Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

A Visible-Residual Gate Finds Easier Reconstructions but Does Not Certify Their Hidden Regions

George Turner[1] · George Barnes[2] · Matthew Fisher[3]

[1] United States · [2] France · [3] United Kingdom

Original publication: November 5, 2026 | Article 685

Abstract

Can a restorer decide which outputs deserve review without seeing the missing content? A threshold on visible-region fitting residuals separated lower and higher hidden-block error in a held digit cohort. A calibration-only gate routed 1,034 of 1,297 records to the automatic group. Their mean hidden MSE was 0.08450, versus 0.12950 among the 263 review candidates. Nevertheless, 29.79% of accepted records exceeded an illustrative hidden-error cutoff of 0.1. The residual supplied useful ranking information, with no implication that accepted reconstructions were correct, readable, or safe to use without task-specific validation.

Illustrative routing analysis. Accepted and review are labels for a numerical policy; no human review, operational safety standard, or guaranteed error control is executed.

A Visible-Residual Gate Finds Easier Reconstructions but Does Not Certify Their Hidden Regions

> Illustrative routing analysis. Accepted and review are labels for a numerical policy; no human review, operational safety standard, or guaranteed error control is executed.

Abstract

Can a restorer decide which outputs deserve review without seeing the missing content? A threshold on visible-region fitting residuals separated lower and higher hidden-block error in a held digit cohort. A calibration-only gate routed 1,034 of 1,297 records to the automatic group. Their mean hidden MSE was 0.08450, versus 0.12950 among the 263 review candidates. Nevertheless, 29.79% of accepted records exceeded an illustrative hidden-error cutoff of 0.1. The residual supplied useful ranking information, with no implication that accepted reconstructions were correct, readable, or safe to use without task-specific validation.

A routing decision with an unobserved target

An image restorer sees the available context but cannot inspect the true content of the missing region at decision time. A workflow may still want to send some outputs for review while allowing others to continue automatically. That requires an observable signal related to hidden-region error. A low fitting residual over visible inputs is one candidate, but its meaning must be tested rather than assumed.

Our research question asks whether an observable reconstruction-residual threshold separates lower and higher hidden-block errors before hidden features are revealed. We use a training-derived latent restorer, a separate calibration subset for the score threshold, and a held evaluation subset for measuring hidden outcomes. The operational comparison is between the groups selected by the gate, not a randomized effect of human review.

Meng et al., 2025 develop a zero-shot inpainting model for oracle-bone inscription rubbings. We do not execute that model, claim its outputs need this gate, or assess historical readability. The connection is an accountability question adjacent to restoration: what observable evidence could justify routing a reconstruction, and how much uncertainty remains after a favorable signal?

Selective classification in Geifman and El-Yaniv, 2017 provides a framework in which predictions can be accepted or withheld. Our policy borrows only the explicit distinction between coverage and error. It is not that neural selective-classification method and inherits no risk bound from it. The word accepted means below a declared residual threshold, not independently certified correct.

Visible fit can be informative without being sufficient

The latent model estimates coefficients that explain observed features. If a record fits the learned representation well, its hidden features might also be easier to predict. That is a plausible association, not an identity. Two records can have similar visible coordinates and different hidden content; a model can fit its inputs while extrapolating the missing region incorrectly.

The representation perspectives in Pearson, 1901 and Tipping and Bishop, 1999 locate the fitted low-dimensional structure. The regularization rationale in Hoerl and Kennard, 1970 explains why a good visible fit is not obtained by freely interpolating every observation. Nevertheless, neither the basis nor its ridge penalty converts a residual into a calibrated probability of hidden correctness.

Confidence calibration in Guo et al., 2017 addresses relationships between confidence statements and observed correctness. Our score is not a classification probability, and no temperature scaling or confidence model is fitted. A quantile of visible residuals calibrates a distributional threshold for that score. It does not, by itself, calibrate the hidden-error rate among accepted records.

Conformal approaches discussed in Shafer and Vovk, 2008 and Lei et al., 2018 build prediction sets from specific nonconformity scores and assumptions. Our gate constructs no hidden-feature prediction interval. It does not use calibration hidden targets to choose an error radius. Thus, an eighty-percent score quantile should not be described as eighty-percent reconstruction coverage or any other conformal guarantee.

A frozen reconstruction and a separate score sample

The source data are Alpaydin and Kaynak, 1998, optical digits represented by sixty-four block-count features divided by sixteen. The official 3,823 training records determine the mean and sixteen principal directions. A fixed ridge coefficient of 0.01 governs latent fitting. These quantities are not adapted to calibration or held hidden outcomes.

The central two-by-two feature block is hidden, leaving sixty observed coordinates. The latent model fits those coordinates and predicts the four hidden values, clipped to their known zero-to-one range. It also reconstructs the visible coordinates through the same latent coefficients. The gate score is the average squared visible reconstruction residual after the declared clipping convention.

We permute the official 1,797 test records with seed 20261108, assigning five hundred to calibration and 1,297 to evaluation. The same split appears in another mask-calibration study in this project, which is disclosed as shared source evidence rather than an independent replication. Here only visible residual scores determine the routing thresholds.

The calibration subset's hidden values are not used to optimize the gate. This is a stricter information boundary than merely excluding held labels: the threshold cannot quietly be selected for a favorable hidden-error profile in calibration. Leakage concerns in Kapoor and Narayanan, 2023 and model-selection concerns in Varma and Simon, 2006 motivate making that distinction explicit.

Three score quantiles and one primary group comparison

We calculate the calibration score's 0.50, 0.80, and 0.95 quantiles using the higher empirical order-statistic convention. The primary rule accepts an evaluation record when its visible residual does not exceed the 0.80 threshold. The two other thresholds describe how coverage and observed error change along the same score ordering; they are not searched to choose the most favorable held result.

The primary cutoff is 0.012885 in mean squared visible-feature units. The lower and upper illustrative cutoffs are 0.008651 and 0.018808. These are data-derived score thresholds, not validated archival quality standards. Since calibration and evaluation contain different records, their realized accepted fractions need not equal the nominal calibration quantiles exactly.

We evaluate two hidden outcomes. The first is average squared error across the four hidden features. The second is whether that hidden MSE exceeds 0.1. The binary cutoff is an explicitly numerical illustration; no perceptual study establishes it as a readable-character boundary, a failure of historical meaning, or an acceptable field error rate.

The threshold perspective in Fawcett, 2006 highlights that different operating points can change coverage and error tradeoffs. Cost-sensitive analysis in Elkan, 2001 explains why choosing one requires

consequences beyond a curve. We have no review-cost model, and the eighty-percent rule is not claimed to maximize utility or minimize an application's expected loss.

An association between selected groups

The primary accepted and review groups are compared using all 1,297 evaluation records. Model 1 regresses hidden MSE on an accepted indicator and intercept. Model 2 uses the hidden-error-above-0.1 indicator as a linear-probability outcome with the same design. The accepted coefficient is the observed mean difference between the two groups, not a treatment effect.

Both models use HC3 covariance and t1295 intervals, following White, 1980. No controls or fixed effects enter the equations. The intercept gives the review-group mean; adding the accepted coefficient gives the automatic-group mean. In this two-group design those fitted means remain empirical outcome averages, rather than an extrapolated probability for a new covariate combination.

Bias-reduced likelihood in Firth, 1993 concerns a different binary-estimation problem. We do not apply such a correction or infer that changing the link function would validate the routing rule. The measurement issue is whether the observable score separates relevant outcomes, followed by whether those outcomes and costs support the intended decision.

Cluster-inference concerns in Cameron and Miller, 2015 and dependence-aware evaluation in Roberts et al., 2017 limit the interpretation of image-level intervals. Writer identifiers are unavailable, and the fitted basis, calibration split, and selected quantile are treated as fixed. The intervals therefore omit uncertainty from alternative training cohorts and threshold-calibration samples.

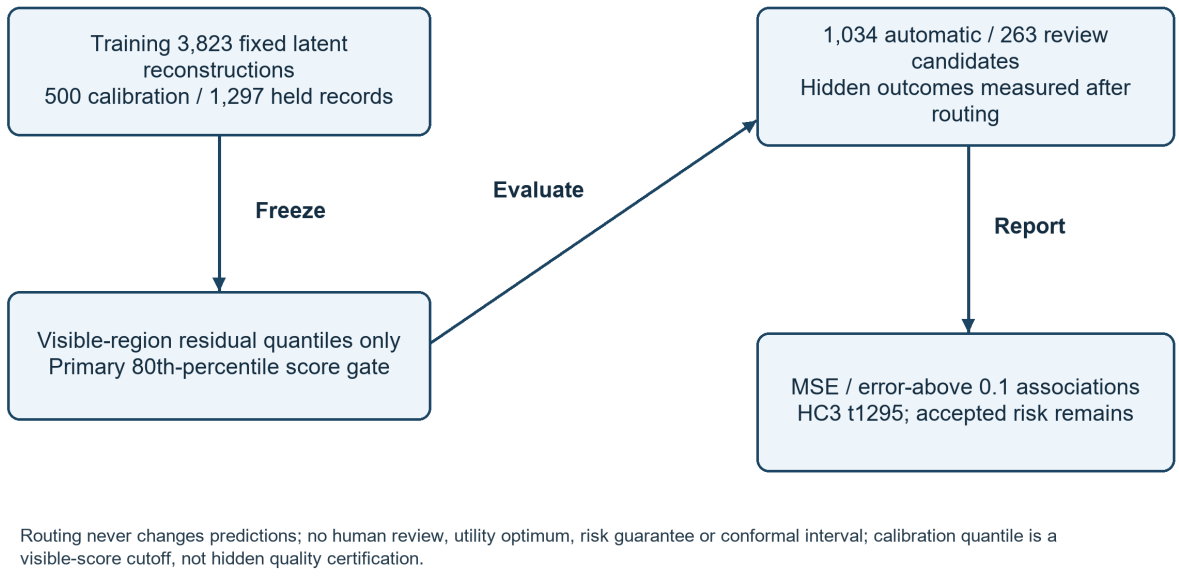

Figure 1. *Hidden features evaluate the gate after routing; they never choose the threshold. Automatic and review are illustrative route labels, not observed human interventions.*

Lower mean error with substantial remaining accepted risk

The primary threshold accepted 1,034 of 1,297 evaluation records, a fraction of 0.797224. The remaining 263 became review candidates. Mean hidden MSE was 0.084500 among accepted records and 0.129496 among review candidates. Model 1 estimates an accepted-minus-review difference of -0.044996, with interval -0.060676 to -0.029316.

The binary illustrative error fraction was 0.297872 in the accepted group and 0.448669 in the review group. Model 2 estimates a difference of -0.150797, with interval -0.217336 to -0.084257. The gate therefore separated easier and harder hidden reconstructions on both declared outcomes, while leaving a nontrivial numerical error fraction among records it accepted.

Table 1. *Held hidden-error associations with the primary automatic-route indicator. N equals 1,297 records; OLS or linear-probability regression with HC3 covariance and t1295 intervals. Intercepts represent the 263 review candidates. No controls or fixed effects; routing is not randomized.*

Model	Term	Estimate (SE)	95% CI	N
M1	Review-route mean	0.129496 (0.007450)	[0.114880, 0.144112]	1297
M1	Automatic minus review	-0.044996 (0.007993)	[-0.060676, -0.029316]	1297
M1: fit	R-squared = 0.0322	-	-	1297
M2	Review-route mean	0.448669 (0.030785)	[0.388274, 0.509064]	1297
M2	Automatic minus review	-0.150797 (0.033918)	[-0.217336, -0.084257]	1297
M2: fit	R-squared = 0.0167	-	-	1297

Model labels: M1 = Hidden-feature MSE association with acceptance; M2 = Hidden MSE above 0.1 association with acceptance.

For context, mean hidden MSE across the entire evaluation set was 0.093624, and the error-above-0.1 fraction was 0.328450. These unconditional summaries use a different denominator from the accepted group. A lower accepted mean does not imply that the hidden error of any record was reduced by routing; the gate selects records without altering their predictions.

At the 0.50 calibration quantile, the rule accepted 673 evaluation records, or 0.518890, with mean hidden MSE 0.075382. At the 0.95 quantile, it accepted 1,247, or 0.961449, with mean MSE 0.091804. These descriptive operating points show a coverage-error pattern within the held records. No additional interval or optimal-threshold claim is attached to them.

What routing has not accomplished

The accepted group still contained records with substantial hidden numerical error. The score orders some of the available difficulty but does not observe the actual hidden content. A workflow that needs a very low failure rate could not infer acceptability from the favorable group difference alone. It would need a task-specific error definition, tolerable rate, and evidence that the required operating point is achieved.

The scoring perspective in Gneiting and Raftery, 2007 keeps the objective explicit. Mean squared feature error and the chosen binary cutoff are two analyst-declared outcomes; neither is a validated semantic loss. Measurement principles in Stevens, 1946 and error analysis in Carroll et al., 2006 caution against relabeling these quantities as the percentage of historical information retained or the probability a restored character is correct.

Distribution shift discussed in Ovadia et al., 2019 and covariate-shift conformal analysis in Tibshirani et al., 2019 also matter if the gate is transported. The observed residual distribution and its relationship to hidden loss could change under a new mask, writer population, or imaging process. The present within-dataset split does not establish a transported conditional error guarantee.

Agreement perspectives in Bland and Altman, 1986 and Lin, 1989 reinforce that a favorable association does not make visible residuals interchangeable with hidden loss. The model can fit observed context unusually well while predicting the missing block poorly. The residual is an observable proxy whose usefulness remains empirical and bounded by the destination data and task.

Accountability requires a separate human-workflow study

An accepted route and a review route are convenient descriptions of a possible policy, but no reviewer participated in this experiment. We do not know whether human inspection would correct the difficult cases, how long it would take, or whether the visible context would let a person recognize the model's error. Those questions require additional observations rather than a numerical reinterpretation of the selected groups.

Image-completion work such as Suvorov et al., 2021 illustrates a substantially different reconstruction mechanism, which could produce a different visible-to-hidden error relationship. Our gate cannot be assumed to transfer to such a model, still less to the target zero-shot architecture. The completed experiment is limited to one fixed linear latent restorer on block-count features.

Neyman, 1934 examines representative sampling and purposive selection, underscoring the need to state how an inferential population relates to a selected group. The caution in Wasserstein and Lazar, 2016 also discourages using a small p-value as a complete decision criterion. Our group associations are distinguishable from zero, yet their significance does not certify accepted items. Influence reasoning in Cook, 1977 motivates retaining high-error accepted records in the analysis instead of deleting them as inconvenient exceptions.

The gate ranks risk without repairing it

At the primary threshold, the accepted group contained 308 records above the illustrative hidden-error cutoff, while the review group contained 118. The accepted fraction was lower, but its larger size meant that most cutoff-exceeding records still traveled through the automatic route. A workflow interested in finding every problematic reconstruction would need to examine this count as well as the favorable difference between group percentages.

The distinction between enrichment and capture is practical. The review group enriched high-error records relative to the automatic group, which could make inspection more efficient under some cost structure. It did not capture all high-error records, and the experiment does not show that every review candidate actually needs intervention. Both quantities matter when translating an association into a decision policy, and neither should be hidden by reporting only a mean error reduction in the accepted subset.

The calibration sample also estimates a score distribution rather than fixing a universal threshold. Another five hundred source records could yield a different empirical quantile, and a destination collection with a different visible-score distribution could accept a different fraction. We condition on the realized threshold in the displayed regression intervals. No repeated-calibration analysis was executed to estimate that additional source of policy variability.

Routing itself leaves the restored vector untouched. If the model is wrong on an accepted record, the gate does not correct it; if a review candidate is accurate, the gate does not make it worse. Changes in overall error after a human workflow would depend on what happens next. Since no such intervention occurred, the current results support prioritization evidence rather than an estimated improvement in final human-approved outputs.

This boundary also prevents a shortcut in interpreting the hidden cutoff. The numerical event above 0.1 is defined only for evaluation against known target values. A deployed gate cannot check it directly. Its observable residual is useful to the extent that the tested association remains relevant, with uncertainty about new inputs, tasks, and review procedures left visible.

Conclusion

A visible-residual threshold separated lower and higher hidden errors in a held reconstruction cohort, but substantial error remained among accepted records. Documentation in Gebru et al., 2021 and reuse principles in Wilkinson et al., 2016 support preserving score definitions, calibration thresholds, route membership, and hidden outcomes. The result supports evaluating an observable gate as a routing aid, with its coverage and remaining risk stated explicitly. It does not substitute for validated quality criteria or evidence about a human review process.

References

1. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
2. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
3. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
4. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
5. Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.
6. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.
7. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
8. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
9. Firth, David. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27-38. https://doi.org/10.1093/biomet/80.1.27.
10. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
11. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
12. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
13. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.
14. Neyman, Jerzy. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society,97(4),558-606. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.
15. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.

16. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

17. Cameron, A. Colin, and Douglas L. Miller. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

18. Shafer, Glenn; Vovk, Vladimir. (2008). A Tutorial on Conformal Prediction. Journal of Machine Learning Research,9(12),371-421. https://www.jmlr.org/papers/v9/shafer08a.html.

19. Lei, Jing; G'Sell, Max; Rinaldo, Alessandro; Tibshirani, Ryan J.; Wasserman, Larry. (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2017.1307116.

20. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804.

21. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

22. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

23. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

24. Elkan, Charles. "The Foundations of Cost-Sensitive Learning." *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

25. Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. https://doi.org/10.1080/14786440109462720.

26. Alpaydin, E.; Kaynak, C. (1998). Optical Recognition of Handwritten Digits [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.

27. Tibshirani, Ryan J.; Barber, Rina Foygel; Candès, Emmanuel J.; Ramdas, Aaditya. (2019). Conformal Prediction Under Covariate Shift. arXiv:1904.06019. https://arxiv.org/abs/1904.06019.

28. White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

29. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.

30. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.