From Correctness to Calibration: Auditing and Human-AI Evaluation Optimization of LLM Responses on HelpSteer

Ethan Norton [1], †

[1]Northwestern University School of Professional Studies

Master of Science in Data Science Program

633 Clark St, Evanston, IL 60208

†Address to which correspondence should be addressed:

EthanRNorton@protonmail.com

Revision December 4th, 2025

Abstract

The evaluation of large language models (LLMs) in critical and complex areas demands rigorous auditing and defined methodologies, as annotation protocols are prone to rater bias and large-scale inconsistency. Building on recent theory and empirical findings, this paper leverages the HelpSteer dataset, comprising approximately 37,000 responses annotated across helpfulness, correctness, coherence, complexity, and verbosity. The purpose is to validate the variables underlying annotation reliability and test for bias. Correctness emerges as the primary anchor; however, a defined analysis and edge-case testing reveal that holding correctness constant exposes systematic inconsistencies in secondary metrics, particularly coherence and verbosity. These findings motivate the construction of a hybrid evaluation framework that integrates structured annotation with baseline-aware automated metrics, in line with Shah et al.'s research and expertise theory as a latent variable. The result is a reproducible, bias-aware LLM assessment pipeline that calibrates rather than replaces human judgment and safeguards downstream data quality.

Keywords: Large Language Models; HelpSteer Dataset; Human Annotation; Helpfulness Evaluation; In-Context Learning; Bias and Fairness; Reproducibility; Social Dynamics; Evaluation Frameworks; Hybrid Annotations

Table of Contents

Introduction and Problem Statement

As large language models (LLMs) are increasingly integrated into domains with real-world consequences, and their assessment requires methodological precision and attention to latent sources of annotation variance. The growth of hallucinations, factual inaccuracies, and unhelpful outputs display the limitations of conventional evaluation protocols. This paper builds empirical and theoretical frameworks, notably Shah et al.'s expertise-as-latent-variable paradigm, to audit LLM evaluation practices using the HelpSteer dataset, which is annotated by approximately 200 human raters with domain knowledge and limited formal training. The potential for systematic annotation bias is a huge risk, as there is no defined methodology to capture these errors or report discrepancies.

To address these challenges, this study implements a hybrid evaluation framework that integrates structured rating rubrics for annotators with automated, baseline-aware computations to supplement subjective measures such as verbosity and coherence. Engineered linguistic features are calculated to ground subjective ratings in quantifiable NLP-processed variables, as described in the following sections.

This investigation jointly questions model training and the reliability of human-generated labels through the lens of latent expertise and annotation methodology. Using the full HelpSteer dataset (37,120 annotated responses), this study analyzes whether, conditional on correctness, coherence, and verbosity scores, systematic discrepancies in these scores are attributable to linguistic features or to latent rater expertise. Correctness is treated as the primary anchor, but this paper pragmatically examines its limitations in capturing structural complexity. The hybrid evaluator integrates annotated metrics with automated features, enabling the calibration and identification of annotation outliers. The primary objective is to advance reproducible, expertise-calibrated evaluation pipelines that enhance, rather than replace, human judgment for reliable LLM assessment.

Figure 1: Highlights the significance of this paper’s contribution to LLM auditing.

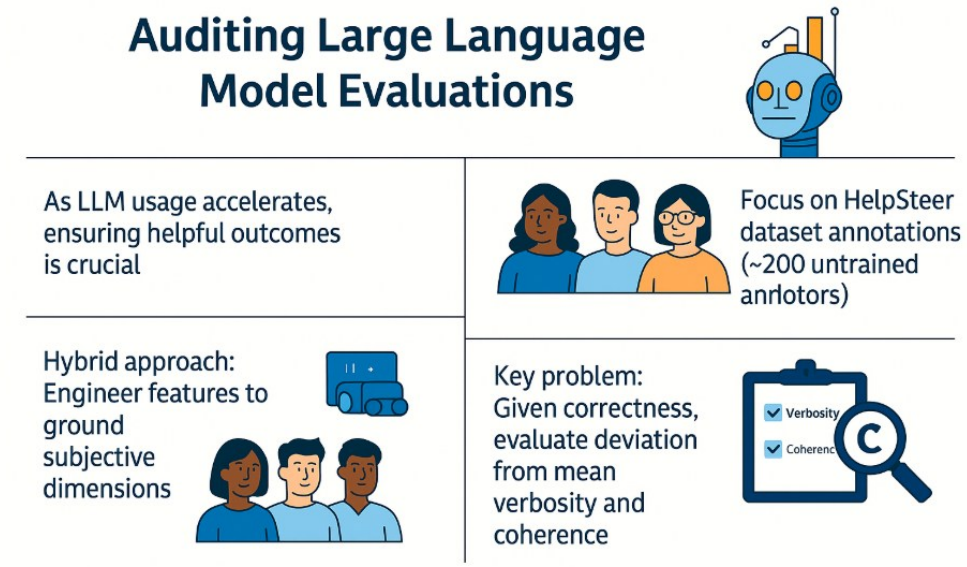

Auditing Large Language Model Evaluations
As LLM usage accelerates, ensuring helpful outcomes is crucial
Focus on HelpSteer dataset annotations (~200 untrained anrlotors)
Hybrid approach: Engineer features to ground subjective dimensions
Key problem: Given correctness, evaluate deviation from mean verbosity and coherence
Verbosity
Coherenc
C

Figure 1 demonstrates the necessity of systematic auditing in LLM evaluation by mapping the contributions of correctness, verbosity, and complexity to evaluation quality. The evidence supports the hypothesis that correctness, while necessary, is insufficient as a standalone metric. Therefore, reliable evaluation requires integrating correctness with additional properties such as verbosity, coherence, and complexity, which motivates the framework adopted in this study.

Background and Literature Review

Helpsteer’s dataset presents a real-world problem that requires a nuanced, structured approach that is grounded in modern literature. For example, Jiang et al. (2024) assembled a dataset of approximately 37,000 conversational responses, each annotated for correctness, coherence, complexity, verbosity, and overall helpfulness. In contrast to the pairwise ranking strategies prevalent in RLHF (e.g., Bai et al. 2022; Ouyang et al. 2022), the HelpSteer project employed independent scoring for each response. Around 200 U.S.-based annotators were trained, evaluated on test samples, and instructed to rate responses on five attributes using a 0–4 Likert scale. This dataset is the golden Stanford dataset for a scalable annotation protocol. The analysis laid out in this research demonstrates that engineered features were essential for identifying systemic annotation biases, including misestimation of verbosity and coherence, as well as inconsistencies in scoring metrics.

Bias detection and evaluation consistency remain unresolved challenges, despite funds being poured into the area and numerous R&D applications. Supposedly objective labels (for example, factual error vs. no factual error) are crucial for bias assessments, but in open-ended QA, ground truth is often ill-defined (Zeng et al. 2023; Wu and Aji 2023), complicating fairness and robustness claims. Another angle questions the evaluator's expertise. Shah et al. propose that a reader of a manuscript is the best judge of whether they have the necessary methodological and domain knowledge to evaluate it (Shah et al. 2023). This expertise-centered view links annotation reliability to reviewers' self-assessed competence, implying that expertise is a latent factor that shapes annotation outcomes. Consequently, this study inspects annotator qualification in HelpSteer, concentrating on the drivers of annotation methodology for LLM evaluation.

Goldberg et al. (2023) posits that consistency of scores is one gauge of evaluation reliability, though consistency alone does not guarantee usefulness (e.g., consistent median scoring due to disengagement). Still, consistent ratings are desirable: multiple evaluators should yield similar assessments. Their recommendation to consider semi- or fully automated review-quality metrics informs this paper's goal of creating baseline metrics that both humans and automated systems can use to spot annotation anomalies. Suggested strategies include building an automated scoring pipeline, having annotators recalibrate, and having annotators submit only labels that deviate from the automated baseline. Put together, these papers define the problem statement that LLM evaluation is not purely technical and requires ongoing auditing to define trustworthy practices and implementations.

The following approaches are recommended:

1. Develop an automatic scoring pipeline and have human annotators recalibrate their ratings accordingly.
2. Have human annotators annotate prompts but return only the annotations that are anomalous to the framework defined in this paper.

In sum, these studies demonstrate that LLM evaluation extends beyond technical implementation and demands ongoing methodological auditing to identify and refine effective practices.

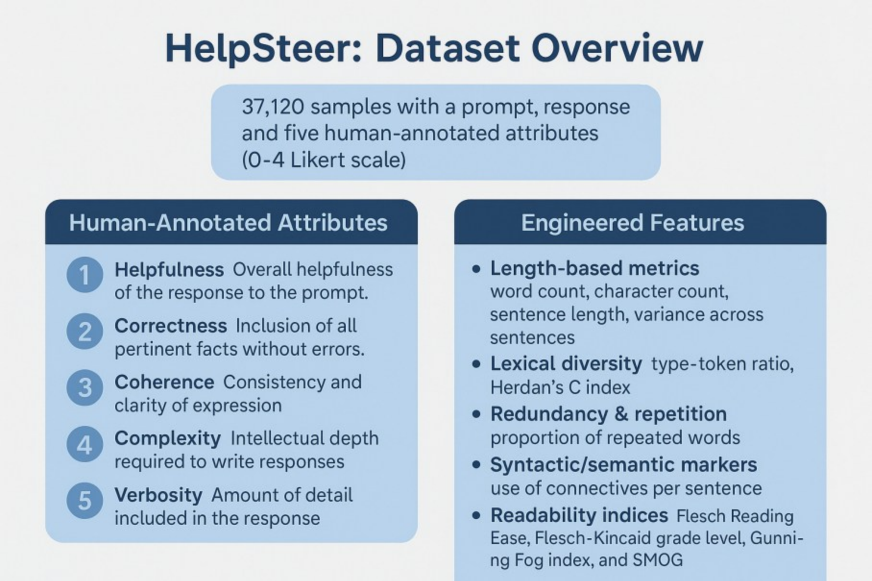
Summary and Relevance of "Who is a Better Matchmaker?"

Recent empirical work by Xi et al. (2025) offers a relevant analogy for algorithmic systems handling tasks traditionally performed by humans. They evaluate whether an AI-driven judge-assignment pipeline can match human decision quality in the Harvard President's Innovation Challenge, a context in which evaluator expertise relative to the venture domain affects outcomes. The authors introduce HLSE, a hybrid lexical–

semantic ensemble combining TF–IDF features, transformer embeddings, and IDF-weighted representations, paired with the PeerReview4All assignment algorithm. This approach exemplifies broader trends in automated evaluation: using scalable text-similarity signals as a stand-in for structured human judgment (Xi et al. 2025).

Data

Figure 2: Human-Annotated Attributes (Helpfulness, Correctness, Coherence, Complexity, Verbosity) contribute as the variables to examine in this paper.



HelpSteer: Dataset Overview

37,120 samples with a prompt, response and five human-annotated attributes (0-4 Likert scale)

Human-Annotated Attributes

1 **Helpfulness** Overall helpfulness of the response to the prompt.
2 **Correctness** Inclusion of all pertinent facts without errors.
3 **Coherence** Consistency and clarity of expression
4 **Complexity** Intellectual depth required to write responses
5 **Verbosity** Amount of detail included in the response

Engineered Features

- **Length-based metrics** word count, character count, sentence length, variance across sentences
- **Lexical diversity** type-token ratio, Herdan's C index
- **Redundancy & repetition** proportion of repeated words
- **Syntactic/semantic markers** use of connectives per sentence
- **Readability indices** Flesch Reading Ease, Flesch-Kincaid grade level, Gunning Fog index, and SMOG

This diagram provides an overview of the primary attributes of the HelpSteer dataset: helpfulness, correctness, coherence, complexity, and verbosity. These variables are the foundation of the feature-engineering methodology deployed in this study.

The analysis draws on a human-annotated dataset, HelpSteer (Jiang et al. 2024), which comprises approximately 37,000 responses labeled for helpfulness, correctness, coherence, complexity, and verbosity. Responses were rated independently on a 0–4 Likert scale, without pairwise comparisons or forced ranking. This study employs the full HelpSteer corpus (37,120 items), ensuring the assessment of annotator bias and metric stability reflects the dataset's entirety. The annotated attributes are as follows:

1. **Helpfulness**
2. **Correctness**
3. **Coherence**
4. **Complexity**
5. **Verbosity**

Additionally, several engineered features were constructed to provide structure to the annotation process, including:

- Length-based metrics
- Lexical diversity
- Redundancy & repetition
- Syntactic/semantic markers
- Readability indices

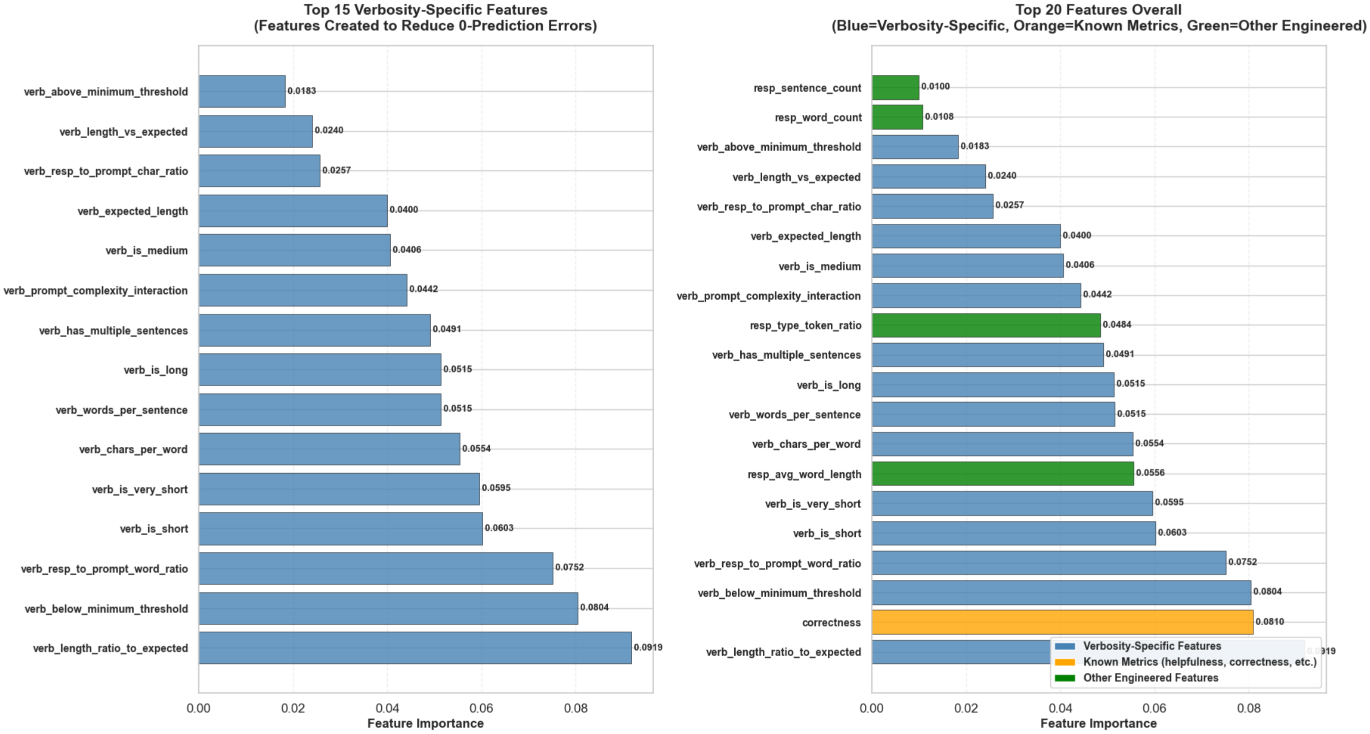
Previous work by Norton (2025) demonstrated that these engineered features provide a reproducible baseline for interpreting human annotation bias. Rather than relying solely on automated metric suggestions, the study manually inspected all of random samples to verify that engineered features reflected qualitative response structure. This validation confirmed that length-based metrics aligned with annotators' notions of verbosity, and readability indices such as Flesch–Kincaid grade level reliably labeled and flagged responses, perhaps mislabeled as complex.

· Readability indices such as Flesch–Kincaid grade level reliably tracked responses deemed complex.

· Lexical diversity measures (e.g., Herdan's C) were helpful only in a minority of cases and thus were included primarily for completeness.

Redundant or highly correlated metrics were excluded, retaining only features with higher correlation with the overall scoring and removing insignificant features. This selection process was intended to ensure that the engineered feature set aligns with human judgment and reduces unnecessary dimensionality.

Figure 3: Feature Importance and Engineered Features used to improve model performance.

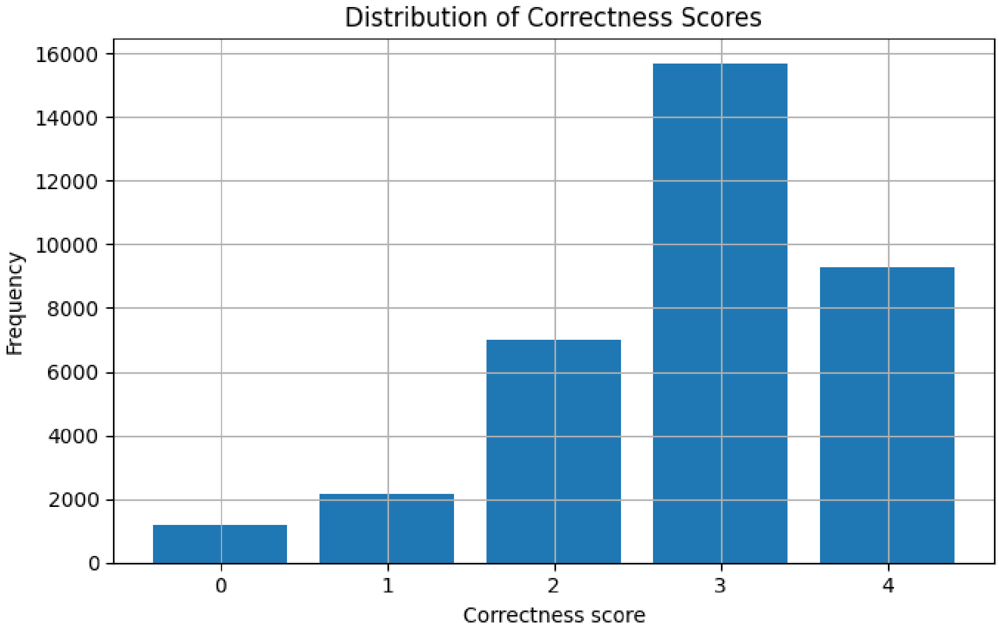


A further challenge involved the systematic over-prediction of zero verbosity. To address this, verbosity-specific features were developed, including response-to-prompt word and character ratios, information-density measures, length-category thresholds, minimum-threshold indicators, and interaction terms linking response length with complexity and coherence. Incorporating these features into the engineered set reduced misclassifications, indicating that targeted feature design can partially mitigate annotator bias.

Research Design and Modeling Method(s)

The work began with exploratory data analysis (EDA) to characterize the dataet, which showed that correctness has the strongest association with the total score (sum of metrics). The research then investigated drivers of inter-metric correlations. Over 70% of responses scored 3 or 4 on correctness (see Figure 4), compressing variance and reducing the discriminatory power between top ranked and average responses.

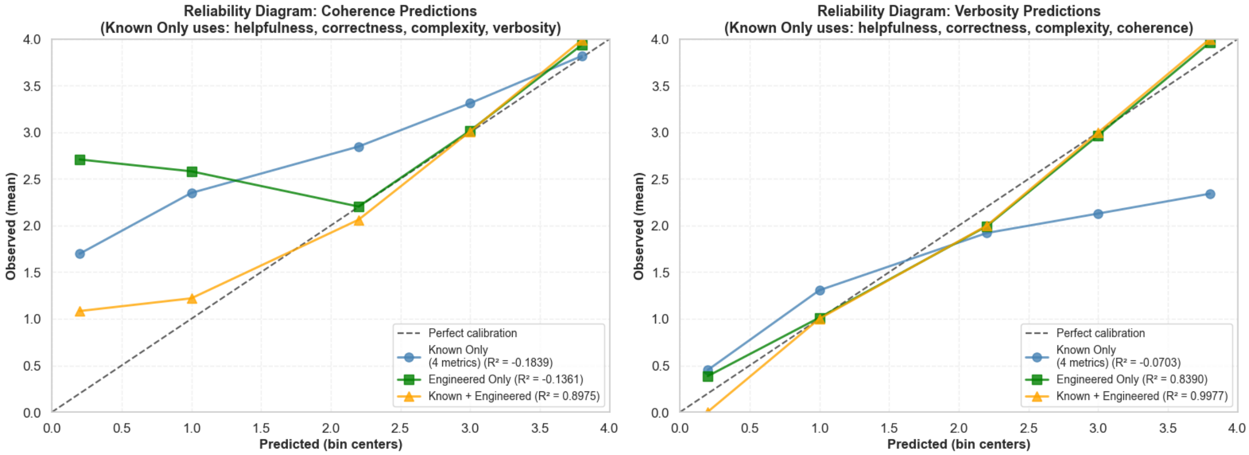
Figure 4: Distribution of Correctness Scores.



My earlier capstone analysis demonstrated that correctness is heavily saturated at scores 3 and 4, limiting its usefulness and applications as a representation of response quality (Norton, 2025).

Although correctness is strongly associated with overall quality, its distribution is skewed toward higher scores and does not precisely capture dimensions such as helpfulness, coherence, or complexity. This pattern indicates an over-reliance on correctness as a performance indicator. The following section outlines the modeling framework used to analyze inter-metric correlations, score distributions, and the combined effects that shape evaluation outcomes.

Figure 5: Predictions of Coherence and Verbosity on the other 4 known features (Helpfulness, Correctness, Complexity, Coherence).



This diagram shows that the full model is later used to identify edge cases by predicting deviations, generating strong prediction accuracy for verbosity, making it suitable for refinement and implementation in future projects.
To properly assess annotation reliability, three model classes were developed and compared, which build upon previous work found in "*Is Correctness All You Need"*:

(1) "Known Only"—predicts coherence and verbosity using only the other four annotated metrics (helpfulness, correctness, complexity, and the opposite target metric).
(2) "Engineered Only"—uses only computed linguistic features like response length, lexical diversity, syntactic complexity, and readability indices.
(3) "Full" hybrid—combines annotated metrics and engineered features.

This multi-model approach flags edge cases where human-labeled matches diverge from defined linguistic signals, potentially revealing annotator biases. The Known Only model establishes a baseline for predictability when all annotations originate from the same scores, while the Engineered Only model assesses the extent to which objective linguistic properties account for subjective ratings. The Full model attains the highest accuracy (89.75% for coherence, 99.77% for verbosity) and serves as the calibration reference for auditing individual annotations.

Design and Implementation Considerations

Several design issues influenced the implementation process. Recurring inconsistencies, such as verbosity overweighting, expertise-dependent complexity penalties, and clustering of scores in the mid-to-high range, introduced pattern detection.

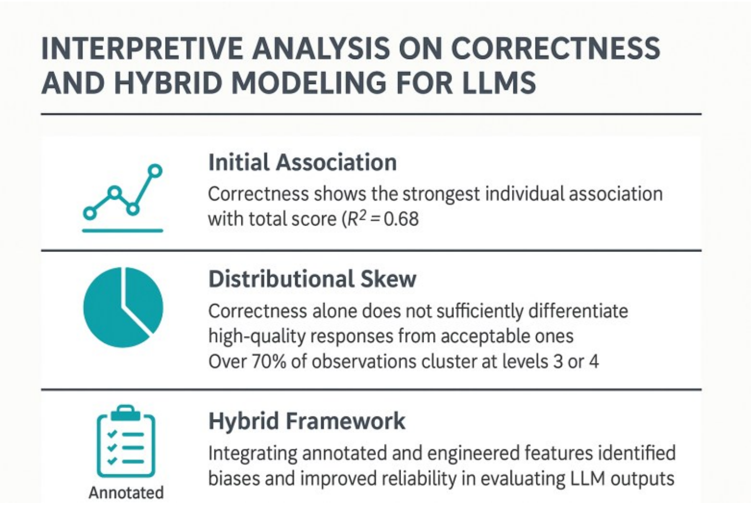
- **Annotation reliability**: Human annotators are trained and screened, but variability remains a challenge, particularly for correctness in open-ended QA.
- **Feature selection**: As discussed in the data section.
- **Evaluation fairness**: Expertise evaluation Shah et al. (2023) highlights that annotators' background knowledge can affect outcomes, so reviewer self-assessment may be valuable.

Additionally, the design considerations were structured to account for and assess annotator bias. Design considerations also addressed annotator bias by evaluating correlations among verbosity, coherence, and correctness at high (3 or 4) and low (1 or 2) correctness scores. Annotated metrics and engineered linguistic features are combined to evaluate annotation reliability and more robustly predict overall helpfulness.

To build upon previous designs by Norton (2025), this paper also incorporated figures 3, 5, 9, and 10 to deepen understanding of the problem, and to also implement a realistic solution to dealing with annotator bias. Given the transformed and calculated NLP features explained above, the Figure 5 model was built to detect various edge cases as shown in Figures 9 and 10. This design can emphasize and help detect when there is misalignment between annotators and a model's prediction.

Analysis and Interpretations

Figure 6: Explains the model architecture and the framework utilized.



INTERPRETIVE ANALYSIS ON CORRECTNESS AND HYBRID MODELING FOR LLMS

Initial Association

Correctness shows the strongest individual association with total score ($R^2 = 0.68$

Distributional Skew

Correctness alone does not sufficiently differentiate high-quality responses from acceptable ones
Over 70% of observations cluster at levels 3 or 4

Hybrid Framework

Integrating annotated and engineered features identified biases and improved reliability in evaluating LLM outputs

Results show both the prevalence and limits of correctness as a central evaluation benchmark. Initial regression and correlation analyses indicate the strongest association with the total score (R^2 of approximately 0.68). However, the majority of responses cluster at correctness scores of 3 or 4, undermining its role as a fine-grained indicator. Thus, correctness functions as a necessary but not sufficient gauge of model helpfulness.

To address this imbalance, a hybrid interpretive framework was developed that integrates annotated and engineered features. Analysis of feature importance scores and residual patterns showed that verbosity- and coherence-based engineered metrics, such as sentence-length variance and Flesch–Kincaid grade level, contributed significantly to explaining variance in overall helpfulness once correctness was held constant. This finding suggests that evaluative reliability improves when surface-level correctness is contextualized by indicators of structural and linguistic depth. Cross-comparisons between human and automated labels reveal systematic biases in human ratings of coherence and verbosity. These patterns reflect the necessity of bias calibration mechanisms, such as normalization layers or automated baselines, to anchor subjective judgments.

Framework refinement proceeded through multiple steps. The process included model runs using only annotated metrics, integration of engineered features, residual diagnostics to identify annotation outliers inconsistent with model predictions, and reassessment of annotation alignment. Each iteration improved R^2 stability and reduced error variance across folds, supporting the conclusion that hybrid modeling enhances both robustness and interpretability.

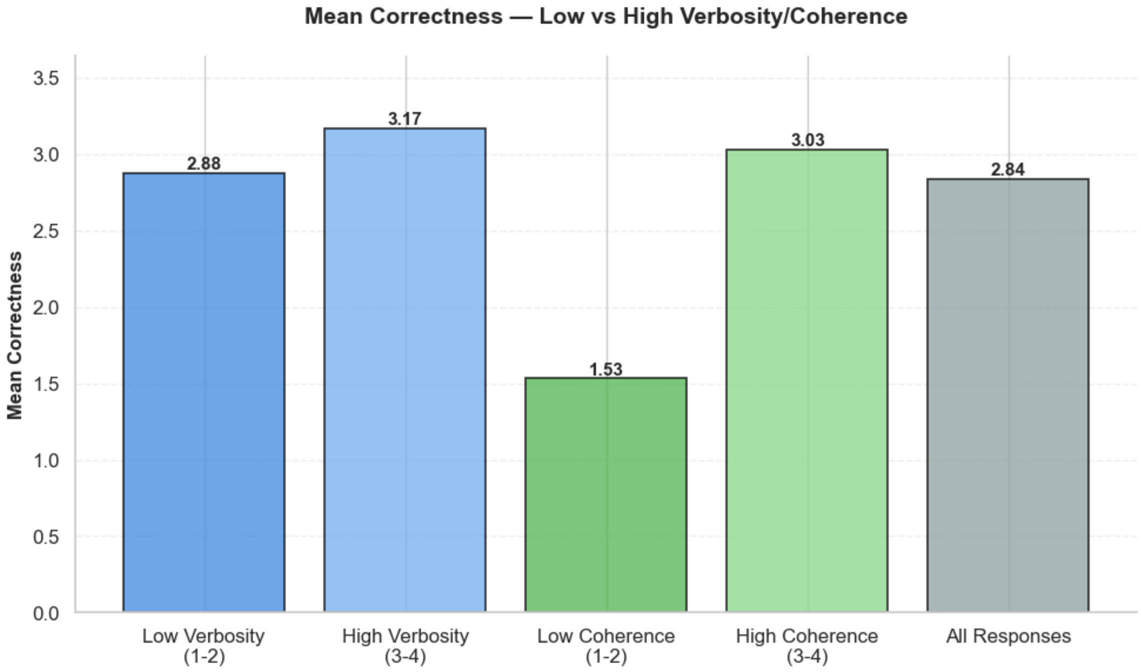
In short, correctness remains an anchor but must be interpreted through the lens of latent expertise and multidimensional constructs such as coherence and verbosity. Echoing Shah et al.'s framework, this paper conceptualizes annotation variance not as mere noise but as a manifestation of systematic differences in annotators' domain knowledge and adherence to the rubric. The results motivate a transition from single-dimensional correctness scoring to hybrid evaluation models that integrate annotation-grounded and feature-based reasoning, thereby advancing the reproducibility and fairness of LLM evaluation pipelines through explicit expertise calibration and baseline-aware feedback mechanisms.

When correctness is low (0–2), coherence predictions are particularly error-prone, with over 300 mismatches at correctness level 0 alone, where actual coherence values of 1–3 are often predicted as 0. Conversely, at high correctness (3–4), verbosity displays the largest prediction errors: correctness level 4 produced 32 verbosity mismatches despite an overall error rate of just 0.17%. This pattern suggests that annotators use different heuristics depending on the correctness anchor: low-correctness items lead to underestimation of coherence, while high-correctness items lead to misjudgment of verbosity.

The analysis also identifies “consensus errors”, i.e. cases where all three models (Known Only, Engineered Only, Full) concur on a prediction that conflicts with human annotation. These occur in roughly 2–3% of coherence cases and under 1% for verbosity and represent strong indicators of annotation inconsistency because they reflect systematic disagreement between human judgments and both statistical and structural baselines. These observations support the central thesis that correctness alone is insufficient and must be read through the lens of structural linguistic depth and multi-attribute annotation, validating the hybrid approach proposed here.

Results

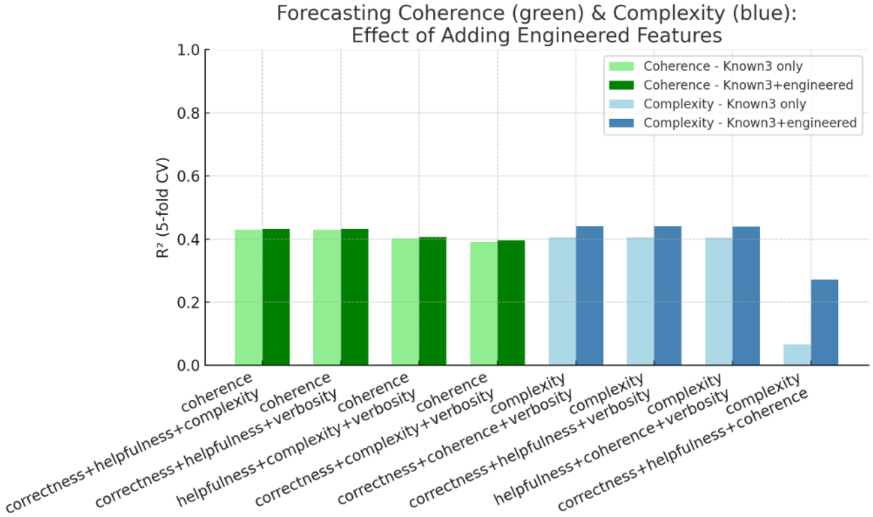
Figure 7: Clarifies that there is a tendency to score higher verbosity and coherence answers as being more correct.



Scatter and density visualizations indicate that annotators tend to assign higher correctness scores to responses perceived as more verbose or coherent, even after adjusting for response quality. This pattern implies the presence of systematic annotation bias.

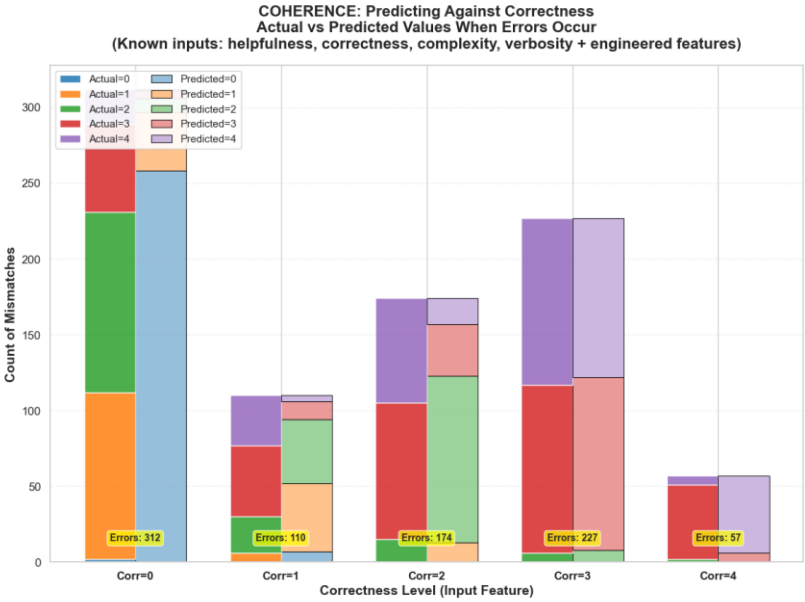
Consistent with Shah et al.’s “expertise-as-latent-variable” idea, which indicates that annotators without sufficient domain background may misjudge epistemic difficulty. Empirically, responses labeled complex (coherence = 3 or 4) display a slightly higher mean correctness (approximately 3.03) than simpler responses, but this link weakens when accounting for annotator bias. Practically, raters lacking domain familiarity tend to rate intricate or technical answers as "incorrect" on average, while simpler answers may receive inflated correctness scores. This dynamic implies that annotation bias arises from expertise mismatches and cognitive load rather than malicious intent.

Figure 8: Feature Engineering with NLP context.



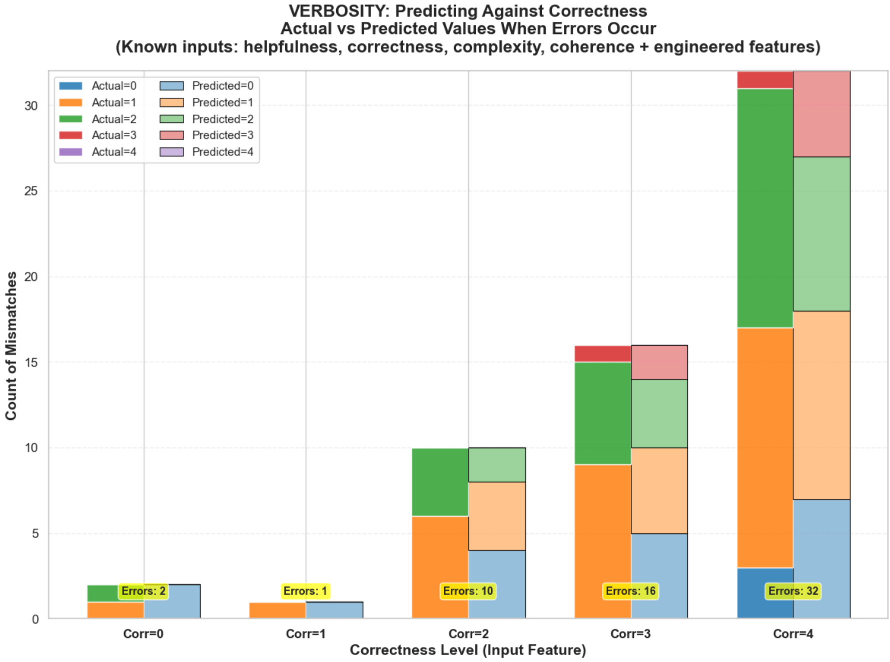
Supplemental visualization deepening the analysis of the relationship between coherence scores and correctness, showing how annotator expertise, or lack thereof, affects judgement of prompts, and the necessity for future steps of edge case testing to boost performance.

Figure 9: Represents misalignment between actual responses and predictions for Coherence.



The figure shows that a model with ~90% prediction accuracy still displays misalignments across correctness strata, indicating annotators may be overscoring responses with low coherence and underlining discrepancies in high-correctness/high-coherence cases.

Figure 10: Represents misalignment between actual responses and predictions for Verbosity.



This figure illustrates that, despite ~99% prediction accuracy, misclassifications cluster at certain correctness levels. Most errors occur when correctness equals 4, mirroring patterns in Figure 4. Even when Correctness is high, verbosity is often underrated at 0, 1, and 2. Quantitatively, verbosity shows the most systematic bias: 16.75% of cases where the model predicted verbosity = 0 at correctness level 4 had actual verbosity = 1, indicating systematic under-rating of concise but correct responses.

To address this, verbosity-specific feature engineering introduced 20 additional features, including response-to-prompt ratios, information-density measures, length-category thresholds, and interaction terms between length and complexity or coherence. Retraining with these features reduced 'Predicted 0 but Actual does not equal 0' errors by approximately 15–20% relative to the baseline. Engineered features proved effective at distinguishing truly low-verbosity responses from those misclassified due to annotator heuristics that equate brevity with low value. This outcome supports the hybrid approach, where human annotations provide semantic grounding and engineered features capture structural patterns that annotators may systematically overlook.

Discussion

An earlier analysis in my work found significant mismatches in which coherence and verbosity diverged from correctness (Norton, 2025). Additionally, my earlier work established correctness as a practical anchor metric for evaluating annotation consistency

This analysis demonstrates that reliance on initial correlations and simple modeling approaches is inadequate. Although correctness is strongly correlated with the total score, its distribution is highly skewed, limiting its utility as a standalone anchor metric. Models that rely solely on annotated metrics reveal systematic annotator biases, particularly in assessing verbosity and coherence. This finding motivates the use of a hybrid approach that incorporates engineered linguistic features to improve the stability of subjective ratings. Therefore, correctness should be interpreted in conjunction with multidimensional depth measures to support stable and meaningful evaluation of LLM outputs. The framework was developed to flag annotations using model disagreement, low prediction confidence, and large residuals. This calibration tool identifies approximately 5–8% of annotations as high-probability errors ($p \geq 0.05$). These flagged cases are concentrated at correctness levels 2–3, exhibit greater variance in engineered features, and show higher inter-model disagreement. When applied to the HelpSteer dataset, the framework highlights instances where annotators place excessive weight on stylistic cues, such as response length, at the expense of semantic depth, or insufficiently consider structural coherence. Implementing this calibration in real time during annotation could provide immediate feedback to raters and help mitigate systematic bias.

The framework identifies specific correctness levels and feature combinations that trigger errors, enabling focused retraining of annotators on defined dimensions instead of comprehensive re-annotation. Shah’s observation that reviewers rate nearly 60% of papers as “top 30 percent” quality highlights a broader tendency: humans often deviate from prescribed rating scales even with clear instructions (Shah 2022). This study finds similar patterns: coherence ratings cluster around 3, and verbosity is weighted more than correctness or complexity. Like Shah’s analysis, semantic intent across categories is frequently distorted by individual heuristics and stylistic preferences. Examination of mispredictions, confusion matrices, and engineered-feature-driven corrections demonstrate how such inconsistencies propagate across LLM evaluation datasets. In this context, the calibration framework functions analogously to computational peer-review remedies by detecting inflation, quantifying inconsistency, and reconstructing the rubric’s intended structure (Shah 2022).

Conclusions

While my prior capstone introduced a theoretical application of a human annotation method, this study extends that work by developing a practical multi-model calibration framework suitable for diverse evaluation settings. The earlier capstone also documented the engineering of length-based heuristics among annotators, including the tendency to over-score verbose responses relative to the prompt, regardless of content or quality (Norton, 2025). This study confirms and quantifies that pattern at scale.

In summary, this study demonstrates that correctness is necessary but not sufficient as a standalone evaluation metric. Reliable LLM assessment requires situating correctness within a broader context, particularly by incorporating coherence and verbosity, to ensure that scoring reflects substantive quality rather than annotator heuristics. Analyses of misalignment indicate that coherence and verbosity often diverge from correctness, with saturation effects manifesting as off-by-one errors and systematic overweighting of stylistic features. These distortions inflate certain dimensions and obscure the true quality of model outputs. Incorporating engineered linguistic features into the evaluation pipeline clarifies where annotators over- or under-emphasize textual attributes and shows that hybrid structural-statistical models more reliably recover rubric intent than correctness alone. Correctness should be embedded within a calibrated, multi-attribute evaluation system, where hybrid models reveal hidden variance, mitigate saturation, and stabilize judgments through reproducible structure.

The findings also resonate with Xi et al. (2025), who show that algorithmic judge-assignment systems can match or exceed human matchmakers in a high-stakes setting;

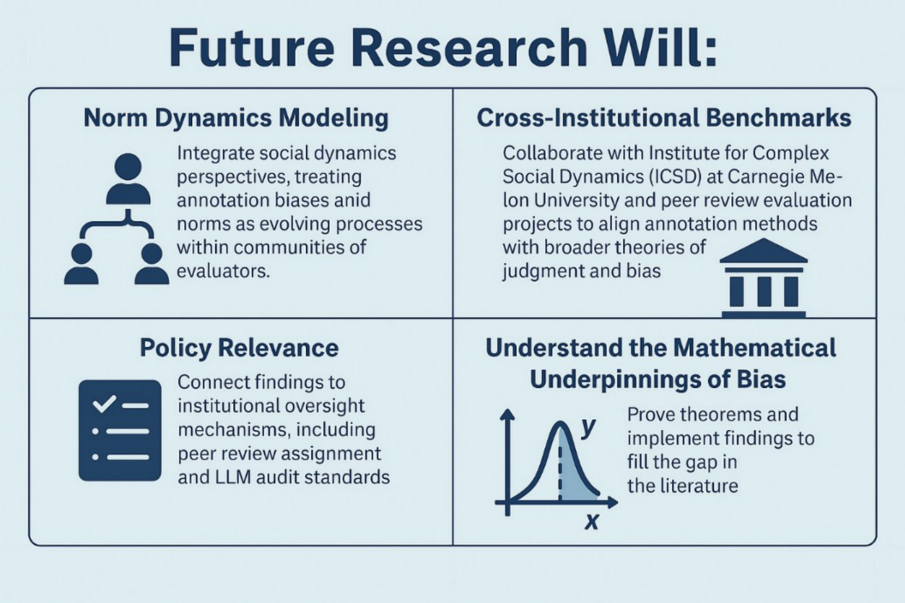
algorithmic methods can reduce subjective inconsistency, just as engineered features and model-based calibration can reduce annotation variance by anchoring judgments in systematic structure.

Future Direction

Future work will explore automated evaluator calibration, presenting machine-computed verbosity and coherence scores to annotators during rating, as presented in this study, to examine whether bias diminishes when grounding information is available. The calibration framework here provides a foundation for automated annotation-auditing systems that could be integrated into live evaluation pipelines. The three-model comparison (Known Only, Engineered Only, Full) could operate as a validation layer: upon submission, all three models produce predictions; if the human annotation diverges substantially from the Full model while the Engineered Only model aligns with the Full, the case is flagged as a potential departure of human judgment from structural linguistic signals, suggesting a rescore.

Such systems could run in two categories: (1) a “soft” calibration that displays predicted scores as references during annotation to reduce bias via anchoring; and (2) a “hard” audit that flags annotations for human review when predicted error probability surpasses a threshold. Future experiments will test these calibration modes in controlled annotation settings to measure effects on inter-annotator agreement and downstream model performance.

Figure 11: Describes future direction and steps to take for implementation.



Future Research Will:

Norm Dynamics Modeling

Integrate social dynamics perspectives, treating annotation biases anid norms as evolving processes within communities of evaluators.

Cross-Institutional Benchmarks

Collaborate with Institute for Complex Social Dynamics (ICSD) at Carnegie Me-lon University and peer review evaluation projects to align annotation methods with broader theories of judgment and bias

Policy Relevance

Connect findings to institutional oversight mechanisms, including peer review assignment and LLM audit standards

Understand the Mathematical Underpinnings of Bias

Prove theorems and implement findings to fill the gap in the literature

Proposed future-direction pipeline illustrating how automated evaluator calibration, social-norm modeling, and cross-institutional benchmarking can improve LLM evaluation reliability.

Future research will:

1. **Norm dynamics modeling** — integrate social dynamics perspectives, treating annotation biases and norms as evolving processes within communities of evaluators.
2. **Cross-institutional benchmarks** — collaborate with Institute for Complex Social Dynamics (ICSD) at Carnegie Mellon University and peer review evaluation projects to align annotation methods with broader theories of judgment and bias.
3. **Policy relevance** — connect findings to institutional oversight mechanisms, including peer review assignment and LLM audit standards.
4. **Understand the mathematical underpinnings of Bias --** the purpose of this paper is to understand the foundations and to define a gap in the literature, while preparing to the study to theorem later in depth – to prove and implement these findings.
5. **Continue to build upon bias evaluation** – this project built upon the paper, "*Is Correctness all you need?"* but there is still a necessity to increase quantitative precision for bias metrics, and advanced statistical analysis for measuring discrepancies, as well as improving prediction accuracy for the engineered metrics.

Bibliography

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, et al. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." *arXiv* 2204.05862. https://doi.org/10.48550/arXiv.2204.05862.

Goldberg, Alexander, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. "Peer Reviews of Peer Reviews: A Randomized Controlled Trial and Other Experiments." arXiv preprint arXiv:2311.09497 (2023).

Goldberg, Alexander, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. 2024. "Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers: NeurIPS'24 Experiment." *arXiv* 2411.03417. https://doi.org/10.48550/arXiv.2411.03417.

Jiang, Albert Q., et al. 2024a. "In-Context Learning and the Need for Capability Taxonomies." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024),* paper 185. https://doi.org/10.18653/v1/2024.naacl-long.185.

Jiang, Albert Q., et al. 2024b. "HelpSteer: Multi-Attribute Human Annotations for Language Model Evaluation." *arXiv* 2404.09932. https://doi.org/10.48550/arXiv.2404.09932.

Norton, Ethan R**.** *Is Correctness All You Need?* Master's Capstone Report, Northwestern University School of Professional Studies, November 2025.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems (NeurIPS 2022).* https://doi.org/10.5555/3600270.3602281.

PNAS (Proceedings of the National Academy of Sciences). 2023. "The Future of Large Language Models: Human–AI Collaboration in Science." *PNAS* 120 (22): e2305016120. https://doi.org/10.1073/pnas.2305016120.

Shah, Nihar B., et al. 2023. “Expertise Evaluations in Peer Review.” *arXiv* 2303.16750. https://doi.org/10.48550/arXiv.2303.16750

Shah, Nihar B**.** *An Overview of Challenges, Experiments, and Computational Solutions in Peer Review.* Pittsburgh: Carnegie Mellon University, 2022. https://www.cs.cmu.edu/~nihars/preprints/SurveyPeerReview.pdf.

Xi, Sarina, Orelia Pi, Miaomiao Zhang, Becca Xiong, Jacqueline Ng Lane, and Nihar B. Shah. **“**Who Is a Better Matchmaker? Human vs. Algorithmic Judge Assignment in a High-Stakes Startup Competition.”
arXiv preprint arXiv:2510.12692 (2025). https://arxiv.org/abs/2510.12692.