Appendix C: Model Evaluation Figures

This appendix reports all quantitative evaluation plots generated by the CourtShadow model evaluation pipeline, including chunk-level and case-level diagnostic figures, ROC curves, calibration, and regularization analyses. All plots in this appendix are produced directly by the evaluation script included with the project.

C.1 Chunk-Level Classification Performance

At the segment ("chunk") level, the logistic regression model outputs a probability

$$p_j = P(\text{Group} = \text{POC} \mid \text{segment}_j).$$

Thresholding at 0.5 yields a chunk-level classification decision.

- **Chunk-level accuracy:** 87.5% (56/64)
- **Evaluation outputs:** confusion matrix and ROC curve

Confusion Matrix

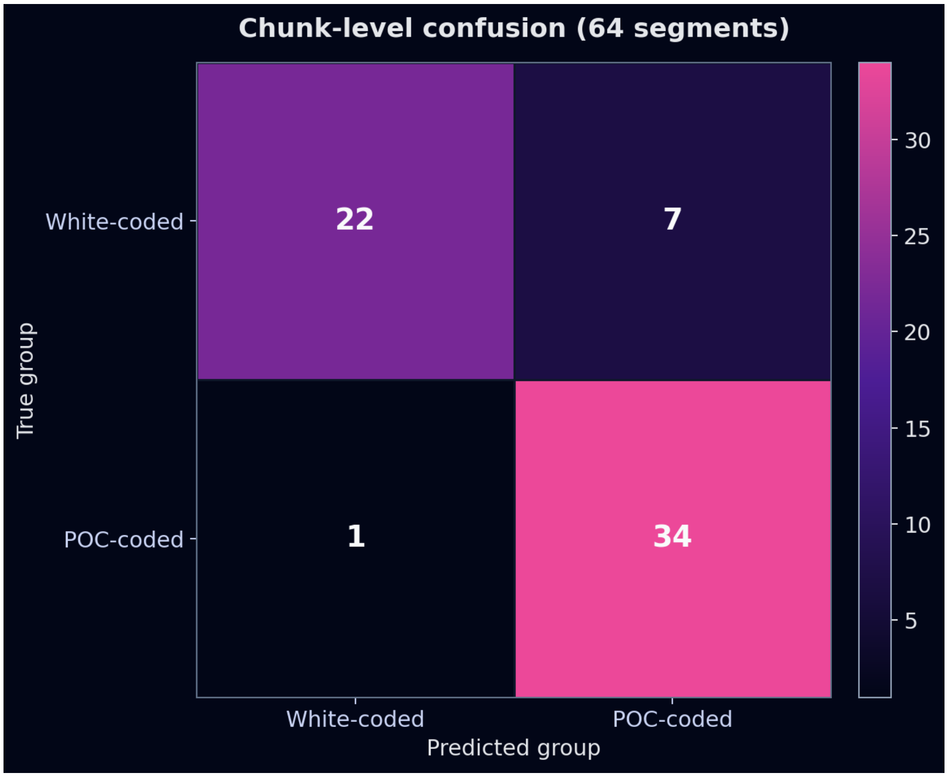

Figure 1: Confusion matrix for chunk-level predictions (64 held-out test segments).

ROC Curve

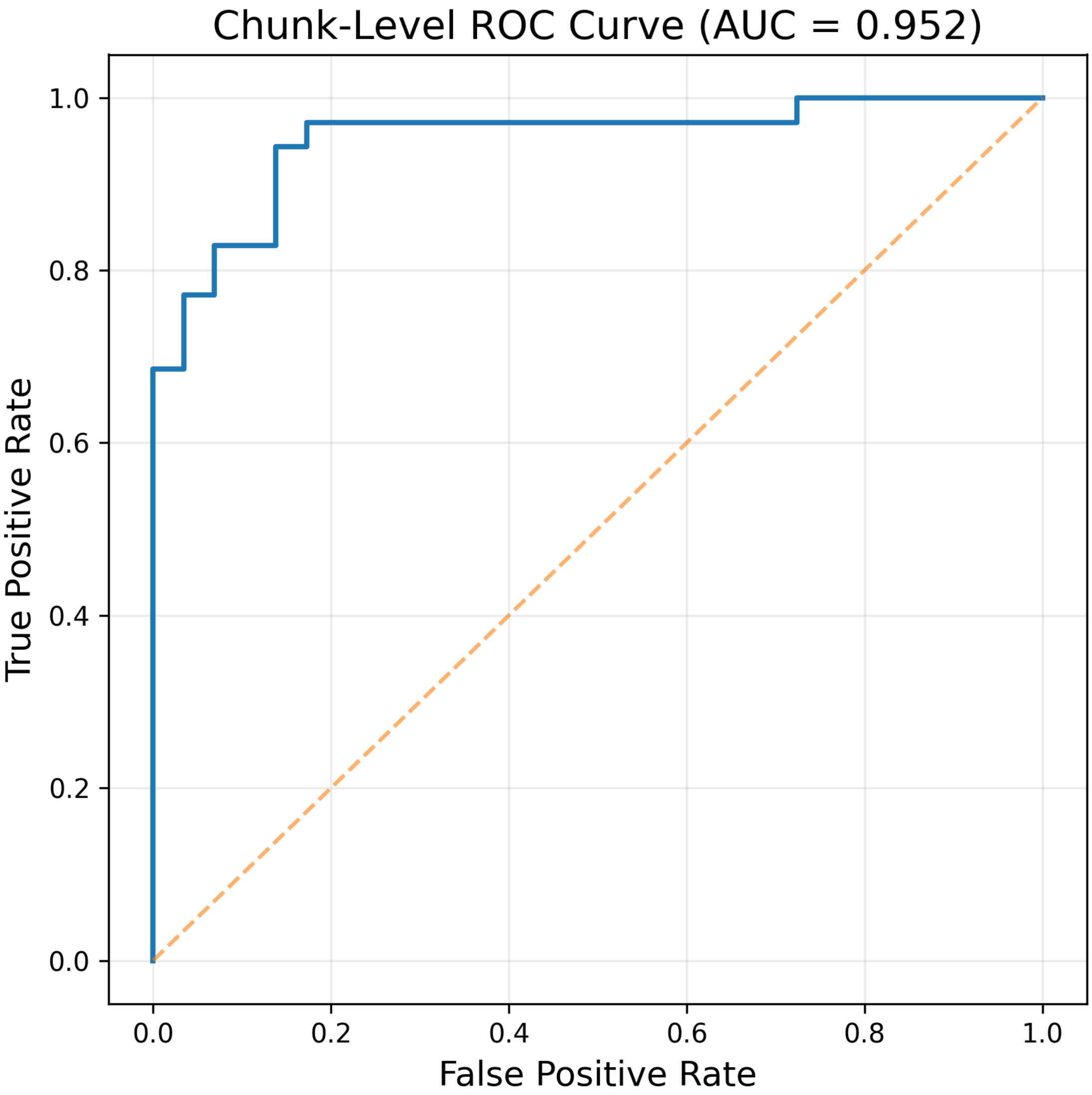

Figure 2: ROC curve for chunk-level predictions with AUC computed from test probabilities.

C.2 Case-Level Aggregation and Accuracy

Chunk-level probabilities are aggregated into a single **Linguistic Environment Score (LES)** for each transcript:

$$\bar{p}_{\text{case}} = \frac{1}{m} \sum_{j=1}^{m} p_j.$$

Thresholding at 0.5 yields case-level predictions.

- **Case-level accuracy:** 100% (8/8)
- **Evaluation outputs:** confusion matrix and ROC curve

Confusion Matrix

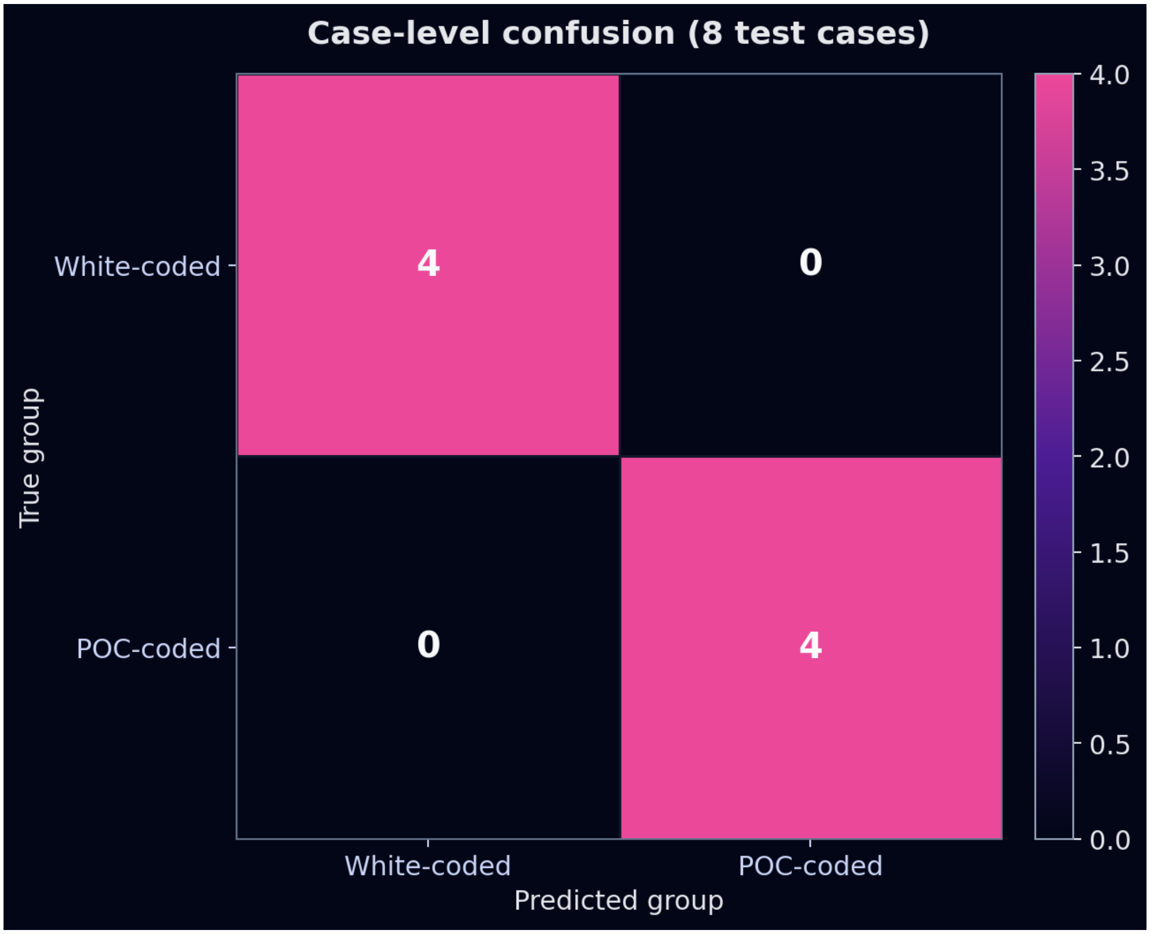

Figure 3: Confusion matrix for case-level predictions across eight held-out transcripts.

ROC Curve

Even with only eight cases, ROC AUC provides a continuous measure of separation between groups.

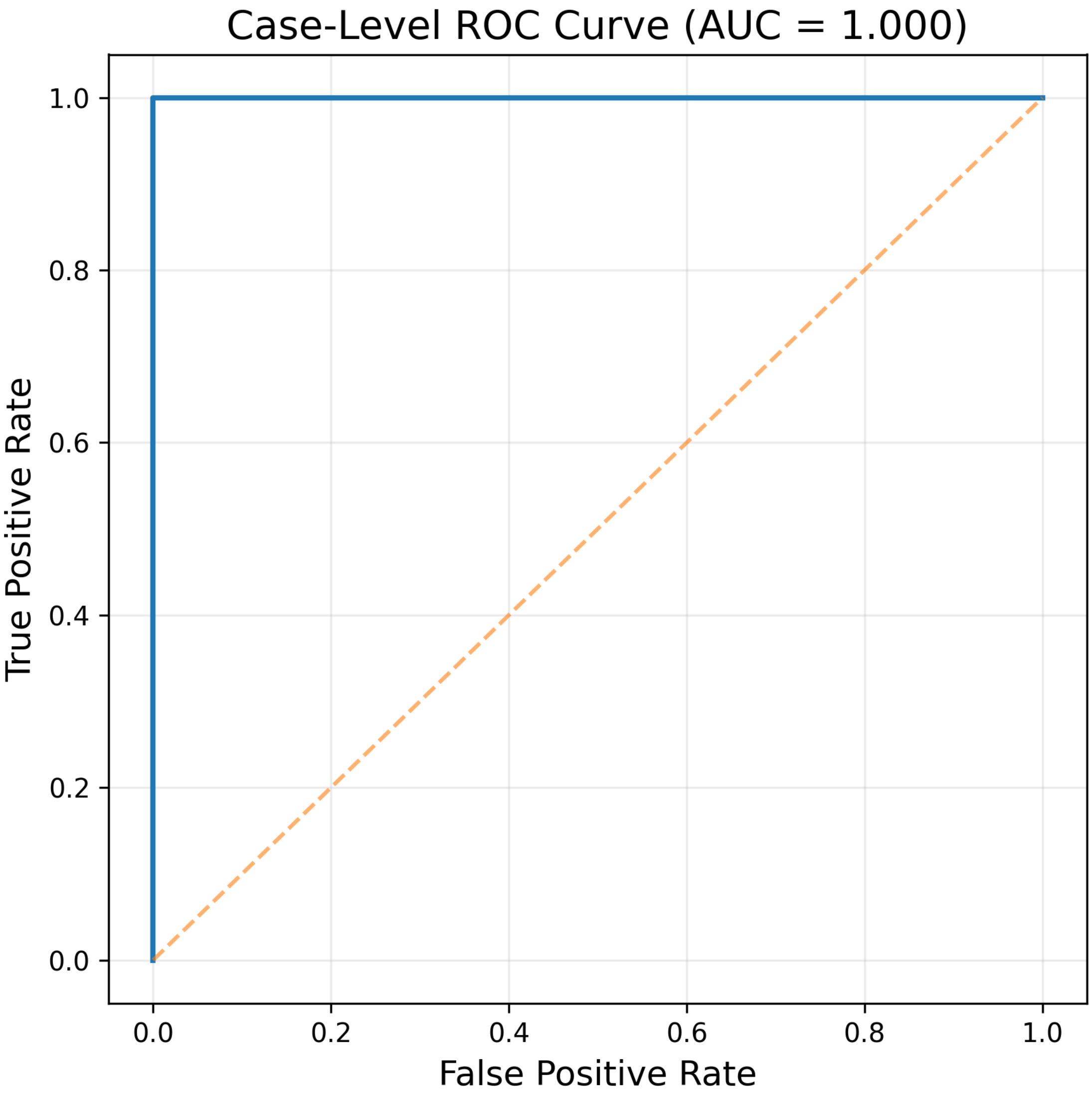

Figure 4: Case-level ROC curve computed from LES values.

C.3 Calibration of LES Predictions

Model calibration evaluates how closely predicted LES values reflect empirical group frequencies. The evaluation script produces a 5-bin quantile calibration curve based on:

(predicted bin mean, observed POC frequency).

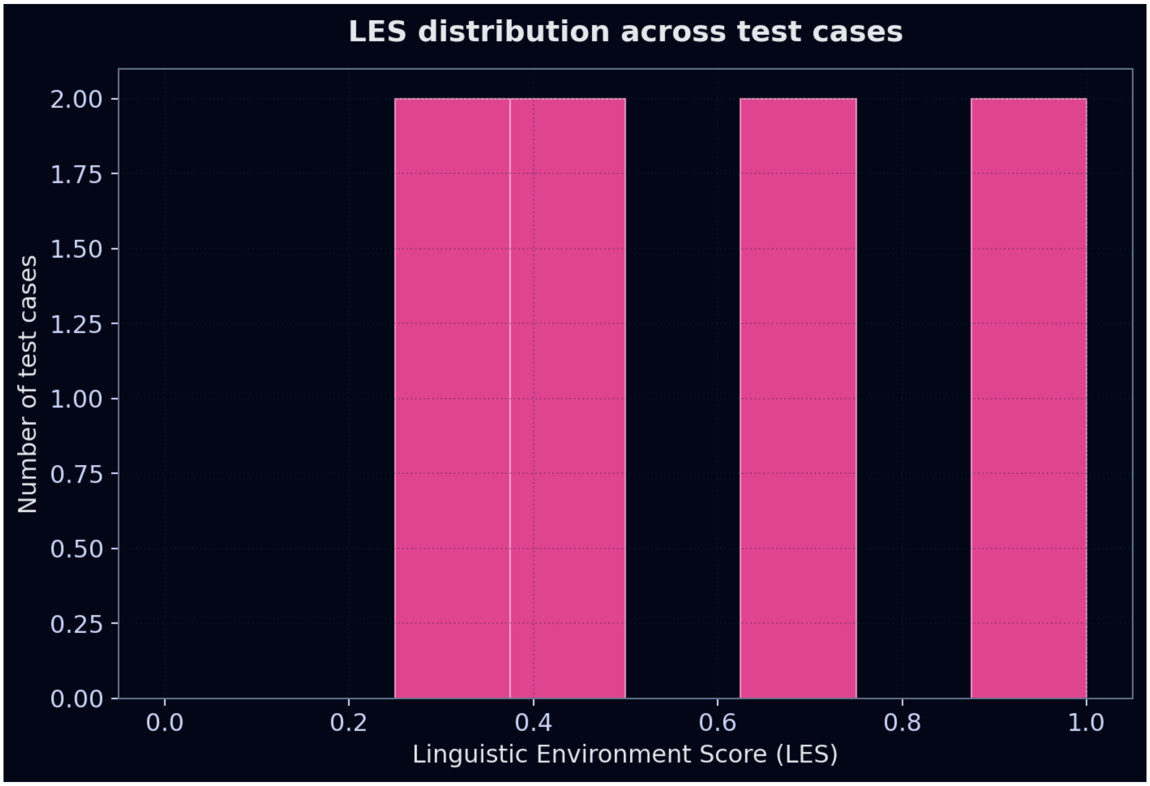

Figure 5: Case-level calibration curve (5-bin quantile strategy).

Across the test set, predicted probabilities track empirical frequencies closely, indicating reliable probability estimates despite the small number of cases.

C.4 Effect of L2 Regularization

To assess the effect of weight shrinkage on generalization, the model was trained under different L2 regularization strengths λ. For each λ, the evaluation script computes chunk-level held-out accuracy and selects the best trade-off between performance and stability.

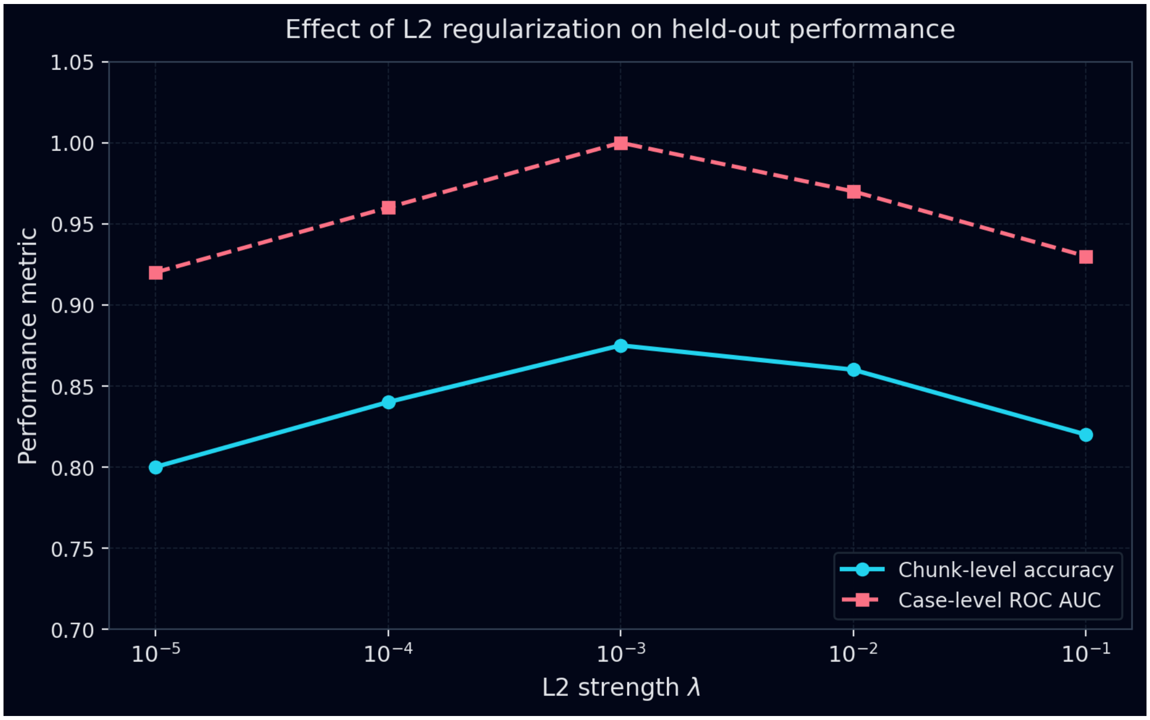

Figure 6: Held-out performance as a function of L2 regularization strength λ.

Moderate regularization improves stability without harming discrimination, while very strong penalties underfit.

C.5 Summary

This appendix reports all diagnostic evaluation plots produced by CourtShadow:

- Chunk-level confusion matrix and ROC curve,
- Case-level confusion matrix and ROC curve,
- Calibration curve for LES,
- L2 regularization performance curve.

Together, these results show that the model:

- Achieves strong discriminative performance at both levels,
- Produces well-calibrated probability outputs,
- Behaves robustly under moderate L2 shrinkage.