Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

A Familiar Prompt Is Additional Information: Rating-Prediction Protocols in Generated Images

Mason Hall[1] · Simon Cooper[2] · Jason Reed[3]

[1] Latvia · [2] Lithuania · [3] Luxembourg

Original publication: November 5, 2026 | Article 593

Abstract

Knowing how other images with the same prompt were rated helped predict alignment more than quality. Across 2,982 AGIQA-3K records, an explicit peer-rating correction reduced alignment MSE from 0.38363 under whole-prompt withholding to 0.32744. Quality MSE instead changed from 0.42503 to 0.43670, with an interval that included no difference. The same images were held out under every protocol. These outcomes describe different information conditions, showing why performance on familiar prompts should be distinguished from evidence about prompts with no prior rated examples.

Cross-fitted metadata baselines on released aggregate ratings. Same-prompt peer scores are deliberately available in one protocol; no neural quality model or new human-rating experiment is run.

A Familiar Prompt Is Additional Information: Rating-Prediction Protocols in Generated Images

> Cross-fitted metadata baselines on released aggregate ratings. Same-prompt peer scores are deliberately available in one protocol; no neural quality model or new human-rating experiment is run.

Abstract

Knowing how other images with the same prompt were rated helped predict alignment more than quality. Across 2,982 AGIQA-3K records, an explicit peer-rating correction reduced alignment MSE from 0.38363 under whole-prompt withholding to 0.32744. Quality MSE instead changed from 0.42503 to 0.43670, with an interval that included no difference. The same images were held out under every protocol. These outcomes describe different information conditions, showing why performance on familiar prompts should be distinguished from evidence about prompts with no prior rated examples.

1. A held image can still have a familiar prompt

Generated-image datasets often contain several images associated with the same textual prompt. Withholding one image can therefore leave other rated images from that prompt in the fitting data. This is a valid setting when the intended task concerns additional images for familiar prompts. It is a different setting from predicting ratings for a prompt whose entire group has been withheld. The distinction concerns the information available to the predictor, not whether two source images are pixel duplicates.

We ask how a metadata-only prediction comparison changes when same-prompt rating information is explicitly supplied. The analysis uses the same held image records under three protocols. One estimates only generator means after withholding the entire prompt group. Another withholds only the individual image. A third adds the mean residual rating of the held image's remaining exact-prompt peers to that image-only reference. The last procedure deliberately measures the usefulness of familiar-prompt information.

Meng, 2026 separates semantics from distortions in vision-language quality alignment. Its substantive distinction motivates analyzing quality and text-alignment ratings separately here. We do not run its two-stream architecture or interpret a metadata baseline as a substitute for its evaluation. Our question is whether the information in an evaluation protocol can help one rating construct more than another before image content is modeled at all.

The analysis is exploratory. A same-prompt correction could help if prompts share systematic residual difficulty after generator means are removed. It could hurt if the remaining peer ratings mostly add noise or if prompt effects differ strongly across generators. We therefore do not assume a common improvement in both outcomes. Both are reported in their original score units, without combining them into a favorable composite.

2. Several visual scores can share a prompt without sharing a construct

AGIQA-3K Li et al., 2023 supplies paired aggregate quality and alignment scores for the records analyzed. AIGIQA-20K Li et al., 2024 represents another generated-image assessment resource, but its rows are not pooled here. Dataset size alone does not determine whether a prompt has appeared in fitting data, whether the images share a generator, or which rating construct a model is expected to predict.

Compositional evaluations such as GenEval Ghosh et al., 2023 and T2I-CompBench Huang et al., 2023a ask more structured questions about text-to-image generation. TIFA Hu et al., 2023 provides another way

to assess text-image faithfulness. These methods motivate treating prompt-related information as substantive, rather than as an arbitrary grouping variable. Our peer residual is not an implementation of any of these evaluators and contains no question-answering or object-detection evidence.

Human-preference approaches including ImageReward Xu et al., 2023, HPS Wu et al., 2023, and Pick-a-Pic Kirstain et al., 2023 target further evaluative relationships. CLIP Radford et al., 2021 provides a learned cross-modal representation relevant to text-image comparisons. The present study uses no embedding or preference model. A predictable alignment MOS is not automatically a predictable human preference, and an average peer score does not establish semantic correctness for the held image.

Perceptual measures such as LPIPS Zhang et al., 2018 and DISTS Ding et al., 2020 concern visual relationships distinct from the protocol variable examined here. Learned aesthetic assessment through NIMA Talebi and Milanfar, 2017, multi-scale quality assessment through MUSIQ Ke et al., 2021, and attention-based assessment through MANIQA Yang et al., 2022 provide other modeling contexts. CLIP-based image-quality assessment Wang et al., 2022 further connects representation and evaluation. These sources explain the breadth of possible predictors, while our deliberately small baseline isolates information availability from architectural complexity.

3. One captured table and three fitting boundaries

The source table contains 2,982 unique image names, 300 exact prompt strings, and complete aggregate quality and alignment MOS fields. Generator identity is taken from the documented filename prefix. The six resulting groups contain 1,200 Stable Diffusion 1.5 records, 592 Midjourney records, 300 AttnGAN records, 300 GLIDE records, 300 xl2.2 records, and 290 DALLE2 records. Exact-prompt group sizes range from seven to ten.

Prompt membership uses the source string exactly, retaining case and punctuation. We do not merge paraphrases, remove adjectives, or infer latent semantic classes. Consequently, prompt withholding is a clearly reproducible grouping rule rather than proof of semantic novelty. Two different strings can express similar content, and the captured metadata do not establish the absence of such relationships across groups.

For each held image, the whole-prompt-withheld predictor removes every row carrying its exact prompt. It then predicts the held rating using the mean of the remaining rows from the same generator. This is the novel-prompt reference in the local experiment. It has no access to ratings from the held prompt group, although it uses other prompts from the same generator. It is not a predictor for an unseen generator.

The image-only predictor instead removes the held row alone and estimates the same generator's mean. All other rows remain available. This protocol measures the small difference caused by retaining the other prompt-group records in the generator mean. It does not automatically exploit exact prompt identity in the prediction. Keeping it separate prevents us from attributing every difference between fitting boundaries to an explicit prompt-memory mechanism.

The peer-corrected predictor begins with those image-only fitting generator means. For every other image sharing the held exact prompt, it subtracts that peer's fitting-generator mean from the peer's rating. The mean of these peer residuals is added to the held image's fitting-generator mean. The held rating enters neither the reference generator means nor the residual correction. Peers can come from different generators, as allowed by the captured group.

4. A disclosed information advantage is not an accusation of leakage

Leakage analysis Kapoor and Narayanan, 2023 motivates naming which information would be available for an intended use. The peer correction is legitimate if rated examples for the same prompt are genuinely available before a new image is assessed. It would support a weaker claim if presented as performance on wholly new prompts. We make this information contrast explicit and do not infer that any published quality model or dataset has used an improper evaluation split.

Structured validation guidance Roberts et al., 2017 and distribution-shift reasoning Ovadia et al., 2019 explain why the unit withheld should match the generalization claim. Withholding a prompt group addresses one kind of novelty; withholding a generator, image subject, or source collection would address others. Our design keeps the held image set fixed so that an observed difference cannot be explained by changing which images are evaluated.

The prediction procedures also differ in their available fitting-row count. Whole-prompt withholding removes seven to ten records, while image-only withholding removes one. We report the image-only versus whole-prompt contrast to expose this component. It is not a perfect equal-size intervention isolating every aspect of prompt information, but it shows how much the simple generator reference changes before a peer correction is added.

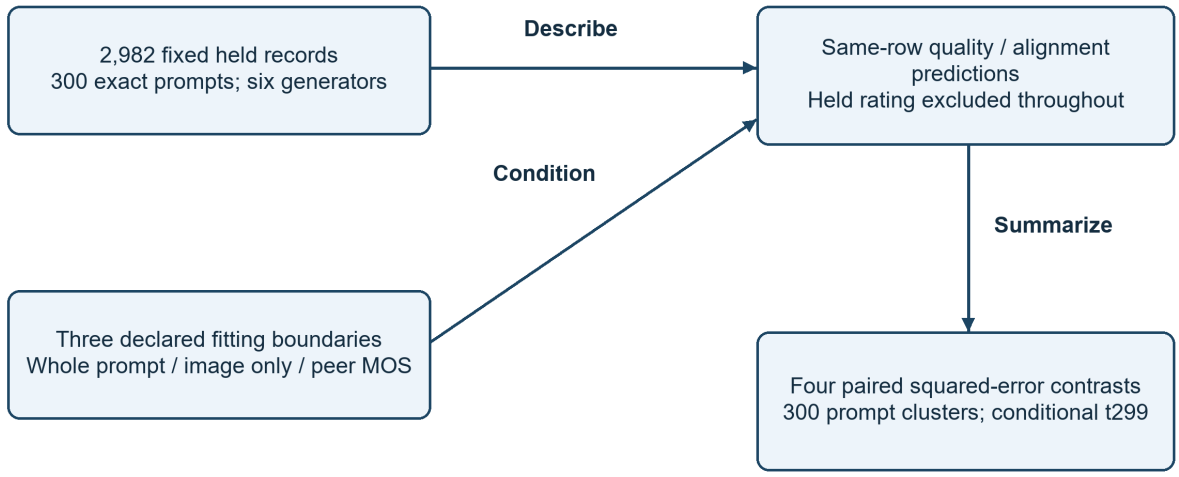

Figure 1. *Whole-prompt and image-only generator references define different fitting boundaries. The explicit peer route adds other exact-prompt residuals, with the held image's MOS excluded throughout.*

Validation-selection cautions Varma and Simon, 2006 are relevant if a researcher repeatedly changes the correction after seeing held loss. Here the unshrunk peer residual rule is fixed in the executed calculation. We do not search a shrinkage multiplier, clip predictions to improve a score, or discard unfavorable prompt groups. Missing-data reasoning Little and Rubin, 2002 also limits extrapolation: a prompt without available peer ratings cannot simply be assigned the same information as the peer-supported protocol.

5. Outcomes and working uncertainty

Each held image receives squared prediction errors under all three procedures. We calculate peer-corrected minus whole-prompt error and image-only minus whole-prompt error separately for quality and alignment. Four intercept-only OLS models summarize the paired differences. The raw scales remain separate, so a loss difference for alignment is not directly interpreted as the same substantive improvement as an equal numerical difference for quality.

Standard errors are clustered by the 300 exact prompt groups, with finite-sample correction and t299 intervals. Robust covariance White, 1980 and clustered-inference principles Cameron and Miller, 2015 motivate this grouping. Nevertheless, fitting sets overlap extensively and all records share six generator categories. The intervals are conditional working summaries, not full uncertainty for refitting all procedures on independent datasets. They also omit uncertainty in the original aggregate ratings because individual observer records are unavailable.

Squared-error scoring Gneiting and Raftery, 2007 supplies the declared numerical objective. Probability calibration Guo et al., 2017 is a different concept: the MOS predictions here are not probabilities of an event, and the peer correction does not become a calibrated confidence estimate. Measurement analysis Carroll et al., 2006 similarly cautions that improved aggregate-score prediction is not proof that a predictor has captured the latent construct that produced those scores.

6. The two rating constructs respond differently

For alignment, whole-prompt withholding produces MSE 0.383635. The image-only generator reference yields 0.383211, while the explicit peer correction gives 0.327440. The primary peer-minus-whole-prompt contrast is -0.056195, with 95% interval [-0.077704, -0.034686]. The source therefore contains useful same-prompt residual information for alignment under this particular unshrunk correction.

For quality, the corresponding MSE values are 0.425033, 0.424975, and 0.436703. The peer-minus-whole-prompt difference is +0.011671, with interval [-0.001478, 0.024819]. Its point estimate favors the simpler reference, but the interval includes zero. We do not describe this as demonstrated quality deterioration in a general population, and we retain it beside the stronger alignment result.

Table 1. Paired MSE differences under alternative prompt-information protocols.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.011671 (0.006681)	[-0.001478, 0.024819]	2982
M1: fit	R-squared = 0.0000	-	-	2982
M2	Mean outcome/contrast	-0.000058 (0.000048)	[-0.000152, 0.000036]	2982
M2: fit	R-squared = 0.0000	-	-	2982
M3	Mean outcome/contrast	-0.056195 (0.010930)	[-0.077704, -0.034686]	2982
M3: fit	R-squared = 0.0000	-	-	2982
M4	Mean outcome/contrast	-0.000424 (0.000053)	[-0.000528, -0.000319]	2982
M4: fit	R-squared = 0.0000	-	-	2982

Model labels: M1 = Quality MOS: Known-prompt peer correction minus novel-prompt reference; M2 = Quality MOS: Image-only generator mean minus novel-prompt reference; M3 = Alignment MOS:

Known-prompt peer correction minus novel-prompt reference; M4 = Alignment MOS: Image-only generator mean minus novel-prompt reference.

Models 1 and 2 concern quality; Models 3 and 4 concern alignment. Within each outcome, the first compares explicit peer correction with whole-prompt withholding, and the second compares image-only with whole-prompt generator means. All models use 2,982 observations, no controls, prompt-clustered standard errors, and effectively zero R-squared. Intervals are pointwise exploratory summaries across four reported contrasts rather than simultaneous family-wise guarantees.

The image-only versus whole-prompt differences are small: -0.000058 for quality and -0.000424 for alignment. The alignment interval excludes zero under the working calculation, yet its magnitude is much smaller than the peer-correction contrast. This distinction illustrates why a numerically detectable result need not explain the main empirical phenomenon. The explicit use of peer residuals matters much more here than their slight contribution to a generator's average.

Mean absolute errors tell a consistent descriptive story without replacing the declared squared-error objective. Alignment absolute error falls from 0.477151 under whole-prompt withholding to 0.436852 under peer correction. Quality absolute error rises from 0.488845 to 0.498051. These summaries are retained as secondary descriptions; no new selection of a preferred loss or correction rule follows from them.

7. Interpreting familiar-prompt evidence

The result does not establish that prompt difficulty is a stable causal property. The peer residual can combine semantic content, generator-by-prompt interactions, and rating practices. Causal-diagram reasoning Greenland et al., 1999 cautions against identifying a mechanism from a useful predictor alone. We observe no randomized intervention on prompt content or individual-rater behavior, and the correction's success for alignment does not prove the source of its information.

Technical-debt arguments Sculley et al., 2015 motivate preserving the exact fitting boundary with an evaluation result. A simple label such as held-out image is insufficient to reconstruct whether peer scores from its prompt were available. Dataset documentation Gebru et al., 2021 and FAIR principles Wilkinson et al., 2016 support retaining the prompt identifiers, excluded-row rules, predictions, and processed losses. The record also preserves that repeated prompts are grouping information, not evidence of duplicate pixels.

In a prospective system, peer ratings themselves would have a collection cost and a timing requirement. The current experiment treats released peer MOS as available information in one declared condition. It does not measure the expense of obtaining those ratings, test whether they precede a new generation in a real workflow, or establish a cost-effective annotation policy. Those operational questions should remain separate from the observed prediction difference.

The primary regressions weight images equally, rather than giving each prompt one observation. Because prompt groups contain seven to ten records, larger groups contribute slightly more to the point estimate. Clustering changes the uncertainty calculation but does not change that weighting. An equal-prompt analysis would answer a related question and should be named explicitly rather than assumed to result automatically from prompt-clustered standard errors.

Peer availability also varies with group size. Each held image has between six and nine other rated examples for the exact prompt, and their generator composition can differ. We use their unweighted residual mean rather than fitting a prompt-specific variance or a generator-specific reliability weight. The

quality result may therefore reflect limitations of this simple correction as well as weak shared prompt information. It should not be generalized to every possible way of using familiar-prompt data.

No text similarity is estimated between different prompts. The whole-prompt-withheld reference is intentionally crude, and a model using prompt meaning could outperform it on genuinely new strings. Its role here is to make the information boundary explicit. Demonstrating a benefit from existing peer ratings is not the same as demonstrating a benefit from understanding a previously unseen prompt's content.

8. Conclusion

The same-prompt peer correction improves alignment prediction but provides no corresponding quality improvement in the captured data. Familiar-prompt and whole-prompt-withheld evaluations therefore describe materially different information tasks, with different consequences for the two rating constructs. Reporting that boundary makes the result interpretable without accusing an existing system of leakage or treating a metadata correction as an image-quality model.

References

1. 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
2. Huang, Kaiyi, et al. "T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation." *arXiv preprint arXiv:2307.06350*, 2023. https://arxiv.org/abs/2307.06350.
3. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
4. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
5. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
6. 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.
7. Wang, Jianyi, et al. "Exploring CLIP for Assessing the Look and Feel of Images." *arXiv preprint arXiv:2207.12396*, 2022. https://arxiv.org/abs/2207.12396.
8. 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.
9. 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.
10. 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.
11. Li, Chunyi; Kou, Tengchuan; Gao, Yixuan; Cao, Yuqin; Sun, Wei; Zhang, Zicheng; Zhou, Yingjie; Zhang, Zhichao; Zhang, Weixia; Wu, Haoning; Liu, Xiaohong; Min, Xiongkuo; Zhai, Guangtao. (2024). AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. arXiv:2404.03407. https://arxiv.org/abs/2404.03407.
12. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
13. 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.

14. Ghosh, Dhruba, et al. "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment." *arXiv preprint arXiv:2310.11513*, 2023. https://arxiv.org/abs/2310.11513.

15. 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.

16. Meng, Zijie. (2026). Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment. arXiv:2606.16799. https://arxiv.org/abs/2606.16799.

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. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

19. 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.

20. Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.

21. 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.

22. Hu, Yushi, et al. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering." *arXiv preprint arXiv:2303.11897*, 2023. https://arxiv.org/abs/2303.11897.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

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

29. Xu, Jiazheng, et al. "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation." *arXiv preprint arXiv:2304.05977*, 2023. https://arxiv.org/abs/2304.05977.

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