve*, Radiology, 143 (1982), pp. 29–36.

[9] C. E. Loeffler, J. Hyatt, and G. Ridgeway, *Measuring self-reported wrongful convictions among prisoners*, Journal of Quantitative Criminology, 35 (2019), pp. 259–286.

[10] M. L. McHugh, *The chi-square test of independence*, Biochemia medica, 23 (2013), pp. 143–149.

[11] A. Nellis and C. Barry, *A matter of life: The scope and impact of life and long term imprisonment in the united states*, The Sentencing Project, (2025).

[12] S.-L. T. Normand, M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil, *Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores*, Journal of Clinical Epidemiology, 54 (2001), pp. 387–398.

[13] J. C. Platt, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, in Advances in Large Margin Classifiers, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, eds., MIT Press, Cambridge, MA, 1999, pp. 61–74.

[14] P. R. Rosenbaum and D. B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika, 70 (1983), pp. 41–55.

[15] M. Semanchik, *When innocence costs you freedom: What a year of parole hearings reveals*, The Innocence Center, (2026).

[16] A. E. Valojerdi and L. Janani, *A brief guide to propensity score analysis*, Medical journal of the Islamic Republic of Iran, (2018).

[17] E. B. Wilson, *Probable inference, the law of succession, and statistical inference*, Journal of the American Statistical Association, 22 (1927), pp. 209–212.