Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

A Face-Identification Accuracy Needs an Unknown-Identity Denominator

Blake Knight[1] · Logan Reed[2] · Jason Reed[3]

[1] Switzerland · [2] Netherlands · [3] Finland

Original publication: November 7, 2026 | Article 647

Abstract

Rejecting uncertain queries can improve an accepted-only accuracy while reducing the number of known people identified successfully. In a small public face collection, a frozen nearest-gallery rule identified 78 of 100 known-identity test queries correctly. Adding a calibrated rejection threshold left 64 accepted queries, 60 correctly identified, while accepting 5 of 100 queries from unseen identities. Thus, 93.75% accuracy among accepted known queries coexisted with only 60% accepted-correct coverage of all known queries. Separate known and unknown denominators make this tradeoff visible.

Public-image protocol audit with a fixed pixel representation and a calibration-only rejection threshold. This is not an evaluation of generated video or a modern face-recognition system.

A Face-Identification Accuracy Needs an Unknown-Identity Denominator

> Public-image protocol audit with a fixed pixel representation and a calibration-only rejection threshold. This is not an evaluation of generated video or a modern face-recognition system.

Abstract

Rejecting uncertain queries can improve an accepted-only accuracy while reducing the number of known people identified successfully. In a small public face collection, a frozen nearest-gallery rule identified 78 of 100 known-identity test queries correctly. Adding a calibrated rejection threshold left 64 accepted queries, 60 correctly identified, while accepting 5 of 100 queries from unseen identities. Thus, 93.75% accuracy among accepted known queries coexisted with only 60% accepted-correct coverage of all known queries. Separate known and unknown denominators make this tradeoff visible.

1. Which queries enter the accuracy claim?

An identity system can answer two different questions: which enrolled person best matches an image, and whether the image belongs to anyone enrolled at all. A closed-roster identification evaluation answers the first question by assuming the second. When unfamiliar people are later admitted to the query stream, forcing every query to match a known identity creates a different error process. A known-only accuracy cannot reveal that process because unfamiliar identities never enter its denominator.

We ask what changes when a known-roster face-identification report includes both successful acceptance of known queries and erroneous acceptance of unknown identities. The experiment uses a fixed, simple nearest-gallery pixel representation. We calibrate one rejection threshold before evaluating held queries, then preserve three denominators: all known queries, accepted known queries, and all unknown queries. The purpose is reporting clarity rather than a new recognition algorithm.

Subject-preserving generation provides one application where an identity score can be interpreted too broadly. Meng et al., 2026 uses stacked multi-view identity mosaic injection for subject-preserving video generation. Our study does not run that method or score its generated subjects. It examines a more elementary issue around identity evaluation: a matcher that distinguishes enrolled identities still needs an explicit protocol for queries outside the roster.

The main proposition is that accepted-only identification accuracy and accepted-correct coverage need not move together. A rejection rule may discard some queries that the forced matcher would have identified correctly. A second proposition is that a calibration constraint on observed unknown acceptance does not certify the same rate for a new set of unknown identities. Both propositions concern the declared finite protocol and its interpretation.

2. Recognition representations do not remove protocol choices

Early face recognition includes eigenfaces Turk and Pentland, 1991, discriminant projections Belhumeur et al., 1997, and identity-model parameterization Samaria and Harter, 1994. These sources provide historical context for small controlled image collections. More recent representation learning through FaceNet Schroff et al., 2015, ArcFace Deng et al., 2018, and CosFace Wang et al., 2018 offers different methods for organizing identity-related feature spaces. None is fitted or evaluated in the present pixel-based experiment.

Quality-sensitive face representations such as AdaFace Kim et al., 2022 and MagFace Meng et al., 2021 further motivate attention to the relationship between an image and its recognition score. Photometric processing Tan and Triggs, 2010 addresses another source of variation. Our per-image centering and normalization do not reproduce those methods or establish invariance to physical illumination, image quality, or a generated subject's identity. They merely define a fixed representation for the reporting audit.

Dataset design also matters. LFW Huang et al., 2008, VGGFace2 Cao et al., 2017, and MegaFace Kemelmacher-Shlizerman et al., 2015 contain different identity populations and comparison settings from the small collection used here. FairFace Kärkkäinen and Joo, 2019 explicitly concerns demographic representation, which is not established by our source's forty identities. PhotoMaker Li et al., 2023a illustrates an identity-conditioned generation setting, but the relationship between reference images and generated identities would require a separate, application-specific evaluation.

Open-set recognition theory Scheirer et al., 2013 distinguishes recognition with unknown classes from a closed-world assumption. OpenMax Bendale and Boult, 2016 supplies a particular deep-network approach to that problem. We use neither its model nor its theoretical bounds. A distance threshold on centered pixels should not be described as inheriting formal open-space guarantees simply because both procedures can reject a query.

3. Four hundred images with declared roles

We use the modified Olivetti face matrix distributed by Sam Roweis AT&T Laboratories Cambridge, 2026. It contains 400 grayscale images, with ten source images for each of forty identities according to the maintainer's ordering. Every image is transformed to the same orientation, centered by its own pixel mean, and divided by its nonzero Euclidean norm. No learned embedding, identity-specific normalization, or test-selected image preprocessing is introduced.

A fixed random seed assigns twenty identities to the known roster, ten to calibration unknowns, and ten different identities to test unknowns. For each known identity, three randomly ordered images enter the gallery, two enter calibration, and five enter the test set. Thus, the gallery has sixty images, known calibration has forty, and known testing has one hundred. Each unknown group contributes all ten images from each of its ten identities, giving one hundred calibration and one hundred test unknowns.

All four roles are disjoint at the image level. Unknown calibration and unknown testing are also disjoint at the identity level. Known calibration and known testing deliberately share the enrolled identities, because the task is to identify later images of those same people. This is an enrolled-roster evaluation, not identity-disjoint training of a general recognition representation. All query groups share the same fixed gallery.

For every query, we compute squared Euclidean distance to each normalized gallery image and take the nearest. Its gallery label is the forced identification output, and its distance is the rejection statistic. A query is accepted when that distance is at or below a single global threshold. No unknown identity is added to the gallery, and test labels never change the nearest-gallery rule or threshold.

4. Choosing the threshold before testing

Candidate thresholds are the distinct calibration-query nearest distances, together with a value that rejects everything. We retain candidates whose empirical acceptance among the one hundred calibration unknown queries is at most five percent. Among these candidates, we maximize the number of correctly identified and accepted known calibration queries. If several thresholds achieve that objective, we choose the smallest. This tie rule is part of the protocol, not a decision made after observing test errors.

The selected squared-distance threshold is 0.409440. The forced matcher correctly identifies thirty of forty known calibration queries, while the accepted-correct count at the chosen threshold is twenty-two. It accepts three calibration unknown queries. The constraint need not bind at exactly five percent because the smallest threshold attaining the best known-query objective can lie below the constraint boundary. This explains the observed calibration rate without adjusting the procedure retrospectively.

Selective-classification work Geifman and El-Yaniv, 2017 clarifies the distinction between predictions made and predictions deferred. Cost-sensitive classification Elkan, 2001 emphasizes that the preferred operating point depends on the costs of different errors and rejections. We choose the stated calibration objective only as a transparent example. It is not claimed to be a generally optimal threshold for authentication, generated-content evaluation, or any deployment with different costs.

Calibration research Guo et al., 2017 and proper scoring rules Gneiting and Raftery, 2007 address related but different properties. The nearest distance here is not a calibrated probability of identity correctness. ROC analysis Fawcett, 2006 and precision-recall reasoning Saito and Rehmsmeier, 2015 further show that a threshold is one operating point in a larger set of possible tradeoffs. We do not choose another point because the first test result is inconvenient.

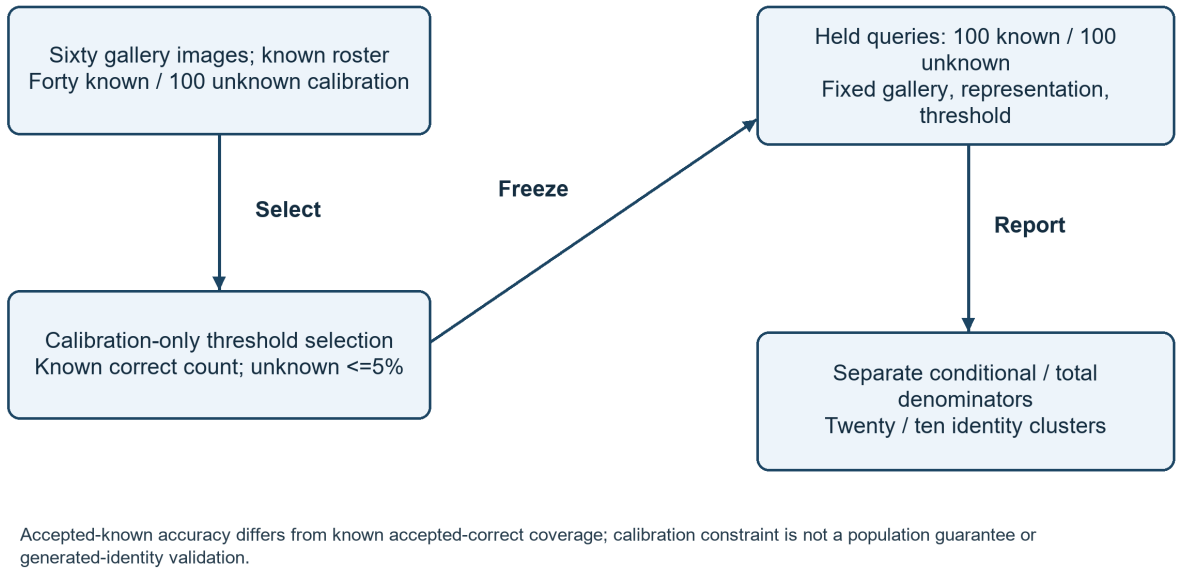

Figure 1. *Calibration known and unknown queries determine the threshold. Held known queries assess accepted-correct identification, while disjoint unknown identities assess false acceptance against the same gallery.*

5. Held results with the denominators retained

The forced closed-roster matcher identifies seventy-eight of one hundred known test queries correctly. After thresholding, sixty-four known queries are accepted: sixty correct and four incorrect. Thirty-six are rejected, including eighteen that the forced matcher would have identified correctly. The accepted-correct fraction of all known queries is therefore sixty percent, compared with seventy-eight percent correctness before rejection.

Among accepted known queries, accuracy is sixty divided by sixty-four, or 93.75 percent. That value is arithmetically correct but answers a conditional question. Comparing it directly with the original seventy-eight percent would conceal that fewer queries receive an answer. The paired loss of correctly identified known queries is eighteen percentage points. It is reported as a loss of successful identification, not as proof that deferral is necessarily worse under every decision objective.

For the unknown test identities, five of one hundred queries are accepted and assigned an incorrect enrolled label. Ninety-five are rejected. The observed unknown acceptance rate is therefore five percent. Its equality to the nominal calibration constraint is a realized result, not a finite-sample guarantee. Calibration used different unknown identities and had only three accepted unknown queries.

Table 1. Separate held known-query loss and unknown-query acceptance.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.1800 (0.0647)	[0.0445, 0.3155]	100
M1: fit	R-squared = 0.0000	-	-	100
M2	Mean outcome/contrast	0.0500 (0.0307)	[-0.0195, 0.1195]	100
M2: fit	R-squared = 0.0000	-	-	100

Model labels: M1 = Known-query correct-identification loss after rejection; M2 = Unknown-query acceptance rate under frozen threshold.

Model 1 uses the one hundred known queries and estimates the mean paired indicator of a formerly correct identification being rejected. Model 2 uses the one hundred unknown queries and estimates their acceptance rate. Both are intercept-only linear descriptions. Uncertainty is clustered by query identity: twenty clusters with t19 for Model 1, and ten clusters with t9 for Model 2. Their R-squared values are effectively zero.

The known-query loss coefficient is 0.1800, with standard error 0.0647 and 95% interval [0.0445, 0.3155]. The unknown-acceptance coefficient is 0.0500, with standard error 0.0307 and interval [-0.0195, 0.1195]. The latter approximate linear interval extends below zero. We retain and label that limitation rather than truncating the interval or presenting it as an exact binomial interval.

6. Why one combined accuracy can be misleading

Pooling known and unknown queries requires a mixture weight. In this constructed test set there are equal numbers of each, but a real query stream need not have that composition. Counting accepted-correct known queries and correctly rejected unknown queries as successes would produce a mixture-specific outcome. Changing the proportion of unknowns would change that outcome even if the two component rates remained exactly the same.

An accepted-only denominator is mixture-sensitive as well. Sixty-four known and five unknown queries receive an identity output in the present test set. A report limited to the sixty-four known accepted queries omits the five erroneous unknown acceptances. This is why a recognition report should specify whether its accepted-query accuracy includes unknown identities, rather than using acceptance and accuracy as if their denominators were self-evident.

Measurement reasoning Carroll et al., 2006 cautions against treating a thresholded pixel distance as the identity construct itself. The nearest-gallery procedure can respond to pose, expression, or image appearance that overlaps with identity. The study has no independent human identity-preservation score, no generated video, and no temporal continuity measurement. A well-specified acceptance rate does not by itself validate the representation for another identity-related task.

Structured validation guidance Roberts et al., 2017 explains why the identity-level split is stated explicitly. Image-level separation alone would not create new unknown identities if the same people appeared in calibration and testing. Conversely, sharing enrolled identities across known-query calibration

and testing is consistent with this particular task. The appropriate separation depends on which population the claim addresses, so neither split should be called universally leakage-free without reference to the intended use.

7. What the uncertainty leaves unresolved

Robust-covariance ideas White, 1980 and clustered-inference guidance Cameron and Miller, 2015 motivate grouping images from the same query identity. Nevertheless, all queries use one gallery, and the analysis conditions on its selected photographs. The displayed intervals do not include uncertainty from gallery resampling, representation choice, or selection of the forty people. With only ten unknown test identities, a population claim about rare false acceptances would be particularly weak.

The ASA statement Wasserstein and Lazar, 2016 also supports interpreting the unknown-acceptance interval without equating a probability threshold with practical safety. The point estimate of five percent is direct evidence about this held set. Its uncertainty and narrow source population remain relevant regardless of whether an approximate test crosses a conventional significance cutoff. A larger, independent unknown-identity cohort would be needed to evaluate a stringent operating requirement.

Dataset documentation Gebru et al., 2021 motivates preserving source checksums, image-role assignments, exact threshold rules, and all test counts. The research model and result table use no face photographs, and public availability is not treated as permission for unrestricted image redistribution. The resulting record supports inspection of the protocol while keeping claims about demographic representation, contemporary recognition accuracy, and generated-subject preservation outside the evidence actually collected.

The threshold's calibration sample is another source of uncertainty outside the displayed test intervals. Repeating the identity assignment could change both the unknown-distance distribution and the known-query objective, producing a different selected threshold. The current intervals condition on the single threshold already fitted. They should not be described as covering the variability of the complete calibration-and-testing procedure. Estimating that variability would require repeating the entire split and selection process, with a declared resampling population.

Finally, unfamiliar identity is only one form of novelty. The held unknown queries come from the same small image collection and share its capture conventions. They do not represent arbitrary images, nonface inputs, severe corruptions, or synthesized identities. The experiment therefore isolates a roster-membership reporting issue within a common source. Passing its unknown test does not establish general out-of-distribution rejection, and failing it would not identify the cause of a failure in a different application.

8. Conclusion

One frozen threshold changes several legitimate summaries in different ways. Accepted-known accuracy reaches 93.75%, while accepted-correct coverage of all known queries falls to 60%, and unknown acceptance is 5%. Reporting these rates separately reveals both the benefit of rejection and the identification it forgoes. A known-only face-matching score should not stand alone as evidence about unfamiliar identities or a broader identity-preservation application.

References

1. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.

2. Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.

3. Samaria, F.S.; Harter, A.C.. (1994). Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. https://doi.org/10.1109/acv.1994.341300.

4. 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.

5. Wang, Hao, et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.

6. Bendale, Abhijit; Boult, Terrance E.. (2016). Towards Open Set Deep Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.173.

7. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.

8. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

9. 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.

10. 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.

11. Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T. E.. (2013). Toward Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.2012.256.

12. 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.

13. 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.

14. Saito, Takaya and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

15. 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.

16. Kemelmacher-Shlizerman, Ira, et al. "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale" *arXiv preprint arXiv:1512.00596*, 2015. https://arxiv.org/abs/1512.00596.

17. Li, Zhen, et al. "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding." *arXiv preprint arXiv:2312.04461*, 2023. https://arxiv.org/abs/2312.04461.

18. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.

19. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

20. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.

21. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.

22. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

23. 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.

24. Xiaoyang Tan; Triggs, Bill. (2010). Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2010.2042645.

25. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.

26. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

27. AT&T Laboratories Cambridge. (2026). Modified Olivetti faces: Sam Roweis MATLAB version [Data archive; mirror accessed 7 November 2026]. File identity and row-label layout verified against the scikit-learn maintainer loader. https://ndownloader.figshare.com/files/5976027.

28. 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.

29. Kärkkäinen, Kimmo and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age" *arXiv preprint arXiv:1908.04913*, 2019. https://arxiv.org/abs/1908.04913.

30. Cao, Qiong, et al. "VGGFace2: A dataset for recognising faces across pose and age" *arXiv preprint arXiv:1710.08092*, 2017. https://arxiv.org/abs/1710.08092.