Evaluating LLM-as-a-Judge Beyond Score Alignment: A Psychometric Analysis of Residual Judging Difficulty

Longwei Cong[1], Sonja Hahn[1], Sebastian Gombert[1], Leon Camus[1], Fabian Zehner[1,2], Hendrik Drachsler[1,3], Ulf Kroehne[1,4]
[1]DIPF | Leibniz Institute for Research and Information in Education
[2]Centre for International Student Assessment (ZIB)
[3]Faculty of Computer Science, Goethe University Frankfurt
[4]Chemnitz University of Technology
{l.cong,s.hahn,s.gombert,l.camus,f.zehner,h.drachsler,u.kroehne}@dipf.de

Abstract

Large language models (LLMs) are widely used as automatic judges, with validity typically assessed via alignment with human scores. However, aggregate agreement fails to reveal whether humans and LLMs find the same evaluation cases difficult. In this paper, we study this problem in summarization evaluation from a psychometric perspective. We fit Many-Facet Rasch Models separately to human and LLM ratings to decompose scores into latent summary quality, rater severity, dimension severity, and rating-scale thresholds. Building on this decomposition, we define residual hardness as a model-adjusted measure of judging difficulty and compare whether human and LLM judges share the same hardness structure.

Across 17 open-weight LLM judges on SummEval, we find that moderate alignment in latent summary quality does not imply alignment in residual hardness. Human and LLM judges differ in which summary–dimension units remain difficult, and this mismatch is strongly dimension-dependent. Consistency shows a pronounced LLM-hard shift, whereas coherence shows a human-hard shift. We further show that human-easy but LLM-hard cases are partially predictable from observable source–summary properties. These findings suggest that aggregate human alignment reflects only part of LLM-as-a-judge reliability, while psychometric residual diagnostics support more informative judge evaluation and more targeted human–LLM collaboration.

1 Introduction

Large language models (LLMs) have attracted substantial attention for their strong performance across a wide range of natural language processing tasks (Naveed et al., 2025). Their advanced language understanding and reasoning capabilities have also enabled their use as automatic judges for complex tasks, offering a scalable alternative or complement to traditional expert-driven evaluation (Gu et al., 2025). LLM-as-a-judge methods have been applied to social intelligence evaluation (Wang et al., 2024b), multimodal evaluation (Chen et al., 2024), and domain-specific settings such as finance (Yu et al., 2024), law (Cheong et al., 2024), and education (Cong et al., 2026b). In this setting, an LLM is prompted to assign scores, rankings, labels, or pairwise preferences according to task-specific criteria or natural-language rubrics. The appeal is that LLMs could approximate aspects of human judgment while overcoming the cost, time, and scalability bottlenecks inherent in human evaluation (Zheng et al., 2023). Consequently, prior work typically validates LLM judges by measuring their alignment with human judgments, most often through correlations with human ratings or agreement with human preference labels (Gu et al., 2025).

However, agreement on average quality provides only a coarse view of judging behavior and reveals little about the difficulty structure of evaluation (Rodriguez et al., 2021; Cong et al., 2026a). Two judge groups may assign similar overall scores while struggling with different evaluation instances. This distinction is important because an automatic judge should not only reproduce aggregate human scores on benchmarks, but also behave reliably on individual cases in real-world evaluation settings (Li et al., 2025; Zhou et al., 2026). If humans and LLMs disagree about which cases are difficult to judge, aggregate alignment metrics may overstate the reliability of LLM-as-a-judge systems.

We argue that a shift toward a more measurement-oriented perspective is essential for a deeper understanding of LLM judging behavior. Rather than evaluating only whether LLM judges reproduce human scores on average, we need diagnostics that distinguish alignment in output quality from alignment in judging difficulty.

To address this gap, we draw on Item Response

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4682–4697
November 6-10, 2026 ©2026 Association for Computational Linguistics

Theory (IRT), a psychometric framework for modeling observed ratings in terms of latent traits and systematic measurement facets. Specifically, we use a rating-scale Many-Facet Rasch Model (MFRM) (Eckes, 2015) because it separates latent item quality from systematic rating effects such as rater severity, dimension-specific severity, and rating-scale thresholds. Building on this model, we define residual hardness as the average absolute residual between observed ratings and the ratings expected by the fitted MFRM. Conceptually, this metric provides a post-fit diagnostic of local judging instability after accounting for the systematic facets of the evaluation process. It therefore distinguishes low estimated summary quality from large residual deviations in the rating process, without treating residual hardness as an intrinsic latent measure of item difficulty.

We study this diagnostic in a SummEval case study of summarization evaluation, a setting in which each system summary is rated along multiple dimensions, such as coherence, consistency, fluency, and relevance, that place different demands on judges. This makes summarization a useful testbed for analyzing whether human–LLM alignment extends beyond aggregate quality scores to the structure of judging difficulty (Fabbri et al., 2021; Liu et al., 2023).

Our analysis is guided by three research questions.

RQ1 Does alignment between human raters and LLM judges in latent summary quality also imply alignment in residual hardness?

RQ2 How does residual-hardness mismatch vary across evaluation dimensions?

RQ3 Can textual features predict which summaries are easy for humans but difficult for LLMs to evaluate?

These questions are practically important for human–LLM collaboration scenarios. Cases that are likely to be human-easy but LLM-hard can be routed to human review, additional source-grounding checks, or stronger judge models (Bondi et al., 2022; Wang et al., 2024c). A difficulty-aware view therefore supports more targeted human–LLM collaboration in evaluation.

The contributions of this paper are as follows:

- We propose a psychometrically grounded diagnostic that separates latent instance quality from residual judging difficulty, providing a more fine-grained view of LLM-as-a-judge behavior.
- We show that human–LLM alignment in latent instance quality does not imply alignment in residual hardness, revealing substantial differences in the residual-hardness patterns of human and LLM judge panels.
- We demonstrate that human-easy but LLM-hard cases are partially predictable from observable summary-, source-, and system-level features, supporting difficulty-aware human–LLM evaluation pipelines.

2 Background

2.1 LLM-as-a-judge and Human Alignment

As complex evaluation tasks become increasingly difficult to assess with fixed references or surface-form metrics, recent work has turned to large language models as automatic judges (Gu et al., 2025). LLM judges have been used for general-purpose evaluation of model outputs, including summarization, dialogue, and instruction following (Liu et al., 2023; Fu et al., 2024; Zheng et al., 2023), as well as for data annotation (Tan et al., 2024) and domain-specific educational assessment (Gombert et al., 2026).

The validity of LLM-as-a-judge systems is typically assessed by measuring alignment with human judgments. Prior work reports correlations with human quality ratings (Liu et al., 2023; Fu et al., 2024), rank correlations between systems (Liu et al., 2023; Fu et al., 2024), agreement with pairwise human preferences (Zheng et al., 2023; Li et al., 2023; Wang et al., 2024d), or agreement with aggregated human labels (Bavaresco et al., 2025). These studies show that LLMs can serve as useful evaluators in some settings, but also that their reliability varies substantially across datasets, evaluated properties, judge models, and forms of human annotation (Bavaresco et al., 2025).

This literature establishes human alignment as a central criterion for evaluating LLM judges. However, most existing alignment metrics focus on whether LLMs reproduce human scores, rankings, preferences, or labels. They provide less information about whether human and LLM judges share the same structure of judging difficulty. A judge may correlate with humans in aggregate quality

assessment while still diverging on which evaluation instances are unstable, ambiguous, or difficult to rate. This distinction is important for model training (Yu et al., 2025), benchmark construction (Zhou et al., 2026), and human–AI collaboration (Pan et al., 2024), where the reliability of individual judgments matters as much as aggregate agreement. We therefore adopt a psychometric view of LLM-as-a-judge alignment that separates latent quality and systematic rating facets from post-fit local residual variation.

2.2 Item Response Theory

Item Response Theory (IRT) provides a measurement framework for modeling observed responses in terms of latent respondent characteristics and item properties (van der Linden, 2016). The Rasch model is a foundational IRT model in which the probability of a correct response depends on the difference between a respondent's latent ability and an item's difficulty:

$$P(y_{mn} = 1 \mid \theta_m, b_n) = \sigma(\theta_m - b_n),$$

where $y_{mn} \in \{0, 1\}$ denotes the observed response of respondent m to item n, θ_m denotes the latent ability of respondent m, b_n denotes the difficulty of item n, and $\sigma(\cdot)$ is the logistic sigmoid.

The Many-Facet Rasch Model (MFRM) (Eckes, 2015) extends the Rasch model by incorporating additional facets that may systematically influence observed scores. It allows the measurement model to include further sources of variation, such as rater severity, task difficulty, scoring dimensions, or category thresholds.

Recent work has begun to apply IRT to LLM-as-a-judge evaluation in several ways (Ye et al., 2026). Choi et al. use a graded-response IRT model to diagnose LLM-as-a-judge reliability, formalizing reliability in terms of prompt-level intrinsic consistency and alignment with human quality assessments (Choi et al., 2026). Recent educational assessment work has also applied many-facet Rasch measurement to evaluate whether LLM raters exhibit acceptable severity, consistency, and rater effects in automated scoring (Jiao et al., 2025; Wang et al., 2025). These studies establish psychometric models as useful tools for assessing LLM raters. However, they primarily ask whether LLM raters are reliable and human-aligned at the score or rater level. In contrast, we analyze whether human annotators and LLM judges exhibit similar post-fit residual-hardness patterns.

3 Method

We first obtain ratings from a panel of LLM judges on the summaries and dimensions rated by human annotators. We then analyze these ratings with MFRM, modeling summary quality, rater severity, dimension effects, and rating-scale thresholds. Based on the fitted model, we compare human and LLM judges in latent measurement space and introduce a residual-based measure of judging hardness.

3.1 Dataset and LLMs

We use SummEval (Fabbri et al., 2021), a multi-rater benchmark for evaluating generated summaries. SummEval was constructed in the context of news summarization evaluation and contains summaries generated for source articles from the CNN/DailyMail dataset. The human-annotated subset used in our study consists of 1,600 system summaries, obtained from 100 source documents and 16 summarization systems. Each evaluated unit corresponds to a system summary for a source document. The dataset includes summaries produced by both extractive and abstractive summarization systems. Human ratings cover four evaluation dimensions: coherence, consistency, fluency, and relevance. The original annotations include three expert annotators and five crowd annotators, with anonymized rater IDs. The primary human MFRM uses all eight human raters, comprising the three expert annotators and five crowd annotators.

To construct LLM-as-a-judge annotations, we prompt 17 open-weight language models to rate the same summary–dimension units on the same 1–5 scale. The models span multiple developers and model families, including Google (Team et al., 2025), Qwen (Qwen Team, 2026), Microsoft (Abdin et al., 2024b,a), Meta-Llama (Grattafiori et al., 2024), MistralAI (Liu et al., 2026), and OpenChat (Wang et al., 2024a), with parameter sizes ranging from 0.8B to 14B. We choose this panel to obtain a diverse, reproducible, and locally runnable set of LLM judges. Decoding is performed greedily with `do_sample=False` and temperature set to 0. Under this setting, token sampling is disabled, so changing the sampling seed does not produce independent generations. Details of the model set and the prompt are provided in Appendix A.1.

3.2 Many-Facet Rasch Measurement

We fit rating-scale MFRM models separately for human raters and LLM raters. For a rating y_{rid}

assigned by rater r to summary i on dimension d, the adjacent-category logit is defined as:

$$\log \frac{P(y_{rid} = c+1)}{P(y_{rid} = c)} = \theta_i - \rho_r - \delta_d - \tau_c, \quad c = 1, \ldots, 4. \tag{1}$$

Here, θ_i denotes the latent quality of summary i, ρ_r denotes the severity of rater r, δ_d denotes the severity associated with evaluation dimension d, and τ_c denotes the threshold between adjacent rating categories c and $c+1$. Higher θ_i indicates higher estimated summary quality. Higher ρ_r indicates a stricter rater, higher δ_d indicates a dimension on which high scores are harder to obtain after controlling for summary quality and rater severity, and higher τ_c indicates a more difficult transition to the next higher rating category. Details of the estimation procedure are provided in Appendix A.2.1.

We assess the adequacy of the MFRM estimation using standard psychometric diagnostics. First, we examine standardized residuals and report MAE and RMSE as descriptive summaries of rating-level prediction error. Second, we report item separation reliability, which indicates whether the calibrated measures distinguish summaries with sufficient precision relative to measurement error. We additionally verify that all five response categories are used by both human and LLM raters.

3.3 Residual Hardness

Classical agreement metrics measure whether two groups assign similar scores, but they do not directly capture whether the same items exhibit larger local deviations after accounting for the main structure of the rating process. We use the term residual hardness as a post-fit diagnostic of local judging instability. In the MFRM, expected ratings are determined by summary quality, rater severity, dimension severity, and rating-scale thresholds. If ratings for a summary–dimension unit remain far from their expected values after these facets have been accounted for, the unit exhibits high residual hardness.

For each observed rating, the fitted MFRM yields an expected rating:

$$\hat{y}_{rid} = E[y_{rid} \mid \hat{\theta}_i, \hat{\rho}_r, \hat{\delta}_d, \hat{\tau}]. \tag{2}$$

The response residual is:

$$e_{rid} = y_{rid} - \hat{y}_{rid}. \tag{3}$$

For a summary-dimension unit (i, d), residual hardness is defined as the mean absolute residual over raters in group g:

$$H_{id}^{(g)} = \frac{1}{|R_g|} \sum_{r \in R_g} |e_{rid}|. \tag{4}$$

Residual hardness is not low summary quality or a conventional latent item-difficulty parameter. Rather, it summarizes the magnitude of local deviations from the fitted MFRM after accounting for the modeled facets. We therefore interpret it as a post-fit diagnostic of local judging instability rather than as an intrinsic latent measure of item difficulty, ambiguity, or uncertainty. Larger values may reflect local rater disagreement, model misfit, or remaining unmodeled structure. The human and LLM analyses use the same MFRM form but separately estimated parameters, and we compare whether the resulting panels exhibit similar residual-hardness patterns. Appendix A.2.2 relates residual hardness to local rater agreement, and Appendix A.2.6 examines sensitivity to the use of separate calibrations.

3.4 Predicting Human-Easy and LLM-Hard Cases

To examine whether human–LLM mismatch is predictable from observable features, we formulate a binary classification task at the summary–dimension level. The positive class consists of units in the human-easy/LLM-hard quadrant, and all other units are treated as negative. We select features that capture observable properties of the source–summary pair. TF–IDF features capture lexical and topical content, length and compression features reflect information density and omission, and sentence counts provide coarse structural cues. Digit and negation counts approximate factual-detail and verification demands, while specific system IDs are excluded to avoid label leakage. We train class-balanced logistic regression models using these features. A detailed list of input features is provided in Appendix A.3.1.

We evaluate both a pooled model over all dimensions and dimension-specific models using 5-fold stratified cross-validation. In each fold, 80% of the data are used for training and 20% for testing while preserving the class ratio. Performance is reported using AUROC and AUPRC. Because the task is imbalanced, we interpret AUPRC relative to the positive-class rate.

Calib.	MAE	RMSE	Res. M	Res. SD	Item rel.
Human	0.712	0.892	0.005	1.015	0.859
LLM	0.642	0.831	0.012	1.031	0.948

Table 1: MFRM fit diagnostics. MAE and RMSE are descriptive summaries of rating-level prediction error. Res. M and Res. SD denote the mean and standard deviation of standardized residuals. Item rel. denotes item separation reliability.

4 Results

4.1 RQ1

Model adequacy. Before comparing human and LLM judging structures, we verify that the MFRM estimations provide usable measurement models. Table 1 summarizes the main fit diagnostics, including MAE and RMSE as descriptive summaries of rating-level prediction error, mean and standard deviation of standardized residuals, and item separation reliability. Across both estimations, standardized residuals are centered near zero and have standard deviations close to one. Item separation reliability is high in both estimations, indicating that the model distinguishes latent summary quality with sufficient precision. These diagnostics support using the fitted MFRM estimates to compare latent quality and residual hardness across human and LLM judge groups. All five response categories are used by both human and LLM raters, although their usage distributions differ.

Quality alignment does not imply hardness alignment. At the summary level, LLM judges show moderate alignment with human latent summary quality. The correlation between human- and LLM MFRM quality estimates is Pearson $r = 0.372$ and Spearman $\rho = 0.371$ over 1,600 summaries. This indicates that LLM judges capture part of the human ranking of summary quality.

However, this alignment does not transfer to residual hardness. When residual hardness is aggregated at the same summary level, the human–LLM correlation drops to $r = 0.283$ and $\rho = 0.176$. The gap is even sharper at the summary-dimension level, where each point corresponds to a summary evaluated on one dimension. There, human and LLM residual hardness correlate only weakly, with $r = 0.145$ and $\rho = 0.089$ over 6,400 summary-dimension units. This pattern is also visible in Figure 1. The quality estimates form a clear positive trend, whereas the hardness estimates are substantially more diffuse, especially at the summary–dimension level.

Robustness to judge capability and human-panel composition. The weak residual-hardness alignment is not driven solely by the smallest LLM judges. When we refit the LLM MFRM using only the seven judges with at least 7B parameters, quality alignment remains moderate ($\rho = 0.408$), whereas summary–dimension residual-hardness alignment remains near zero ($\rho = 0.039$). We additionally refit the human MFRM using only the three expert annotators and compare it with the independently fitted $\geq$ 7B LLM panel. In this setting, quality alignment is substantially stronger ($\rho = 0.688$), while residual-hardness alignment remains weak ($\rho = 0.092$). These analyses indicate that the contrast between quality alignment and residual-hardness alignment is not explained solely by the smallest LLM judges or by pooling expert and crowd human ratings.

These results answer RQ1 negatively. LLMs partially align with humans in latent summary quality, but this quality alignment does not carry over to the structure of residual hardness. In other words, LLMs can rank summaries in a moderately human-aligned way while exhibiting substantially different residual-hardness patterns.

4.2 RQ2

We next examine whether the weak overall residual-hardness alignment is accompanied by systematic differences across evaluation dimensions. We classified each summary–dimension unit as easy or hard within each judge group by thresholding residual hardness at the group's overall median. This median split is an operational definition rather than a theoretically privileged cutoff. Appendix A.3.2 examines the sensitivity of the downstream RQ3 prediction analysis to a stricter quartile-based quadrant definition.

Table 2 reports the percentage of units classified as hard for each judge group. Consistency shows the largest LLM-hard shift, with 74.7% of units classified as LLM-hard compared with 50.8% human-hard. Coherence shows the opposite pattern, with 57.3% human-hard but only 34.8% LLM-hard. Fluency and relevance are closer to balanced, with much smaller hard-rate gaps. This structure is also visible in Figure 2, where consistency shows a pronounced concentration of points in the human-easy/LLM-hard region, whereas coherence shows the opposite tendency. This pattern also persists un-

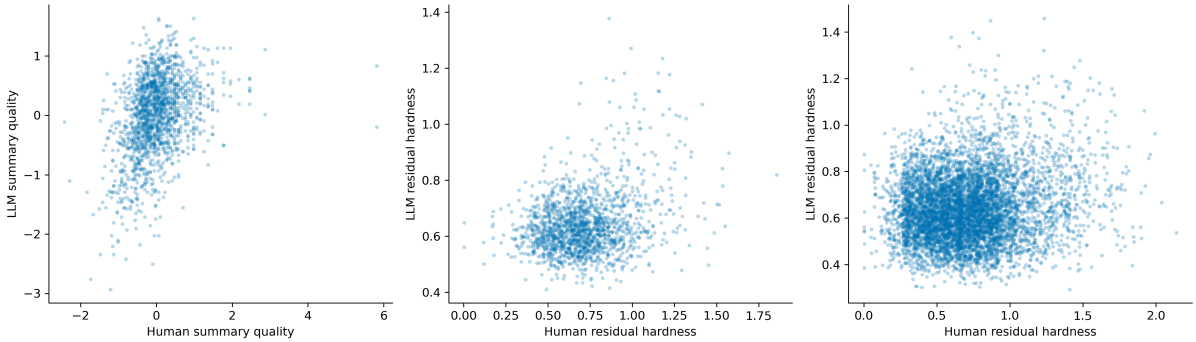


Figure 1: Human–LLM alignment in latent summary quality (left), summary-level residual hardness (middle), and summary–dimension residual hardness (right).

Dimension	Human	LLM	Gap
Consistency	50.8	74.7	+23.9
Coherence	57.3	34.8	-22.5
Relevance	43.0	41.6	-1.4
Fluency	48.9	48.9	-0.1

Table 2: Dimension-level residual hardness gaps. Values are percentages of summary–dimension units classified as hard within each judge group. Gap is LLM minus human hard-rate.

System type	Human	LLM	Gap
Extractive	38.9	51.8	+12.9
Abstractive	53.7	49.4	-4.3

Table 3: System-level residual hardness gaps. Values are percentages of summary–dimension units classified as hard. Gap is LLM minus human hard-rate.

der the expert-only versus $\geq$ 7B LLM comparison, where a $+26.2$ percentage point consistency gap remains.

At the system-type level, the mismatch is weaker than at the dimension level. As reported in Table 3, extractive systems show a modest LLM-hard shift, whereas abstractive systems show a small human-hard shift.

We additionally examined rater-level fit and misfit sensitivity in Appendix A.2.5. The main residual-hardness findings remain stable after excluding raters flagged by descriptive infit or outfit criteria. To assess whether the conclusions depend on direct comparability of raw residual magnitudes across separate calibrations, we also report normalization- and calibration-based sensitivity checks in Appendix A.2.6. The weak residual-hardness alignment and the dominant consistency/coherence contrast remain stable across these alternative specifications.

Task	N	Pos.	AUROC	AUPRC
All dimensions	6,400	1,494	0.631	0.341
Relevance	1,600	356	0.717	0.442
Consistency	1,600	571	0.633	0.465
Fluency	1,600	374	0.631	0.353
Coherence	1,600	193	0.621	0.197

Table 4: Prediction of human-easy/LLM-hard cases from all summary-dimension units. The positive class is the human-easy/LLM-hard quadrant; all other quadrants are negative.

4.3 RQ3

For RQ3, we analyzed whether this mismatch can be predicted directly from observable properties of the source document and summary. Table 4 shows that human-easy/LLM-hard cases are partially predictable from observable features. The pooled all-dimension model reaches AUROC 0.631 and AUPRC 0.341, compared with a positive rate of 23.3%. Relevance is the most predictable task, with AUROC 0.717, followed by consistency, fluency, and coherence. These results suggest that the human-easy/LLM-hard quadrant is not arbitrary. It is associated with observable properties of the evaluated summary and source document. This predictive signal is robust to feature ablations and stricter grouped split strategies. Additional ablations, confidence intervals, grouped-split robustness checks and quadrant-definition sensitivity checks are reported in Appendix A.3.2.

Figure 3 interprets the dimension-specific classifiers using signed SHAP values (Lundberg and Lee, 2017). Because several surface features are correlated, particularly source length, summary length, and compression ratio, we interpret the SHAP values descriptively as model associations rather than as independent or causal feature ef-

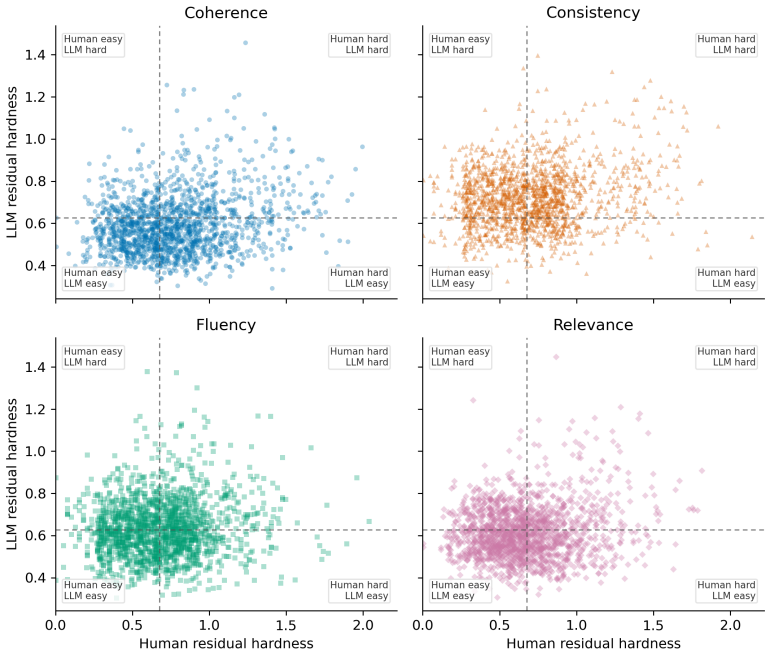


Figure 2: Human–LLM residual hardness by evaluation dimension. Dashed lines indicate the overall median residual hardness within each judge group, partitioning the space into human-easy/LLM-easy, human-easy/LLM-hard, human-hard/LLM-easy, and human-hard/LLM-hard regions.

fects. For relevance, higher compression ratio is associated with SHAP contributions toward human-easy/LLM-hard classification, whereas longer summaries tend to contribute in the opposite direction. Coherence shows a similar pattern, where longer source documents and higher compression ratios are associated with contributions toward the human-easy/LLM-hard quadrant, while longer summaries tend to contribute in the opposite direction.

Consistency exhibits a different profile. Its most influential predictors include digit count, sentence count, source length, negation count, and coarse system type, indicating that mismatch in this dimension is tied more strongly to factual-detail and verification-related cues than to compression alone. Fluency shows a weaker but still interpretable pattern, in which longer summaries contribute most strongly to human-easy/LLM-hard classification, with smaller contributions from system type, digit count, and compression ratio.

5 Discussion

Our results show that human alignment in LLM-as-judge evaluation is not a single property. Although LLM judges moderately align with human latent summary quality, this agreement does not imply that they share the same residual hardness structure. The weak correlation between human and LLM residual hardness indicates that the cases that remain difficult after controlling for summary quality, rater severity, dimension severity, and rating thresholds are only partially overlapping across judge groups. This distinction matters because aggregate quality alignment can make an evaluator appear reliable while masking item-level instability (Lalor et al., 2016, 2018).

The mismatch structure appears at the dimension level. Consistency shows a large LLM-hard shift, whereas coherence shows the opposite human-hard shift. Consistency judgments require checking whether summary claims are supported by the source document (Fabbri et al., 2021; Kryscinski et al., 2020; Wang et al., 2020), which may impose a factual-verification burden on LLM judges. In contrast, coherence judgments depend more on discourse organization, readability, and holistic interpretation (Zhao et al., 2023; Barzilay and Lapata, 2005; Fabbri et al., 2021), where human annotators may apply more nuanced or variable standards. Fluency and relevance show smaller gaps, suggesting that the divergence is not a general failure of the LLM judges but a dimension-specific difference in residual judging behavior.

The prediction results further indicate that part of the human–LLM residual-hardness mismatch is associated with observable source–summary prop-

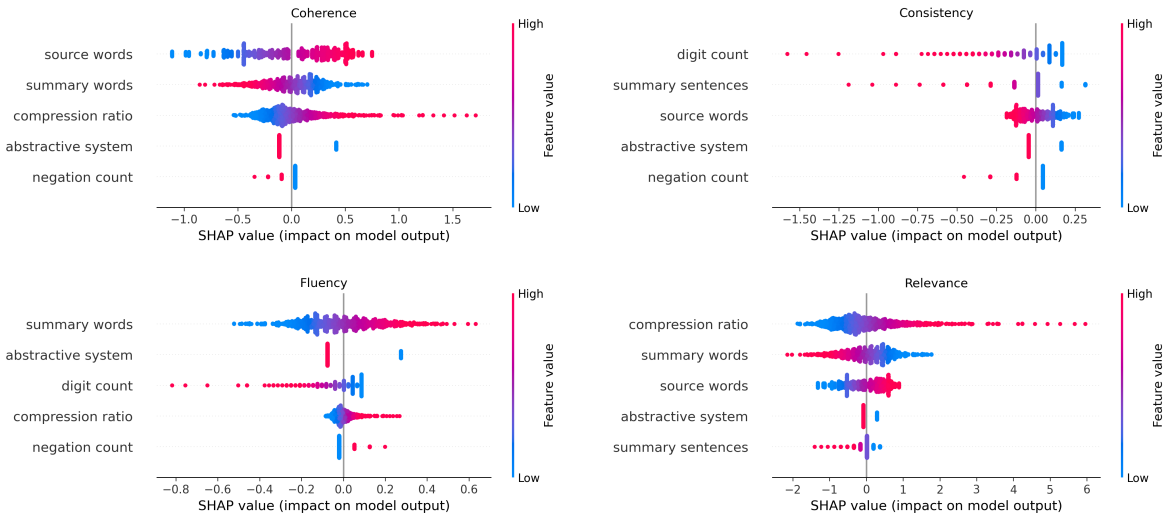


Figure 3: SHAP beeswarm plots for human-easy/LLM-hard classification by evaluation dimension. Positive SHAP values indicate features that push predictions toward the human-easy/LLM-hard quadrant, while negative values indicate the opposite. Color denotes feature value from low to high.

erties, consistent with prior evidence that LLM-as-judge behavior exhibits persistent biases and dimension-dependent unreliability (Ye et al., 2025; Shen et al., 2023). In our setting, the class-balanced logistic regression model predicts human-easy/LLM-hard cases above the positive-class baseline, providing modest evidence that part of the observed mismatch is associated with observable source–summary properties. SHAP analysis reveals two feature profiles. For relevance and coherence, compression-related features dominate, providing modest evidence that some of the observed mismatch is predictable from observable source–summary properties. For consistency, digit count, sentence count, source length, negation count, and system type point to factual-detail and verification-related cues, consistent with prior work framing factual consistency evaluation as verification-intensive (Chen et al., 2023; Gabriel et al., 2021). Overall, part of the residual-hardness mismatch is measurable and partially predictable from observable text properties.

These findings have methodological implications for evaluating LLM judges. Existing validation practices (Liu et al., 2023; Gu et al., 2025) often treat human alignment as a score-level property, measured through correlations, rank agreement, or preference agreement. Our results show that such metrics are incomplete. Two judge groups may agree on the latent quality ordering of outputs while disagreeing on which individual cases are difficult to judge. Psychometric residual diagnostics therefore provide a complementary view of evaluator reliability. In particular, residual hardness and dimension-specific hardness gaps can reveal failure modes that are invisible to aggregate agreement metrics.

Practically, this suggests that LLM judges should not be deployed as uniform replacements for human annotators. Their reliability depends on the evaluation dimension and on properties of the evaluated text. A difficulty-aware evaluation pipeline could use such diagnostics to route cases selectively (Bondi et al., 2022; Wang et al., 2024c). Cases predicted to be human-easy but LLM-hard may be prioritized for human review, additional source-grounding checks, or stronger judge models. In this sense, these diagnostics may support more targeted human–LLM collaboration rather than simply accepting or rejecting LLM judges at the aggregate level.

6 Conclusion

We presented a psychometric analysis of human–LLM alignment in summarization evaluation. Moving beyond aggregate agreement metrics, we asked whether human and LLM judges share the same structure of residual judging difficulty after controlling for latent summary quality, rater severity, dimension severity, and rating-scale thresholds. Using a Many-Facet Rasch Model, we showed that moderate human–LLM alignment in latent summary quality does not imply alignment in residual hardness. In particular, human and LLM judges

differ substantially in their summary–dimension residual-hardness patterns, with especially pronounced mismatch for consistency and coherence.

We further showed that human-easy but LLM-hard cases are partially predictable from observable properties of the source–summary pair, with different predictive patterns across evaluation dimensions. These findings suggest that score-level human alignment provides only a partial view of LLM-as-judge reliability. Psychometric residual diagnostics can reveal failure modes that remain hidden under aggregate agreement metrics.

More broadly, our results argue for a difficulty-aware view of LLM-as-judge evaluation. Rather than treating LLM judges as uniform replacements for human annotators, future evaluation pipelines should account for which cases are likely to remain difficult for LLMs even when overall score alignment appears acceptable. This perspective can support more targeted human–LLM collaboration and more informative evaluation of automatic judges. The empirical findings are specific to the SummEval setting studied here, and establishing whether similar patterns generalize across tasks, domains, and stronger judge populations remains an important direction for future work.

Limitations

Our study has several limitations. First, we examine residual hardness mismatch in summarization evaluation using a single multi-rater benchmark, which limits the generalizability of the findings to other LLM-as-judge settings. However, this benchmark is well suited to our research question because it provides repeated ratings across multiple evaluation dimensions and includes anonymized rater identifiers required for the MFRM. More broadly, public summarization benchmarks with repeated human ratings, identifiable raters, and summary-level multidimensional quality scores are extremely limited, and SummEval is, to our knowledge, the most suitable publicly available benchmark for this analysis.

Second, our results are based on a fixed panel of open-weight LLM judges under a single prompting setup. Different model families, prompting strategies, or evaluation protocols may yield different residual hardness structures.

Third, residual hardness should not be interpreted as a standalone latent construct of judging difficulty or ambiguity. Rather, it is a post-fit diagnostic that operationalizes local judging difficulty through residual deviations from the fitted MFRM. Its magnitude may reflect local rater disagreement, model misspecification, or remaining structure not captured by the model. Accordingly, our analyses establish differences in residual-hardness patterns between the evaluated human and LLM panels, but do not assume that these residuals have identical substantive meaning across panels.

Fourth, the RQ3 prediction analysis identifies correlates of human–LLM mismatch from observable features, but does not establish the causal mechanisms underlying these patterns.

Acknowledgments

This research was conducted within the project "Assessment for Learning with AI (ALwAI)" funded by the Leibniz Association under the Leibniz Competition (project no. T163/2024).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024b. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2022. Role of human-ai interaction in selective prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5286–5294.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *Preprint*, arXiv:2305.14069.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2454–2469, New York, NY, USA. Association for Computing Machinery.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. 2026. Diagnosing the reliability of llm-as-a-judge via item response theory. *arXiv preprint arXiv:2602.00521*.

Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachsler, and Ulf Kroehne. 2026a. Estimating LLM grading ability and response difficulty in automatic short answer grading via item response theory. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 259–271, San Diego, California, USA. Association for Computational Linguistics.

Longwei Cong, Leon Hammerla, Sonja Hahn, Sebastian Gombert, Hendrik Drachsler, and Ulf Kroehne. 2026b. Automatic short answer grading with LLMs: From memorization to reasoning. In *Proceedings of the 16th International Learning Analytics and Knowledge Conference*, New York, NY, USA. Association for Computing Machinery.

Thomas Eckes. 2015. *Introduction to Many-Facet Rasch Measurement*. Peter Lang Verlag, Berlin, Deutschland.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Marcus Kubsch, Daniele Di Mitri, Knut Neumann, and Hendrik Drachsler. 2026. Are rubrics all you need? towards rubric-based automatic short answer scoring via guided rubric-answer alignment. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, LAK '26, page 272–282, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Hong Jiao, Dan Song, and Won-Chan Lee. 2025. Comparing human and ai rater effects using the many-facet rasch model. *Preprint*, arXiv:2505.18486.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing, pages 648–657, Austin, Texas. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

Qian Pan, Zahra Ashktorab, Michael Desmond, Martín Santillán Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for LLM-as-a-judge. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29. ACL.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

W.J. van der Linden, editor. 2016. *Handbook of Item Response Theory: Volume 1: Models*. Behavioral Sciences. Chapman and Hall/CRC.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. In *International Conference on Learning Representations*, volume 2024, pages 57021–57040.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024b. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024c. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024d. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations*, volume 2024, pages 43573–43593.

Yuehan Wang, Jinyan Huang, Lun Du, Yuxin Guo, Ying Liu, and Rong Wang. 2025. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers and Education: Artificial Intelligence*, 9:100481.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2026. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *Preprint*, arXiv:2505.08245.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025. Improve LLM-as-a-judge ability as a general ability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14099–14115, Suzhou, China. Association for Computational Linguistics.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: a synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press.

A Appendix

A.1 LLM Judge Prompt

We use the same task instruction for all open-weight LLM judges. Table 5 lists the 17 open-weight LLM judges used in our main analysis. For models whose tokenizer supports a system role, the prompt is formatted as a chat conversation using the model's native chat template. If the model does not support a system role, the system instruction is prepended to the user message. Decoding is deterministic.

> **System:** You are a careful evaluator of news summaries.
>
> **User:** Rate the summary for the requested evaluation dimension.
>
> Use a 1–5 integer scale where 1 is very poor and 5 is excellent.
>
> Dimension: {*dimension*}
>
> Criterion: {*rubric*}
>
> Source document:
>
> {*source document*}
>
> Summary:
>
> {*summary*}
>
> Return ONLY one number: 1, 2, 3, 4, or 5.
>
> Answer:

The dimension field is one of coherence, consistency, fluency, or relevance. The criterion field contains the corresponding SummEval rubric. We parse the first valid integer in $\{1, \ldots, 5\}$ from the model output. If no valid score is found, we retry once with a longer generation budget and apply the same parser. All prompts included the full source document without explicit truncation. As a length check, the longest rendered prompt contained 694 whitespace-delimited words, with maximum source and summary lengths of 540 and 133

words, respectively. These lengths are far below the context limits of the local judges.

Model
Qwen/Qwen3.5-0.8B
Qwen/Qwen3.5-2B
Qwen/Qwen3.5-4B
Qwen/Qwen3.5-9B
microsoft/Phi-4-mini-instruct
microsoft/Phi-3-mini-4k-instruct
microsoft/Phi-3-medium-4k-instruct
meta-llama/Llama-3.2-1B-Instruct
meta-llama/Llama-3.2-3B-Instruct
meta-llama/Llama-3.1-8B-Instruct
google/gemma-3-1b-it
google/gemma-3-4b-it
google/gemma-3-12b-it
mistralai/Ministral-3-3B-Instruct-2512
mistralai/Ministral-3-8B-Instruct-2512
mistralai/Ministral-3-14B-Instruct-2512
openchat/openchat-3.5-0106

Table 5: 17 open-weight LLM judges used for SummEval annotation in our analysis.

A.2 MFRM

A.2.1 MFRM Estimation Details

We estimate the MFRM parameters by maximum likelihood. For each observed rater–summary–dimension rating, we compute $\eta_{rid} = \theta_i - \rho_r - \delta_d$, construct the five category logits implied by the adjacent-category formulation, and normalize them with a softmax. The objective is the negative log-likelihood of the observed rating categories with weak L_2 regularization. We optimize this objective in PyTorch using AdamW with learning rate 0.05 and weight decay 10^{-4}. For identification, summary, rater, and dimension parameters are mean-centered during optimization. The step thresholds are constrained to be ordered using an ordered parameterization and are initialized as evenly spaced values from -1.5 to 1.5. All other facet parameters are initialized at zero. Optimization uses early stopping when the loss does not improve by 10^{-7} for 300 epochs.

A.2.2 Agreement-based Sanity Check for Residual Hardness

As a sanity check, we examine whether residual hardness covaries with local rater agreement. For each summary–dimension unit i, we compute the observed pairwise agreement

$$P_i = \frac{\sum_c n_{ic}(n_{ic} - 1)}{n_i(n_i - 1)}, \tag{5}$$

Group	Dimension	Exact κ	Weighted κ
Human	Coherence	-0.547	-0.854
Human	Consistency	-0.481	-0.865
Human	Fluency	-0.434	-0.855
Human	Relevance	-0.645	-0.828
LLM	Coherence	-0.401	-0.416
LLM	Consistency	-0.374	-0.615
LLM	Fluency	-0.339	-0.373
LLM	Relevance	-0.385	-0.494

Table 6: Spearman correlations between residual hardness and summary–dimension-level agreement contribution. Negative correlations indicate that units with higher residual hardness contribute less to rater agreement.

where n_{ic} denotes the number of raters assigning category c to unit i, and n_i is the total number of ratings for that unit. We then define the local chance-corrected agreement contribution as

$$\kappa_i = \frac{P_i - P_e}{1 - P_e}, \tag{6}$$

where P_e is estimated separately for the corresponding judge panel and evaluation dimension. This quantity is used as a local agreement diagnostic rather than as the standard global kappa statistic.

Residual hardness is negatively associated with the local agreement diagnostics across all dimensions in both judge panels. This relationship is intended only as a sanity check showing that larger residual deviations tend to coincide with greater local rater disagreement. It does not independently validate residual hardness as a distinct construct or establish that residual hardness has identical substantive meaning across the human and LLM panels. Indeed, the associations are substantially stronger for human ratings than for LLM ratings. Unlike the agreement diagnostic, residual hardness is conditional on the fitted MFRM facets.

A.2.3 Judge-Capability and Human-Panel Sensitivity

Panel comparison	Quality ρ	Residual ρ
All human vs. all LLM	0.371	0.089
All human vs. $\geq$ 7B LLM	0.408	0.039
Experts vs. $\geq$ 7B LLM	0.688	0.092

Table 7: Sensitivity of human–LLM alignment to LLM judge capability and human-panel composition. Residual-hardness correlations are computed at the summary–dimension level.

Restricting the LLM panel to the seven judges with at least 7B parameters does not remove the

contrast between quality alignment and residual-hardness alignment. The contrast is also preserved when the human MFRM is estimated using only the three expert annotators. In the expert-only versus $\geq$ 7B comparison, the consistency hard-rate gap remains $+26.2$ percentage points. These results indicate that the main RQ1 and RQ2 patterns are not driven solely by the smallest LLM judges or by expert–crowd disagreement in the human panel.

A.2.4 Rater-by-Dimension Interaction Sensitivity

To examine whether the weak human–LLM residual-hardness alignment primarily reflects an omitted rater-by-dimension interaction, we refit the MFRM with an additional rater–dimension term. Under this richer specification, human–LLM residual-hardness alignment remains weak ($r = 0.277$, $\rho = 0.122$). Thus, the main RQ1 pattern is not solely attributable to omission of rater-specific dimension effects.

A.2.5 Rater-Misfit Sensitivity

To examine whether the main residual-hardness findings are driven by poorly fitting raters, we conducted a post-hoc sensitivity analysis after excluding raters with infit or outfit mean-square outside the conventional descriptive range of 0.5–1.5. The original MFRM calibrations were kept fixed, and residual-hardness summaries were recomputed using only the remaining raters.

Table 8 shows that the central RQ1 conclusion remains unchanged. Human–LLM alignment in latent quality remains moderate, whereas residual-hardness alignment remains weak at both the summary level and the summary–dimension level. If anything, excluding flagged raters further reduces the residual-hardness correlations, suggesting that the weak human–LLM hardness correspondence is not an artifact of poorly fitting raters.

Outcome	Stat.	All raters	Exclude flagged
Quality	r	0.372	0.372
	ρ	0.371	0.371
Summary hardness	r	0.283	0.222
	ρ	0.176	0.157
Dim. hardness	r	0.145	0.108
	ρ	0.089	0.073

Table 8: RQ1 sensitivity after excluding raters with infit or outfit mean-square outside the descriptive 0.5–1.5 range.

Table 9 shows that the dimension-level pattern is also stable. Consistency remains LLM-hard relative to humans, whereas coherence remains human-hard relative to LLMs. The smaller gaps for fluency and relevance vary more after exclusion, which is expected because their original gaps are close to zero.

Dimension	All raters gap	Exclude flagged gap
Coherence	-22.5	-19.8
Consistency	23.9	18.0
Fluency	-0.1	7.3
Relevance	-1.4	-5.5

Table 9: RQ2 dimension hard-rate gap sensitivity after excluding flagged raters. Values are percentage-point gaps, computed as LLM hard-rate minus human hard-rate.

A.2.6 MFRM Calibration Comparability Sensitivity

Because the main analyses compare residual hardness derived from separately calibrated human-only and LLM-only MFRMs, we examine whether the conclusions depend on direct comparability of raw residual magnitudes across calibrations. We recompute the residual-hardness comparisons under three alternatives: within-group z-scored residual hardness, percentile-rank residual hardness, and residuals from a combined MFRM fit to pooled human and LLM ratings. For the combined-model analysis, absolute residuals are aggregated separately for human and LLM raters after fitting the shared model.

Method	Stat.	Summary	Dim.
Main	r	0.283	0.145
	ρ	0.176	0.089
z-score	r	0.283	0.145
	ρ	0.176	0.089
Rank	r	0.176	0.089
	ρ	0.176	0.089
Combined	r	0.352	0.170
	ρ	0.231	0.104

Table 10: Residual-hardness alignment under alternative calibration and normalization choices.

The main conclusion does not rely on direct comparability of raw residual magnitudes across separately calibrated models. As expected, within-group z-scoring leaves the correlations unchanged because it preserves within-group ordering. Rank-based normalization yields slightly smaller correlations, but the overall pattern remains the same. Under the combined MFRM, residual-hardness

alignment is somewhat higher, yet it remains clearly weaker than the corresponding quality alignment and remains weak in absolute terms at the summary–dimension level. The dominant dimension pattern is also stable across specifications. Consistency remains LLM-hard relative to humans, whereas coherence remains human-hard relative to LLMs.

A.3 Additional Information and Experiments for RQ3

A.3.1 Input Features for RQ3

Table 11 summarizes the input features used for the RQ3 human-easy/LLM-hard prediction experiments. We include these features as interpretable, observable proxies for factors that may affect judging difficulty. TF–IDF features capture lexical and topical content, length and compression features approximate information density and omission, and sentence, digit, and negation counts provide coarse structural and factual-detail cues. Coarse system type captures broad extractive versus abstractive differences, while specific system IDs are excluded to reduce label leakage.

Feature group	Features
Summary text	TF–IDF features from the generated summary, using unigram and bigram features with English stop-word removal.
Source text	TF–IDF features from the source document, using unigram and bigram features with English stop-word removal.
Surface features	Summary word count, source word count, compression ratio, summary sentence count, summary digit count, and summary negation count.
Coarse metadata	System type, coded as extractive or abstractive.

Table 11: Input features used for RQ3 prediction of human-easy/LLM-hard cases. TF–IDF features use `min_df` $= 2$, unigram/bigram ranges, and up to 6,000 features each for the summary and source fields.

A.3.2 RQ3 Prediction Robustness and Ablations

We run additional robustness checks for the RQ3 human-easy/LLM-hard prediction task. All analyses use the pooled summary–dimension dataset ($N = 6{,}400$, positive class $= 1{,}494$, positive rate $= 23.3\%$). As in the main analysis, we retain coarse system type but exclude specific system IDs. Confidence intervals are percentile bootstrap intervals computed from out-of-fold predictions.

Feature set	AUROC	AUPRC
Surface only	0.573 [0.555, 0.590]	0.293 [0.273, 0.314]
TF–IDF only	0.620 [0.605, 0.637]	0.326 [0.305, 0.351]
Metadata only	0.542 [0.524, 0.560]	0.260 [0.245, 0.277]
Full model	0.631 [0.614, 0.648]	0.341 [0.319, 0.365]

Table 12: RQ3 feature ablation for predicting human-easy/LLM-hard summary–dimension units. Brackets denote 95% bootstrap confidence intervals from out-of-fold predictions. Metadata includes coarse system type only; specific system IDs are excluded.

Split strategy	AUROC	AUPRC
Stratified 5-fold	0.631 [0.614, 0.648]	0.341 [0.319, 0.365]
Summary	0.618 [0.602, 0.634]	0.326 [0.306, 0.348]
Document	0.569 [0.553, 0.586]	0.289 [0.269, 0.310]
System	0.618 [0.602, 0.634]	0.329 [0.309, 0.353]

Table 13: RQ3 split robustness for the full model. Brackets denote 95% bootstrap confidence intervals from out-of-fold predictions. Grouped splits hold out all dimensions of the same summary, all summaries from the same source document, or all summaries from the same generation system, respectively.

These checks show that the RQ3 signal is not driven by a single feature family or solely by random-fold leakage across repeated summaries. TF–IDF features provide most of the predictive signal, surface features retain above-chance predictive performance, and the full model consistently improves over the ablations. Performance decreases under stricter grouped splits, especially when holding out entire source documents, but predictive signal remains in all settings. For the surface-feature model, we additionally perform a compression-drop ablation under summary-grouped 5-fold cross-validation. Removing compression ratio changes AUROC only from 0.571 to 0.566 and AUPRC from 0.293 to 0.286. Overall, these results support partial predictability of human-easy/LLM-hard cases that persists under alternative feature sets and split strategies rather than high-accuracy classification.

Task	N	Pos.	AUROC	AUPRC
All dimensions	6,400	349	0.676	0.114
Relevance	1,600	102	0.703	0.206
Consistency	1,600	87	0.672	0.133
Fluency	1,600	79	0.743	0.138
Coherence	1,600	70	0.578	0.067

Table 14: RQ3 quartile-threshold sensitivity for predicting human-easy/LLM-hard cases from all summary–dimension units.

We also test a stricter quadrant definition based on task-specific quartiles. For each prediction task, human-easy units are those whose human residual hardness falls in the lower quartile of that task's evaluation group, and LLM-hard units are those whose LLM residual hardness falls in the upper quartile of the same group. Table 14 shows that the pooled predictive signal remains above chance under this more selective definition (AUROC $= 0.676$), although dimension-specific robustness is uneven. In particular, coherence decreases to AUROC $= 0.578$. We therefore interpret this analysis as supporting pooled predictability rather than uniformly strong robustness across dimensions.