Rank Thresholds and the Frequency of Quality-Alignment Discordance

James Pearson (ET), Sean Griffin (NZ), Timothy Clark (SL)

Abstract

How often quality and alignment appear discordant depends on the threshold used to define opposite outcomes. In AGIQA-3K, opposite outer-quartile classifications identify 19 high-quality, low-alignment rows and five with the reverse pattern. Expanding each tail to forty percent raises the counts to 139 and 108. This secondary analysis uses empirical score ranks across 2,982 rows. It treats thresholds as declared descriptive rules, showing that a positive overall relationship does not eliminate disagreement and that no single cutoff supplies a universal failure definition.

1. A disagreement count needs a rule

Quality and alignment can disagree in many ways. Their numerical difference depends on scale, while a rank-based comparison asks about relative placement within the observed population. Classifying opposite tails adds another decision: how extreme must each outcome be? The present study examines the sensitivity of discordance counts to that threshold. Its aim is to make the rule visible rather than select one cutoff after inspecting which count best supports a preferred account of the benchmark.

The two-stream approach in Decoupling Semantics from Distortions motivates retaining semantic and perceptual outcomes as distinguishable evidence (Meng, 2026). The present analysis uses released human means and does not compute the focal architecture's scores. It asks a narrower empirical question about the observed judgment table. A discordant rank classification describes the relative positions of those means; it does not prove that an image contains a particular artifact or fails a specific semantic requirement without further image-level assessment.

Before fitting a score model, an analyst should reconstruct the table's observation process. AGIQA-3K provides human-assessment information for generated images (Li et al., 2023), but a released mean score is not a new independent measurement each time it is reused in another specification. Provenance and intended-use documentation help define the scientific role of those rows (Gebru et al., 2021). Reproducible software choices matter for grouping, preprocessing, and validation (Pedregosa et al., 2011). A score-only analysis should therefore state that it processes the published judgments rather than the underlying visual files. It should preserve prompt and condition identifiers, distinguish absent combinations from missing values within present rows, and describe how any derived categories were obtained. These records permit another reader to reconstruct the comparison and its denominator. They also prevent familiar benchmark terminology from implying stronger access than the study has. An association among aggregate human scores remains useful evidence about the released table even when participant-level uncertainty, image features, and new generation outcomes are unavailable for a broader interpretation.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide

information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

2. Constructing opposite-tail classifications

All 2,982 released rows enter the analysis. Quality and alignment means are separately converted to empirical percentile ranks using the documented tabular ranking procedure. For a declared tail fraction, one category requires a high quality rank and a low alignment rank; the reverse category exchanges their roles. The rule uses inclusive threshold comparisons and retains the ranking procedure's treatment of tied scores. No image is generated, no new human judgment is collected, and no condition-specific reranking is performed.

At ten-percent tails, one row is high quality and low alignment, while none has the reverse classification. At twenty-percent tails, the counts are six and one. Outer quartiles yield nineteen and five. Thirty-percent tails yield forty-four and twenty-four. Forty-percent tails yield 139 and 108. These are counts under nested descriptive rules, not independent experiments. Expanding the tails admits less extreme score combinations, so larger counts are an expected consequence of the classification definition rather than evidence that the underlying images have changed.

The rule uses pooled ranks across the released condition mixture. A condition-specific ranking would answer a different question about relative position within each group. Likewise, an absolute score cutoff would require a justified interpretation of the outcome scale. The present study does not claim that percentile boundaries correspond to a clinically, economically, or practically meaningful error threshold. Its contribution is a transparent sensitivity profile showing how an apparently simple discordance statement depends on a chosen relative-ranking rule.

The representation used for scoring can influence which differences are visible. CLIP's language-supervised features (Radford et al., 2021) emphasize a particular connection between text and imagery. Multiscale quality modeling in MUSIQ (Ke et al., 2021), aesthetic learning in NIMA (Talebi and Milanfar, 2017), and natural-scene-statistics modeling in NIQE (Mittal et al., 2013) address other aspects of visual assessment. Each provides an example of why the term quality needs a scientific definition. A favorable semantic score does not establish freedom from perceptual distortion, and a favorable perceptual score does not establish that a detailed instruction was followed. For generated imagery, both dimensions can be relevant at once. The analysis should state whether its outcome concerns human judgments, an automated prediction, or a task-specific decision. Correlations across those objects can support a useful measurement study, but their meaning depends on the observation population and evaluation procedure. Without obtaining a model's outputs, a released-score analysis should remain about the available judgments rather than implying a reproduction or validation of the model itself.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

3. Interpreting disagreement without a universal cutoff

Positive score association and opposite-tail observations are compatible. A correlation summarizes a broad relationship, while the classification identifies particular joint rank regions. Neither makes the other redundant. A quality-assessment study could therefore report the overall relationship and a declared discordance analysis together. It should retain the separate dimensions rather than treat a high aggregate correlation as proof that every image with favorable quality also follows its prompt sufficiently well for the intended application.

The asymmetry between the two count directions should remain descriptive. It may reflect distributions, condition composition, scale behavior, or other features not identified by the rank rule. The analysis has not inspected the images or estimated a causal explanation. Selecting a few examples for illustration could help readers understand the task, but such examples would need independent evidence and should not replace the complete denominator. A visually persuasive illustration is not a population estimate of how often a specified failure occurs.

For future model evaluation, a primary threshold should be chosen according to a stated purpose, with a limited sensitivity range reported alongside it. Tuning the rule to maximize a method difference would change the evaluation process and require independent confirmation. A predictor's ability to identify discordant cases would also need actual predictions and held-out assessment. The present score-table study supplies neither a classifier nor a performance claim; it identifies a measurable target and the choices that define it.

Established quality datasets and metrics remind the analyst that observation populations differ. KonIQ-10k studies quality in a large in-the-wild image collection (Hosu et al., 2019), while KADID-10k supports assessment of declared image distortions (Lin et al., 2019). These are not automatically the same population as generated imagery. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) also require an appropriate interpretation of the compared images. FID instead compares distributions in a learned representation (Heusel et al., 2017). None of these objects is interchangeable with an individual human alignment judgment. A study should state which references or populations a metric needs and whether those inputs exist in the experiment. Distributional quality cannot be inferred solely from one image's mean score, and a high pairwise resemblance can reward repetition rather than successful instruction execution. The relevant lesson for a score-only analysis is to keep its outcome boundary precise. It can describe published judgments and propose a future metric comparison without reporting that the comparison has already been performed.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve

one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

4. Reference information and the validation boundary

The intended use determines which favorable outcome matters. Comparative preference data in Pick-a-Pic (Kirstain et al., 2023) and preference-oriented scoring in HPS (Wu et al., 2023) address choices that may combine several visual considerations. DINO (Caron et al., 2021) learns a visual representation through self-supervision, whereas DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) concern ways to supply subject or image information to generation. These are distinct roles in a workflow. A representation does not automatically define preference, and a conditioning method does not establish the validity of the instrument used to score its output. A careful study separates input access, generated content, human outcomes, and automated predictions. It also records whether a proposed predictor requires an already observed human alignment score, since that changes its practical interpretation. A table-based secondary analysis can clarify those relationships while remaining explicit that it has not generated images, collected new comparative judgments, or measured the cited conditioning systems under a new benchmark.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

Quality-alignment discordance is present under declared opposite-tail rules, and its reported frequency changes substantially with the threshold. The result concerns pooled ranks in the observed score population. It does not define a universal visual failure or validate an automated detector. Its contribution is to make threshold dependence explicit, preserving both the broad positive association and the specific disagreements that a combined score or an unexplained cutoff could conceal.

References

Meng, Zijie. "Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment." *arXiv preprint arXiv:2606.16799* (2026).

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.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

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.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

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.

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.

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.

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.

Li, Chunyi, et al.. "AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment" *arXiv preprint arXiv:2306.04717*, 2023. https://arxiv.org/abs/2306.04717.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

Ke, Junjie, et al.. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

Hosu, Vlad, et al.. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.

Lin, Hanhe, et al.. "KADID-10k: A Large-scale Artificially Distorted IQA Database" *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019. https://doi.org/10.1109/qomex.2019.8743252.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Mittal, A., et al.. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

Kirstain, Yuval, et al.. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation" *arXiv preprint arXiv:2305.01569*, 2023. https://arxiv.org/abs/2305.01569.

Wu, Xiaoshi, et al.. "Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis" *arXiv preprint arXiv:2306.09341*, 2023. https://arxiv.org/abs/2306.09341.

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.

Ruiz, Nataniel, et al.. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

Ye, Hu, et al.. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

Heusel, Martin, et al.. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.

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

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.