Beyond Agreement: Compositional Validity in Multidimensional LLM Evaluation

Savita Bhat
TCS Research
IIIT Hyderabad
savita.bhat@tcs.com

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

Abstract

LLM judges are increasingly used to evaluate generated text in multiple quality dimensions, where overall quality is implicitly built from fine-grained trait judgments. However, whether the evaluator combines these judgments as human evaluators do is rarely examined. We argue that this composition is a core property of multidimensional evaluation that agreement cannot capture, and formalize it as '*compositional validity*', the degree to which a judge's trait-to-overall aggregation matches that of human judges. Across summarization, dialogue, and essay evaluation, we introduce a diagnostic framework that investigates the compositional relationship between overall and trait-level scores. We find that overall judgments are highly recoverable from traits (R^2 = 0.675 to 0.945). However, this does not imply validity. Additionally, attribution methods agree with one another far more than any judge agrees with humans (Kendall's τ = 0.47 to 1.00 internally vs. -0.07 to 0.20 against humans). An alignment diagnostic, computed at full sample size, localizes much of this gap to how traits are aggregated rather than how they are scored, and targeted perturbations show that correlational alignment overstates trait-level competence. Rubric geometry, in turn, is a dominant and auditable determinant of compositional validity, explaining 17x more variance than model choice (η^2 = 0.384 vs $\approx$ 0.022). Our results suggest that reliable multidimensional LLM evaluation requires validating not just that judges agree with humans, but also that they base their judgments on the right reasons, and that rubric geometry offers a way to audit this before deployment.

1 Introduction

With powerful models and ready access, LLM judges offer a scalable alternative to multidimensional evaluation that traditionally depends on human experts. However, the majority of LLM-based evaluations are validated solely by agreement with human scores. We argue that this criterion is insufficient for assessing LLM judges' reliability across tasks and evaluation rubrics. We believe that the multidimensional evaluation implicitly rests on a simple premise that an overall judgment has some stable relationship to the fine-grained traits scored alongside it. In other words, the whole is, at least in part, a function of the parts. If this is the case, then two evaluators can broadly agree on the parts and on the overall score, yet reach that overall by weighting the parts differently. Agreement, at either level, will hide the difference in how the whole was built.

Imagine an LLM judge scoring a student essay on an overall quality and six writing traits. It rates the essay highly on grammar and vocabulary, with a strong overall score. A teacher may agree with the LLM judge's vocabulary evaluation but assign a low grammar score, yet weighing the essay's ideas more heavily, still assign a high overall score. The two overall scores still agree, here and across many essays, but they are reached by different rules, i.e., the judge's overall score rests on grammar the teacher considers weak. In such cases, which overall score should be trusted and can the trait score be used for targeted feedback? Agreement, at either level of granularity, cannot answer these questions. The premise that a holistic evaluation maintains a stable relationship with fine-grained constituent attributes is not novel. Judgment analysis models such as information integration theory (Anderson, 1971) represent human evaluative judgment as a weighted combination of attribute values, each consisting of a value and an associated weight, and lens-model studies have used linear models to decode how judges weight traits across medicine, education, and business (Kaufmann et al., 2013). Such models capture a judge's policy without making any claims about the underlying reasoning. We adopt the stance that the composition we estimate

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 895–913
October 17-21, 2026 ©2026 Association for Computational Linguistics

describes how an evaluator weights traits, not how it computes.

We formalize this gap as compositional validity: the degree to which a judge's mapping from trait scores to an overall judgment matches that of a human judge. Agreement asks whether a judge's scores are correct, whereas compositional validity asks whether the overall score is reached using the same logic as humans. Only a compositionally valid judge can be trusted to explain why it scored as it did or to be corrected when it is wrong.

To study it, we propose a diagnostic framework (Figure 1) that decomposes overall evaluations into trait contributions, and investigates, with increasing strictness, whether a judge's overall score is recoverable from its trait scores, whether the recovered composition is stable and human-aligned, and where any misalignment lies. Across summarization, dialogue, and essay evaluation, we find that overall judgments are highly recoverable from traits, but recoverability does not imply validity. LLM Judges are internally consistent across attribution methods, but exhibit no statistically detectable alignment with human trait ordering. The rubric geometry, and not the model choice, governs whether a valid composition exists at all ($\eta^2 = 0.384$ vs. ≈ 0.022) and depends on the rubric rather than the content.

Our contributions are as follows:

1. We posit multidimensional evaluation as a problem of compositional validity, in which we investigate whether a judge composes its overall judgment from traits in the same way humans do.

2. We propose a diagnostic framework that separates recoverability from compositional validity and, via an alignment diagnostic, localizes misalignment to trait scoring versus aggregation.

3. We empirically demonstrate that rubric geometry is a dominant, auditable determinant of compositional validity, explaining far more variance than model choice and depending on the rubric rather than the content.

4. We provide a practical rubric-audit workflow that consists of computing effective rank and condition number before trusting trait-level feedback, and demonstrate through counterfactual perturbation that correlational alignment overstates trait-level competence.

2 Related Work

2.1 LLM-as-a-Judge

LLM-as-a-Judge frameworks have emerged as a scalable alternative to human evaluation across generation tasks (Chiang and Lee, 2023; Yuan et al., 2023; Zheng et al., 2023), with several studies (Gu et al., 2024; Liu et al., 2023) reporting strong overall agreement with human evaluators. Previous work in (Ye et al., 2023; Huang et al., 2025; Wang et al., 2024; Turpin et al., 2023; Gallegos et al., 2024) has also identified important limitations, including hallucination, factual errors, unfaithful chain-of-thought, and systematic biases such as position, verbosity, and self-preference. However, most assess judges primarily through aggregate metrics — correlation or agreement with human scores at the overall and, increasingly, the trait level. We focus instead on the composition that agreement cannot measure.

2.2 Multidimensional Evaluation

Recent multidimensional evaluation frameworks decompose holistic quality into finer-grained criteria to improve interpretability and evaluation quality. Systems such as (Liu et al., 2023) and (Feng et al., 2025) show that structured criteria, hierarchical decomposition, or adaptive rubric construction can improve human alignment and transparency, and FLASK (Ye et al., 2024), LMUnit (Saad-Falcon et al., 2024), HD-EVAL (Liu et al., 2024), and (Yu et al., 2025) demonstrate the effectiveness of decomposed evaluation. However, these works primarily use decomposition as a design choice to improve the evaluation. Recent work examines aggregation and structural problems of multidimensional evaluation. In particular, criterion entanglement as a halo effect bias (Li et al., 2025); dependency-aware aggregation to prevent unlicensed credit in reward modeling (Lv et al., 2026); and calibrated aggregation of rubric criteria into an overall judgment (Hashemi et al., 2024). These treat entanglement as a system-specific bias to mitigate, or as a mapping to fit. In contrast, we investigate diagnostically whether, given an arbitrary rubric, its geometry supports a trustworthy trait-level evaluation at all.

2.3 Attribution and Interpretability

Attribution methods such as SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), and related interpretability tools such as token-SHAP

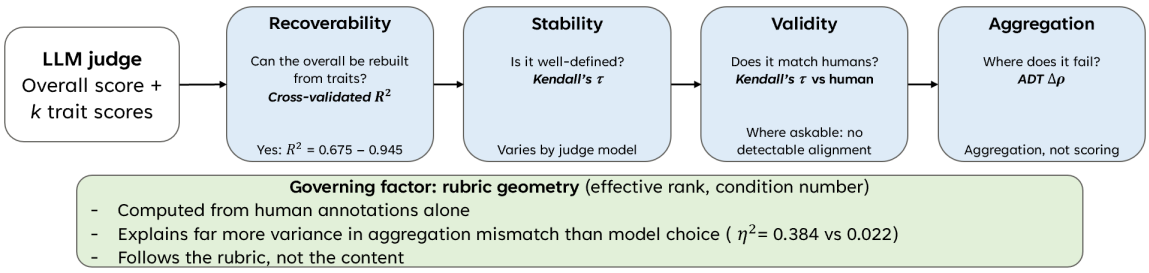


Figure 1: Each stage asks a stricter question of the judge's composition. Recoverability establishes that a composition exists; stability, whether it is well defined; validity, whether it matches the human one; and the alignment diagnostic, where it fails. Rubric geometry governs all four.

(Horovicz and Goldshmidt, 2024) and LAMP (Chen et al., 2025) identify which inputs or features drive model predictions, and attribution-informed methods have also been explored for alignment and control. We repurpose these tools with a different unit of analysis. Rather than attributing predictions to tokens or internal activations, we apply attribution at the level of evaluation traits to measure how a judge maps fine-grained rubric traits onto its overall judgment, and whether this mapping aligns with that of humans. In our setting, attribution is not the end (explaining a prediction) but the means (quantifying composition) to study rubric geometry and diagnostic value of trait-level evaluation.

3 Compositional Validity

3.1 Agreement to Composition

Consider an evaluator who assigns an overall score y along with k fine-grained trait scores $a_1, a_2, \ldots, a_k$. The standard validation process compares these scores with the corresponding human reference scores (y_l versus y_h) to assess their agreement, treating these scores as independent quantities checked one at a time.

In our study, we focus on this combination with the implicit assumption in multidimensional evaluation: that an overall evaluation has a stable and recoverable relationship to the fine-grained traits. In other words, there exists a mapping $f : (a_1, a_2, \ldots, a_k) \mapsto y$ that describes how parts become a whole. Our intuition comes from frameworks that decompose holistic quality into criteria (Liu et al., 2023; Ye et al., 2024), learn a calibrated criteria-to-overall mapping (Hashemi et al., 2024), or aggregate criterion judgments into a scalar reward (Lv et al., 2026). This modeling choice is also observed outside NLG evaluation. Linear models of judgment have long been used to recover how human evaluators weight the components of a holistic decision (Kaufmann et al., 2013). Unlike these, where the mapping is used as a means to an end, we treat the mapping as an object of measurement. We require only that this relationship be systematic rather than arbitrary, so that the mapping can be estimated from the observed scores. It is a modeling choice for diagnostic purposes and not a cognitive claim about the evaluator. Our empirical findings confirm that such a mapping is recoverable across all tasks, rubrics, and judges, including the human reference. The detailed formulation of f and its estimation are in the Appendix.

Estimating compositional mapping Consider an evaluator $s \in \{h, l\}$, viz. the human evaluator or the LLM judge. Each item x_i gets an overall score y_i^s and trait scores $a_{i,1}^s, \ldots, a_{i,k}^s$. We estimate f^s as a linear model,

$$y_i^s = \beta_0 + \sum_{j=1}^{k} \beta_j^s \cdot a_j^s + \epsilon_j \tag{1}$$

where the trait scores are standardized so coefficients are comparable. Each β_j^s denotes the relative contribution of trait j to the overall evaluation. We refer to the vector $\boldsymbol{\beta}^s = (\beta_1^s, \ldots, \beta_k^s)$ as the *composition signature* of the evaluator s. Notably, linearity is the diagnostic choice for better interpretability and not the cognitive claim. Our experiments with linear attribution methods such as linear regression along with non-linear methods such as SHAP and Gradient Boosted Regression Trees do not change the compositional conclusions.

3.2 Defining Compositional Validity

Given a human mapping f^h and an LLM judge's mapping f^l, we define the judge's compositional validity as the degree to which f^l matches f^h. It measures correspondence between how the judge and humans combine traits into an overall score. It

is distinct from the agreement. A judge can match humans on the overall score ($y^l \approx y^h$) and on individual traits ($a^l_j \approx a^h_j$), but still combines them differently ($f^l \neq f^h$). This divides the evaluation problem into two questions that the agreement conflates: 1) Scoring: does the judge assign the correct value to each trait? and 2) Aggregation: does it combine these values into the overall judgment as humans do? A judge can fail in either of these, scoring traits well, but aggregate them by the wrong rule, or aggregating faithfully but with poorly scored traits. Identifying this failure is the central diagnostic focus of our work.

3.3 Diagnostic Setup

We study compositional validity through three complementary lenses, each focusing on a different question about mapping f.

***Recoverability and Stability*:** We first ask whether f^l can be recovered from a judge's trait and overall scores, and, when it can, whether the recovered mapping is stable across attribution methods and aligns with f^h. We report the cross-validated coefficient of determination R^2 of Eq. 1. A high R^2 indicates that the overall score is largely a function of the trait scores, i.e., there exists a composition for further analysis. Importantly, this measures interpretability or recoverability, and not human alignment.

Stability across methods determines whether a mapping is well-defined enough to compare. If different estimators recover different orderings from the same scores, no single mapping is reliably recoverable. We compute Kendall's τ between the induced trait rankings induced by different attribution methods (ridge coefficients, leave-one-trait-out ablation, SHAP). High τ indicates stable, method-independent importance estimates. A low τ indicates that no single mapping can be reliably recovered from the judge's scores, in which case the composition is not well defined, and we do not interpret its human alignment. Therefore, we treat stability as a prerequisite for interpreting compositional validity rather than as an established universal property.

Where the mapping is well-defined, alignment with f^h is the compositional validity itself. We measure compositional validity as Kendall's τ between the trait rankings induced by β^l and β^h over k traits. High recoverability with low alignment is the primary failure we target, viz., a judge whose composition is legible and self-consistent yet unlike that of a human judge.

***Aggregation diagnosis*:** When a judge's composition diverges from humans, the divergence may be in trait scores or in their combination. We investigate the aggregation component by an Alignment Diagnostic Test (ADT) applying human composition weights to the judge's trait scores. The adjusted overall score is computed as,

$$\hat{y}_i^{\text{adj}} = \sum_{j=1}^{k} \beta_j^h \, a_{i,j}^l, \tag{2}$$

and the reweighting benefits are measured as,

$$\Delta\rho = \rho\big(y^h, \hat{y}^{\text{adj}}\big) - \rho\big(y^h, y^l\big). \tag{3}$$

$\Delta\rho > 0$ indicates that the trait scores were informative, but combined by a rule different from that of human evaluators. In such cases, the failure lies in aggregation and not in scoring. Since the ADT requires only β^h and the judge's raw trait scores, it is applicable even to judges whose own composition signature is not stably recoverable.

***Rubric Geometry*:** Whether composition is recoverable, stable, and human-aligned depends on both, the judge and the evaluation rubric. When traits are highly entangled, many weightings fit the same score equally well, and hence there is no single stable or interpretable mapping. We characterize this through the trait-score correlation matrix at two levels: the nominal trait structure (the k traits and criteria an evaluation rubric defines) and the observed trait space (the correlation structure the evaluator's scores actually produce). These need not agree—a rubric may nominally define six independent dimensions while an evaluator collapses them onto few. We find that the same rubric is interpreted differently by different judges and that judges frequently consider more entangled trait spaces than the human reference. From the SVD for this matrix we compute two quantities (Details in Appendix): 1) *Effective Rank*, the exponential entropy of the normalized singular values, measures the trait independence; and 2) *Condition Number*, the ratio of the largest to the smallest singular values, measures the weight estimation stability. The observed trait space of the human reference is what a practitioner can audit before deploying a judge, since it requires only the existing human annotations. We argue that this rubric-level geometry is a dominant and auditable determinant of compositional validity.

Together, these move from establishing that composition exists and can be misaligned, to locating where the misalignment lies, and explaining when a rubric permits valid composition in the first place. A confirmatory probe additionally tests whether trait scores respond specifically to their intended dimension under the targeted perturbation. It validates that the measured trait behavior exhibits genuine sensitivity rather than correlation (details in section 5).

4 Experimental Setup

4.1 Tasks and Datasets

We analyze four datasets representing 3 NLG evaluation tasks (*Summarization*, *Dialogue*, and *Essay*), with human annotations for both overall quality and traits (Table 1). To enable within-domain comparison for further investigation of rubric effects, we consider two essay datasets (ASAP++ and ELLIPSE). This comparison is central to rubric geometry diagnostics, where we separate rubric effects from content effects. Details of trait definitions and annotation guidelines are in the Table 1 and Appendix.

Annotations and Evaluation Rubric All human annotations used in this study are external and originally reported with the datasets used. Each data set is published and peer-reviewed as a benchmark, with scores collected independently and prior to this work. We adopt every trait definition, scoring scale, and evaluation rubric as-is from the original dataset documentation (Stiennon et al., 2020; Mehri and Eskenazi, 2020; Mathias and Bhattacharyya, 2018; Crossley et al., 2023). No post-hoc rubric adjustment, tuning, or trait selection is performed during this study. Rubrics were fixed prior to any experimental run. In summary, the human annotations and evaluation rubric are externally validated benchmarks and our study introduces no new annotation protocol of its own. Detailed information about annotator recruitment, training, and inter-annotator agreement (where reported) can be found in the respective source publications. For ELLIPSE dataset, each essay was scored by two independent raters and the reported human score is their average. Inter-rater agreement is moderate across all six traits (Spearman ρ from 0.62 for vocabulary to 0.69 for conventions), indicating that the human reference is not a single-annotator artifact. One trait requires exclusion. USR's (Mehri and Eskenazi, 2020) understandable trait is constant in the persona subset, viz., all samples are rated 1. It carries no variance and is excluded from the geometry analysis. The USR rubric therefore nominally defines five traits while realizing four. Geometry is computed on complete-case human scores (n = 299/59/97/1031).

4.2 LLM Judges

We evaluate four models (Claude Sonnet 4.5, GPT-5, Gemini 2.5 Pro, Mistral Large 2)[1], with 3 prompting strategies per task (rubric-based, attribute-first, chain-of-thought), totaling 36 configurations across three tasks. Additionally, we consider 3 configurations for ELLIPSE. For counterfactual experiments, we use LLaMA-3.1-8B and Claude Sonnet 4.5. All judges use the default temperature. The prompts follow the published annotation guidelines, providing trait definitions and scoring scales. The output is a structured JSON with per-trait scores, overall score, and reasoning. Full prompts are in Appendix. Additionally, we define competent models as those achieving $\rho \geq 0.3$ with human overall scores across configurations, following the rationale that attribution analysis of non-systematic evaluators would not reveal meaningful compositional patterns. For analysis, Mistral is excluded as the weakest judge.

5 Results

Agreement at overall and trait levels

Alignment with humans is measured across models, tasks and prompting strategies (Figure 2). The strongest configuration varies by task, with no universally best prompting strategy (Appendix Table 4). This shows that multidimensional alignment is jointly shaped by task, rubric, and prompt format, rather than by a single, universally best prompting strategy.

Aggregate alignment is nearly identical at the two levels of granularity, but this masks opposing within-task patterns (Table 2). Summarization shows the strongest overall alignment and the weakest trait-level alignment; essay evaluation reverses this; dialogue is balanced. By attribute type, verifiable traits such as coverage, accuracy, and uses_knowledge align more systematically with human judgments (ρ= 0.55–0.82) than subjective traits such as coherence and naturalness

[1]We use Gemini for Gemini-2.5-pro, GPT for GPT-5, LLaMA for LLaMA-3.1-8B, and Claude for Claude-Sonnet-4.5 unless stated otherwise.

Dataset	N	k	Eff.R.	Cond.	R^2_h	ρ_{max}	Traits	Type
TL;DR	350	3	2.78	2.31	0.885	0.32	coverage, accuracy, coherence	V, V, S
USR	60	5 (4)	3.66	3.77	0.848	0.57	understandable, natural, maintains_ context, engaging, uses_knowledge	-, S, SO, S, V
ASAP++	100	5	1.74	47.3	0.547	0.91	content, organization, word_choice, sentence_fluency, conventions	V, SO, S, S, SO
ELLIPSE	1031	6	2.12	32.0	0.912	0.83	cohesion, syntax, vocabulary, phraseology, grammar, conventions	S, V, SO, S, V, SO

Table 1: Rubric characteristics. Effective rank (Eff.R.) and condition number (Cond.) are computed from the SVD of the human trait-score correlation matrix; higher effective rank indicates more independent traits, higher condition number less stable weight estimation. R^2_h is the cross-validated recoverability of the human overall score from human trait scores. ρ_{max} is the maximum pairwise trait correlation. k is the number of traits the rubric nominally defines; Trait inventory across the four datasets. V = verifiable, SO = semi-objective, S = subjective; For USR, *understandable* is constant across the persona subset and is excluded, so the rubric nominally defines five traits while realizing four and geometry is computed over four traits.N is the released dataset size; geometry is computed on complete cases (4.1)

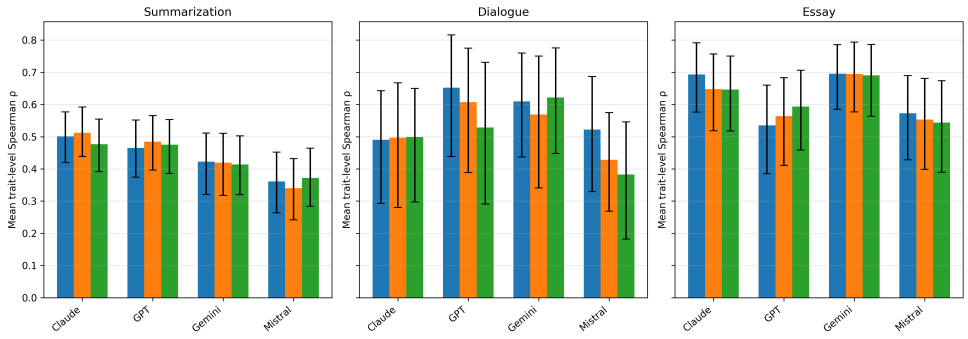


Figure 2: Mean alignment after bootstrapping across 3 tasks, 4 models, and 3 prompt strategies (Rubric: Blue, attribute_first: Orange, CoT: Green)

(0.29–0.48). A deeper analysis of the Essay evaluation (ELLIPSE dataset, Details in Appendix) confirms the same pattern at higher resolution. We observe that agreement at one level of granularity is already a poor predictor of agreement at the other, the rest of this section analyzes what it fails to capture.

Overall judgments are recoverable from trait scores

We test the assumption of 3.1 directly. In the matched cross-task setting, linear trait-to-overall composition reports R^2 = 0.675 to 0.945, across all 36 configurations (Appendix Table 4) indicating that multidimensional LLM evaluation is usually decomposable, although the strength of this decomposition varies across rubric settings. This pattern is stable across prompting strategies for all four judges (Appendix Table 4). On ELLIPSE, cross-validated ridge regression predicts the overall essay scores with high accuracy for both humans and LLM judges: humans (R^2 = 0.912), and LLM judges (Gemini 0.916, Claude 0.915, GPT 0.766). We observe that the assumption holds for every task, rubric, and judge considered, and importantly, that it holds for the human reference as well. However, this establishes only that a composition exists to analyze, not that a judge's composition resembles the human one.

Recoverability does not imply validity

The distinction between predictive fit obtained by regression and correct composition is well established. Anderson (1971) reports that integration models of human judges routinely achieve high correlation with observed responses, while concealing systematic misalignment in how components are combined. Our recoverability observations concur with this and put LLM judges in the same position. Before investigating composition alignment between LLM judge and the human evaluator, we probe whether the judge has a well-defined and

Task	Overall ρ [95% CI]	Trait-level ρ [95% CI]
Summarization	0.577 [0.507, 0.642]	0.437 [0.346, 0.522]
Dialogue	0.513 [0.297, 0.705]	0.534 [0.333, 0.698]
Essay	0.498 [0.338, 0.630]	0.619 [0.484, 0.730]
All 36 configurations	0.542 [0.511, 0.573]	0.541 [0.512, 0.570]

Table 2: Mean overall score and trait-level alignment with human judgments, bootstrapped. Aggregate alignment is nearly identical at the two levels, but the two diverge within tasks in opposite directions.

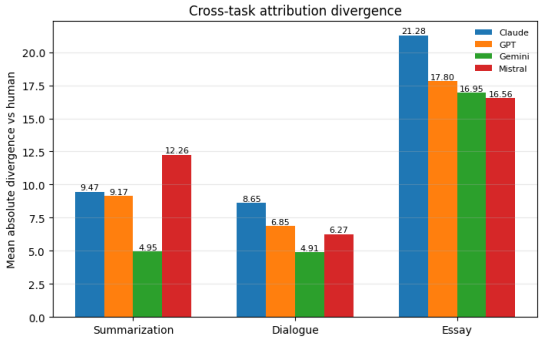


Figure 3: Cross-task attribution divergence from human composition. Bars show the mean absolute difference between LLM-inferred & human trait-importance patterns, averaged across traits and prompting strategies for each task–model pair. Higher values indicate larger departure from the human evaluative structure.

stable composition to compare. We recover each judge's composition using three attribution methods and measure their pairwise agreement using Kendall's τ (3.3). The τ here is computed over k = 6 traits (ELLIPSE dataset) leading to effective sample size 6. Under the exact null obtained by enumerating all 720 orderings, τ has standard deviation of 0.355, and $|\tau| \geq 0.733$ is required for two-sided significance at $\alpha = 0.05$. We therefore report exact permutation p-values throughout rather than interpreting τ magnitudes directly, and treat absence of detectable agreement as distinct from evidence of disagreement [2]. This agreement is not uniform across judges. On ELLIPSE, GPT's three attribution methods recover an identical trait ordering (τ = 1.00, exact p=0.003). Gemini (τ = 0.733, p = 0.056) and Claude (τ = 0.467, p=0.272) do not show significance against the null distribution for a six-trait ranking. We therefore treat only GPT's composition as reliably recoverable. For Gemini and Claude, no well-defined, single composition signature is found.

We find that this variation follows the rubric entanglement in each judge's observed trait space. GPT with highest effective rank (3.24) for its observed trait space is the judge whose composition is stably recovered. Claude (1.84) and Gemini (1.77) both exhibit substantially more entangled trait spaces and accordingly, their composition signatures are unstable. Rubric geometry helps in predicting this behavior. When the traits are entangled, many weightings fit the same score equally well, and no unique signature can be recovered for the evaluator.

Even when the signature is well-defined, we find that it is not human-aligned. GPT's trait ordering shows no statistically detectable agreement with human ordering (τ =0.20, exact p = 0.719). The judge composes its own overall judgment stably using a rule different from the human evaluator. For Claude and Gemini, the corresponding values (τ = -0.07 and 0.20 resp.) cannot be read as divergence, since there is no stable composition to diverge. In these cases, the alignment question is blocked by the geometry instead of answered by it.

The cross-task attribution divergence shows the same departure from human weighting. These are differences between standardized coefficients rather than rank correlations over traits, and hence they are not subject to power limitation as above. In summarization, the human reference is strongly coverage-driven. Rubric prompting comes closest to this pattern, especially for Gemini rubric, whereas attribute-first and CoT assign greater importance to accuracy and coherence and less to coverage, e.g. summarization-claude-attribute_first underweights coverage by -20.83 while overweighting accuracy (+7.25) and coherence (+13.58). In dialogue, judges often overweight naturalness and engagingness while underweighting uses_knowledge and sometimes maintains_context. In essay evaluation, the attribution divergence is the strongest. LLM judges repeatedly underweight content and organization, while overweighting word_choice, sentence_fluency, and conventions (Figure 3, Table 6). We observe that judges learn stable but systematically different weighting rules from humans.

[2]This bound applies only to the trait-ranking comparisons. Overall-score alignment, recoverability, and the ADT gain are estimated over the full sample of texts (n = 1,031 for ELLIPSE) and are correspondingly well powered. See Appendix B.

The failure lies substantially in aggregation

To diagnose whether this mismatch is caused by poor trait scoring or poor aggregation, we apply the Alignment Diagnostic Test (ADT) (3.3), substituting human composition weights into the judge's trait scores. In ELLIPSE, recalibration improves final-score alignment for all three judges, with gains in Gemini +0.044 [95% CI: 0.025, 0.062], GPT +0.077 [0.054, 0.100], Claude +0.049 [0.030, 0.068] (5-fold, essay-level bootstrap). This indicates that at least some part of the misalignment lies not in the trait scores themselves, but in how those scores are combined. This holds for all three judges, including those whose own composition signature is not stably recoverable.

Rubric geometry governs when composition is valid

Principal Component Analysis (PCA) on ELLIPSE trait scores shows that LLM judges produce more entangled trait spaces than humans. For Gemini and Claude, only 2 principal components from PCA explain 90% of the variance, compared with 3 for humans and 4 for GPT, consistent with their lower observed effective ranks reported above. Some rubric dimensions are therefore not independent signal, so a decomposition may look stable while still having a limited interpretive value.

Geometry follows the rubric, not the content

Trait compositional structure could in principle be a result of the content being evaluated rather than the evaluation rubric. We separate the two directly by holding the content fixed and swapping the rubric. The same ELLIPSE essay samples are re-scored under the ASAP++ rubric. If the attribute entanglements were driven by the texts, the geometry should also follow the ELLIPSE content. On the other hand, it should follow the ASAP++ if driven by rubric design. The condition number doesn't essentially change (49.8 vs 50.2), and matches ASAP++'s native value (47.3, Table 1) rather than ELLIPSE (32.0). Rubric structure, not the content, determines whether a stable composition can be recovered, which is what makes rubric geometry auditable before deployment. This control changes the rubric while keeping the context constant. It does not establish that geometry is invariant to content, since the observed trait-correlation structure is a joint property of the rubric and the score distribution a given corpus induces.

Correlational alignment overstates trait-level competence

The decomposability (R^2) and the attribution stability (τ) are correlational. They do not establish that a trait score responds to the dimension it names. We construct a targeted-perturbation experiment as a confirmatory probe. For each of the 5 target ELLIPSE traits (syntax excluded), we generate degraded versions of $\sim$ 300 essays with trait-specific perturbation instructions (e.g., for *grammar*: introduce subject-verb disagreements and tense inconsistencies while preserving ideas and structure). Both original and perturbed essays are scored by Claude and LLaMA on 5 traits plus overall quality. The magnitude of the perturbation varies by design: cohesion (57% character change) and conventions (43%) involve substantial rewrites; grammar (5%) and phraseology (9%) are more targeted.

We do not assume edits are perfectly isolated. We verify the perturbation localization (Details in Appendix). Perturbations are generated with trait-specific instructions and their magnitude is measured per trait. The non-target score change is consistently smaller than the target change for both judges (Table 5), and the vocabulary null control[3] (<0.1% character change) yields negligible drift. The reported specificity (Target Δ - Non-target Δ) represents this localization, net of general drift. Claude responds specifically to targeted perturbations of conventions (specificity = -0.112) and cohesion (-0.072) but not to phraseology (+0.035, worse than the null baseline). In contrast, a far smaller LLaMA shows stronger specificity than Claude for conventions (-0.254) and grammar (-0.164), while both models fail in phraseology. We observe that traits with comparable correlational alignment can exhibit very different targeted sensitivity, and that model capacity does not predict which. Decomposable and correlationally-aligned outputs need not reflect trait-level competence.

6 Discussion

Rubric design is a primary factor, not a secondary annotation choice

Rubric design defines what traits are available and whether those traits behave like distinct evaluation dimensions. Because observed trait spaces are of substantially lower dimension than rubrics nominally define (2–4 principal components explain

[3] null control is a perturbation small enough to produce no detectable score change in a well-calibrated judge

90% of variance), a richer rubric does not automatically yield better multidimensional evaluation. When traits are highly correlated, multidimensional outputs may overstate how much unique information each trait actually contributes. The cross-rubric control confirms this structural instability follows rubric design rather than content—which is why rubric geometry should be treated as a primary factor and audited before deployment.

A practical audit workflow

Our framework is diagnostic, but it provides concrete directions for mitigation.

- Audit the rubric: Compute the effective rank and condition number of the evaluation rubric from the human trait-correlation matrix. No LLM judge is needed. A high condition number with an effective rank below the actual trait count k indicates that the trait-level outputs are not independent signals (e.g., ASAP++, cond. 47 and eff. rank 1.74 over five traits collapse to 2).
- Prune or merge entangled traits: Merge collinear pairs (e.g., 'word_choice'–'sentence_fluency' into a single "language use" dimension) until the condition number drops.
- Diagnose before correcting: When trait scores are individually informative but aggregated differently from humans (positive ADT gain, e.g. all three ELLIPSE judges ($+0.04$ to $+0.08\,\rho$)), apply human-weight recalibration rather than pruning. The ADT applies even without a recoverable judge signature.
- Prefer rubric-anchored prompting: Rubric-based prompting is the safest option in summarization and essay evaluation, whereas dialogue is more model-dependent. Attribute-first and CoT prompting amplify aggregation mismatch.
- Do not rely on model capacity: The task and rubric explain far more variance in aggregation mismatch than the choice of model, and a smaller model shows stronger targeted sensitivity than a larger one on two of five traits.

7 Conclusion

LLM judges have emerged as a scalable alternative to human evaluation. However, reliable multidimensional evaluation requires more than high overall agreement, because agreement overlooks how a judge constructs the overall evaluation from fine-grained traits. We formalized this missing axis as compositional validity, the degree to which a judge's trait-to-overall aggregation matches the human one, and proposed a diagnostic framework to measure it. Across three NLG tasks, overall scores are strongly recoverable from trait profiles (R^2 = 0.675 to 0.945), yet this recoverability does not imply validity. Where a judge's composition is stably recoverable, it shows no statistically detectable alignment with the human trait ordering. Our Alignment Diagnostic Test reveals that a substantial component of this misalignment lies in how traits are aggregated rather than how they are scored, which is correctable through recalibration. Additionally, targeted perturbations confirm that traits with strong correlational alignment can fail to respond when specifically degraded — decomposable outputs are not the same as compositionally grounded ones. Rubric geometry is the primary determinant of whether evaluation delivers genuine diagnostic value. It constrains whether a composition can be recovered, explains far more variance in aggregation mismatch than the choice of model (η^2 = 0.384 vs. 0.022), and follows the rubric rather than the content. Rubrics with higher effective rank and lower condition numbers support stable, human-aligned compositions and entangled rubrics produce misleadingly structured outputs regardless of model capability. Trustworthy evaluation requires not only that judges agree with humans, but that they do so for the right compositional reasons.

Limitations

Use of Linear Attribution Methods

Linear models have a long history of being used as descriptive representations of human judges. They uncover a judge's cue-weighting policy without claiming that the judge computes linearly, across various lens-model studies (Kaufmann et al., 2013). Equation 1 is diagnostic in exactly this sense. Linear attribution methods may not fully capture the interplay among attributes. However, there is a distinction that needs to be considered. We use linear methods as a diagnostic tool, not as a predictive modeling tool, to infer how LLMs actually integrate fine-grained attribute information. We use linear methods as a theoretically optimal and interpretable choice for our diagnostic pur-

poses. Theoretically, Lundberg and Lee (2017) prove that Shapley values-computed through linear approximations-provide the unique solution satisfying key properties for additive feature attribution: local accuracy, missingness, and consistency. Works such as (Parcalabescu and Frank, 2024; Horovicz and Goldshmidt, 2024) demonstrate SHAP acceptance and effectiveness in interpreting human-AI alignment for modern LLMs. Additionally, work in (Belinkov, 2022) establishes linear probes as the preferred standard in NLP interpretability research to avoid confounding probe capacity with model capabilities. Empirically, recent research (Chen et al., 2025) directly validates the linear approximation assumption. To validate the assumption of attribute discriminability required for linear decomposition, we examined pairwise correlations between ground-truth human ratings across all dimensions. Linearity is an informed choice for interpretability and not a cognitive claim about the model. Additionally, we investigated the nonlinear decomposition setting with SHAP (Random forest regressors) and Gradient Boosted Regression Trees (GBRT). We find that nonlinear compositional models do not significantly change the core assumption.

Statistical power of the trait-ranking comparison

Compositional validity is computed with reference to two trait orderings. Accordingly, its effective sample size is the number of fine-grained traits k and not the number of samples. For ELLIPSE rubric, k=6 and the null distribution of Kendall's τ over six items has a standard deviation of 0.355. A six-trait ranking must reach $|\tau| \geq 0.733$ before it is distinguishable from chance at $\alpha = 0.05$. This bound is the property of the measurement instrument and not of our data. It applies equally to LLM-Human agreements, and it means there is an absence of detectable alignment but the degree of misalignment cannot be quantified. It does not apply to the overall-score alignment, the recoverability, or the ADT improvements. All of these are estimated over the full sample of texts. Rubrics with more traits would tighten this bound. We report exact permutation p values throughout rather than interpreting τ magnitudes directly.

Recoverability of the composition signature/mapping

The alignment question is appropriate only for judges with stable and recoverable composition signature. On ELLIPSE, only GPT qualifies under this condition. In case of Claude and Gemini, attribution methods provide different orderings from the same scores. Additionally, we do not report an internal stability estimate for the human reference. The ADT depends on the human composition weights and the human condition number of 32.0 in ELLIPSE indicates that the weights are estimated under some collinearity. We treat the human reference as given, with no intention of modifying the source benchmarks. The stability of the human composition signature is identified for future work.

Sample Size

Our analysis spans three diverse tasks, and the per-task sample sizes range from 58 to 350 texts, placing limits on statistical robustness. We have adequate power (1-$\beta \geq 0.80$) to detect large effects (Cohen's $d \geq 0.8$) but limited power for small effects ($d < 0.3$). Our bootstrapped confidence intervals account for uncertainty from limited samples, but future replication with N $>$ 200 per task would strengthen statistical conclusions. We use larger ELLIPSE essay dataset for detailed analysis. This limitation concerns the number of evaluated texts. It is distinct from the trait-count bound discussed above, which a larger text sample does not address.

Prompt Coverage

We test three distinct prompting strategies (rubric-based, attribute-first, CoT), each strategy with fixed phrasing. Our task-level findings are stable across these three strategies, but we cannot claim robustness to arbitrary prompt variations. Future work should systematically vary prompt phrasing within each strategy type to assess sensitivity.

Task Coverage

We consider three tasks for our initial study, but broader validation across additional NLG types would strengthen generalization. Tasks such as machine translation, data-to-text generation, and code generation exhibit unique multicollinearity patterns requiring out-of-the-box mitigation strategies.

Annotator Coverage

As noted in §4.1, our human annotations are taken directly from the published benchmarks. Accord-

ingly, per-sample annotation counts and agreement follow each source's protocol. We agree that multi-annotator data throughout would nevertheless add confidence.

Model Coverage

Our judges are from one generation of different frontier models. Model capability advances rapidly, and a newer judge may compose differently. We note that our central findings relate to this directly, since the task and rubric explain far more variance in aggregation mismatch than the choice of model ($\eta^2 = 0.384$ vs. 0.022), and a smaller model shows stronger targeted sensitivity than a larger one on two of five traits. Our framework is model-agnostic and can be applied unchanged to any judge that emits trait and overall scores.

Behavioral Analysis

Our study focuses on input trait scores and overall evaluations. We characterize how judges combine fine-grained scores into a holistic score. We do not consider mechanistic analysis at the token-level or internal model representations. It is a complementary direction that could explain *why* certain traits are harder to judge or perturb.

References

Norman H. Anderson. 1971. Integration theory and attitude change. *Psychological Review*, 78(3):171–206.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ryan Chen, Youngmin Ko, Zeyu Zhang, Catherine Cho, Sunny Chung, Mauro Giuffré, Dennis L Shung, and Bradly C Stadie. 2025. Lamp: Extracting locally linear decision surfaces from llm world models. *arXiv preprint arXiv:2505.11772*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Kehua Feng, Keyan Ding, Jing Yu, Yiwen Qu, Zhiwen Chen, chengfei lv, Gang Yu, Qiang Zhang, and Huajun Chen. 2025. Samer: A scenario-aware multi-dimensional evaluator for large language models. In *The Thirteenth International Conference on Learning Representations*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834.

Miriam Horovicz and Roni Goldshmidt. 2024. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Esther Kaufmann, Ulf-Dietrich Reips, and Werner W. Wittmann. 2013. A critical meta-analysis of lens model studies in human judgment and decision-making. *PLOS ONE*, 8(12):e83528.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Ryan Koo, Ian Yang, Vipul Raheja, Mingyi Hong, Kwang-Sung Jun, and Dongyeop Kang. 2025. Learning explainable dense reward shapes via bayesian optimization. *Preprint*, arXiv:2504.16272.

Weiyuan Li, Xintao Wang, Siyu Yuan, Rui Xu, Jiangjie Chen, Qingqing Dong, Yanghua Xiao, and Deqing Yang. 2025. Curse of knowledge: When complex evaluation context benefits yet biases llm judges. *arXiv preprint arXiv:2509.03419*.

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.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *ACL (1)*, pages 7641–7660.

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.

Can Lv, Mingju Chen, Heng Chang, and Shiji Zhou. 2026. Mitigating false credit propagation: Probabilistic graphical reward aggregation for rubric-based reinforcement learning. *arXiv preprint arXiv:2606.03361*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests. *CoRR*, abs/2412.13091.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Wenxuan Wang, Juluan Shi, Zhaopeng Tu, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2024. The earth is flat? unveiling factual errors in large language models. *arXiv preprint arXiv:2401.00761*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Fangyi Yu, Nabeel Seedat, Dasha Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. 2025. Beyond pointwise scores: Decomposed criteria-based evaluation of llm responses. *Preprint*, arXiv:2509.16093.

Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023. Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv:2308.01240*.

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.

A Notation

Table 3 summarizes the notation used in Section 3. The composition signature $\boldsymbol{\beta}^s$ is estimated separately for each source $s \in \{h, l\}$ and for each task–model–prompt configuration. Trait scores are standardized within each configuration before fitting Eq. 1, so that coefficients are comparable across traits with different score ranges.

Symbol	Meaning
y^h, y^l	overall score by human, by LLM judge
a_j^h, a_j^l	score for trait j by human, by LLM judge
k	number of traits in the rubric (nominal)
f^s	composition mapping for source s
$\boldsymbol{\beta}^s$	composition signature, viz., the coefficient vector of f^s
ρ	Spearman rank correlation
τ	Kendall rank correlation
$\Delta\rho$	ADT reweighting benefit (Eq. 3)
κ	condition number of the trait-score correlation matrix

Table 3: Notation.

Estimation details. We fit Eq. 1 with ridge regression and report cross-validated R^2. Composition signatures are additionally recovered by leave-one-trait-out ablation, viz., the drop in R^2 when trait j is removed, and by SHAP with Random Forest regressors. The three methods are used jointly to assess stability, as described in Section 3.3. Geometry is computed on complete-case scores.

B Statistical Treatment of Trait-Ranking Agreement

The unit of analysis. Compositional validity is measured as Kendall's τ between two trait rankings, viz., the ordering of traits induced by the judge's composition signature $\boldsymbol{\beta}^l$ and by the human signature $\boldsymbol{\beta}^h$. It is important to distinguish the sample size at which each quantity in this work is estimated. The overall-score agreement (ρ) and the recoverability (R^2) are computed over the evaluated texts. They are therefore well powered ($n = 1{,}031$ for ELLIPSE). The composition signatures are likewise estimated from all texts. However, the *comparison* between the two signatures is a rank correlation over k traits. Its effective sample size is k and not the number of texts. For the ELLIPSE rubric, $k = 6$.

This distinction is easy to overlook and it materially affects the interpretation. A τ computed over six items is a low-power statistic, regardless of how precisely each underlying coefficient was estimated. We therefore report the null distribution of τ explicitly rather than interpreting its magnitude directly.

Exact null distribution. For k items, τ takes values on a discrete lattice with spacing $2/[k(k-1)]$. For $k = 6$, the increments are $0.0\overline{6}$. Since $k! = 720$ orderings are enumerable, we compute the null distribution of τ exactly rather than relying on the normal approximation, which is typically unreliable at single-digit n. Enumerating all 720 permutations against a fixed reference ordering gives a null distribution with mean 0 and standard deviation 0.355. Additionally, we note that the asymptotic expression

$$\mathrm{SE}(\tau) = \sqrt{\frac{2(2k+5)}{9k(k-1)}} \tag{4}$$

evaluates to 0.3549 at $k = 6$. It agrees with the exact standard deviation to four decimal places. The approximation is therefore adequate here, but we report exact permutation p-values throughout.

Under this null, $|\tau| \geq 0.733$ is required for two-sided significance at $\alpha = 0.05$ (exact $p = 0.056$). The next attainable value, $\tau = 0.600$, gives exact $p = 0.136$, so no weaker threshold reaches $\alpha = 0.10$ either. In other words, a six-trait ranking must agree almost perfectly with the reference before the agreement is distinguishable from chance.

Observed values. Table 7 reports the exact two-sided permutation p-values for the observed LLM–human trait-ranking agreements on ELLIPSE. We observe that no judge's trait ordering is distinguishable from a random ordering of the same six traits.

C Detailed Related Work

C.1 LLM-as-a-Judge

LLMs as evaluators- now referred to as LLM-as-a-Judge-are being used to determine whether output is within an established scope. Various surveys (Gu et al., 2024; Liu et al., 2023) provide systematic insights into dimensions such as formal definition and formulation, applications, and limitations. Recent works have demonstrated the use of LLM judges in diverse applications such as story evaluation (Chiang and Lee, 2023), code comprehension(Yuan et al., 2023) and general open-ended tasks like Chatbot Arena and MT-Bench(Zheng et al., 2023). Despite their promising performance and high alignment with human evaluators in diverse settings (Kim et al., 2024; Liu et al., 2023; Zheng et al., 2023; Sottana et al., 2023), even powerful LLM judges still exhibit known issues of hallucination (Ye et al., 2023; Huang et al., 2025), factual errors (Wang et al., 2024), unfaithful chain-of-thought (Turpin et al., 2023), and bias(Gallegos et al., 2024). Almost all these works report aggregate evaluation (overall ρ, accuracy, K). We extend this work beyond the overall score by asking how do judges compose their holistic judgments from fine-grained criteria?

C.2 Multidimensional Evaluation

Previous evaluation studies focused primarily on coarse-grained evaluations, limiting interpretability. Recent studies like G-Eval have studied multidimensional evaluation in their framework. G-Eval(Liu et al., 2023) proposes a foundational framework that uses GPT-4-turbo with chain-of-thought reasoning to decompose given evaluation criteria into structured steps. Feng et al. (2025) introduces SaMer, a scenario-aware multidimensional evaluator that is capable of adapting evaluation dimensions based on query context. Works such as FLASK (Fine-Grained Language Model Evaluation) (Ye et al., 2024), LMUnit (Saad-Falcon et al., 2024), HD-EVAL (Liu et al., 2024), and Yu et al. (2025) demonstrate the effectiveness of decomposed evaluation enabling interpretable assessment and improved human-LLM evaluation agreement. However, these frameworks focus on improving evaluation quality through structured criteria, not on diagnosing why and where misalignment occurs. These frameworks use decomposition to improve evaluation quality by designing better rubrics or better prompting. None of them systematically explore the compositional relationship between traits and overall scores or explore if decomposition adds genuine diagnostic value. We complement these approaches by providing analytical tools to determine whether decomposition is worthwhile for a given rubric before deploying it.

C.3 Attribution and Interpretability

Recent work on LLM explainability has focused on different granularities of interpretation. SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) provide foundational feature attribution methods, token-level attribution methods (Horovicz and Goldshmidt, 2024) apply SHAP to individual tokens, while (Chen et al., 2025) extract locally linear decision surfaces from LLM internal representations. Attribution-informed methods have been successfully explored for human-AI alignment. Turner et al. (2024) introduces activation engineering that modifies activations in inference time to steer the model outputs in a specific direction. They suggested that activation engineering can flexibly re-target LLM behavior without damaging general performance. The SHAP / LIME-based dense reward shaping framework in (Koo et al., 2025) shows a 10–15% improvement in the RLHF pipeline for LLM alignment. This technique directly addresses credit assignment, precisely the challenge facing multidimensional evaluation agents that must attribute overall quality to specific attributes such as helpfulness, harmlessness, and honesty. Previous work considers attribution at token or representation level. We apply attribution methods at the evaluation trait level.

D Attribute Types- Illustrative examples

- Verifiable: These traits have concrete, checkable referents in the source material: accuracy (can verify claims against source), coverage (can check which key points are included), uses_knowledge (can verify factual content), content (can assess idea coverage against prompt). These consistently report higher alignment (ρ = 0.55–0.82).

- Semi-Objective: Semi-objective traits are intermediate — they have partial verifiable markers but also require judgment: conventions (grammar/spelling errors are checkable, but what counts as "good conventions" is partly subjective), organization (structure can be observed, but quality of structure is judged).

- Subjective:Subjective traits generally represent holistic judgment without a concrete external referent: coherence (logical flow — no checklist), natural (human-likeness — inherently subjective), engaging (reader experience), phraseology (expression quality). These consistently achieve lower alignment (ρ = 0.29–0.48).

E LLM Judges

For each model-task combination, we test three prompting strategies:

Rubric-based We provide a detailed definition and explicit scoring criteria for all fine-grained evaluation measures and the overall quality measure.

Attribute-First In this setup, we explicitly instruct LLM judges to score fine-grained measures before evaluating the overall quality, forcing bottom-up aggregation.

Chain-Of-Thought (CoT) We introduce explicit step-by-step reasoning for every dimension before scoring.

We follow the annotation guidelines provided in the relevant work to design evaluation prompts for LLM judges. We adopt evaluation frameworks from dataset documentation (Stiennon et al., 2020; Mehri and Eskenazi, 2020; Mathias and Bhattacharyya, 2018), which are used and validated through previous works. Output is returned as structured JSON containing scores and reasoning for each trait and overall quality.

F Rubric Geometry

We characterize rubrics through two structural properties derived from the singular value decomposition (SVD) of the trait score correlation matrix:

Effective rank. The number of independent dimensions in the trait space, computed as the exponential of the Shannon entropy of the normalized singular values:

$$\text{eff_rank} = \exp\left(-\sum_i p_i \log p_i\right), \quad p_i = \sigma_i / \textstyle\sum_j \sigma_j \tag{5}$$

An effective rank of 1.0 means all traits measure essentially the same underlying quality; an effective rank equal to k (the number of traits) means all traits are fully independent. This provides a continuous, principled measure of trait independence.

Condition number. The ratio of the largest to the smallest singular value ($\sigma_{\max}/\sigma_{\min}$), indicating the numerical stability of weight estimation. When the condition number is high, multiple weight configurations fit the data equally well, making it impossible to reliably determine which traits the evaluator considers most important.

G Targeted Perturbation Validation

Correlational alignment establishes *association* between LLM and human trait scores but not trait-specific sensitivity: an LLM judge might assign "cohesion" scores from surface cues that correlate with cohesion in the training distribution without tracking cohesion-specific quality changes. To test this, we construct targeted perturbation experiments on the ELLIPSE dataset.

Perturbation generation. For each of 5 target traits (cohesion, conventions, grammar, phraseology, vocabulary) and each of 300 essays, we generate a perturbed version that specifically degrades the target trait while preserving other qualities. Perturbations are generated by rules with trait-specific degradation instructions. We measure perturbation magnitude through character change percentage, word change percentage, and sequence similarity ratio, confirming that perturbations vary in intensity across traits (from $<1\%$ for vocabulary to 57% for cohesion).

We do not assume edits are perfectly isolated. We verify the perturbation localization as follows: 1) The perturbations are generated with trait-specific instructions, and their magnitude is measured per trait (character/word change and sequence similarity; refer to Appendix), confirming the intended intensity and that the null control is near-zero; 2) It is observed that the non-target score change is consistently smaller than the change for the target trait across both judges (refer to Non-target mean Δ in Table 5). This suggests concentrated degradation in the intended dimension rather than uniform spread; and 3) The null control of the vocabulary ($< 0.1\%$ character change) yields negligible drift, thus ruling out spurious global responses.

Evaluation. Both original and perturbed essays are scored by two LLM judges: Claude Sonnet (a competent judge from our main analysis) and LLaMA-3.1-8B (a smaller model, testing the effect of capacity). Judges score 5 traits plus overall quality for each essay.

H ELLIPSE alignment details

On deeper analysis of the essay evaluation (ELLIPSE), Gemini shows the strongest overall alignment with humans (ρ = 0.739), followed by GPT (0.698) and Claude (0.681). At the trait level, mean correlation is 0.692 for Gemini, 0.637 for Claude, and 0.597 for GPT.

I Interpretation and scope

We draw a deliberately narrow conclusion from trait-ranking τ values. The correct reading is not that the LLM–human trait agreement is weak, since the data cannot support such a claim at this power. The correct reading is that no human-aligned trait ordering is detectable, viz., the observed agreements are consistent with chance. This is the weaker claim statistically and the stronger claim substantively, since it withholds any assertion that a faint alignment exists.

Two scope conditions follow. First, this power limitation applies specifically to the trait-ranking comparison, which is bounded by k. It does not apply to the overall-score alignment, the recoverability, or the ADT reweighting benefit. All three are estimated over the full sample of texts ($n = 1{,}031$) and are correspondingly well powered. The conclusions drawn from those quantities are unaffected.

Second, the same bound applies to the internal attribution-agreement values, which are also τ over $k = 6$ traits. Individually, only GPT's perfect internal agreement reaches significance ($\tau = 1.00$, $p = 0.003$). Gemini's ($\tau = 0.733$, $p = 0.056$) and Claude's ($\tau = 0.467$, $p = 0.272$) do not. We therefore do not claim that each judge's attribution is individually significantly self-consistent. Importantly, the load-bearing observation is the *contrast* between the two sets of values, viz., internal agreements of 0.47 to 1.00 against human agreements of -0.07 to 0.20. Both are estimated by the same statistic on the same k items. The contrast is therefore not attributable to the power of the estimator, since both sides share it.

Table 4: Trait-to-overall compositionality across tasks, models, and prompts

Task	Model	Prompt	ρ	MAE	Within-1	Bias	R^2	Top Dimension
Dialogue	Claude	CoT	0.587	1.638	40.00%	1.638	0.834	engaging (34.8%)
Dialogue	Claude	Rubric	0.496	1.383	58.30%	1.383	0.764	natural (28.8%)
Dialogue	Claude	attribute-first	0.572	1.475	51.70%	1.475	0.797	engaging (36.8%)
Dialogue	GPT5	CoT	0.5	1.017	78.30%	0.983	0.857	maintains_context (31.3%)
Dialogue	GPT5	Rubric	0.571	0.917	88.30%	0.85	0.676	maintains_context (32.0%)
Dialogue	GPT5	attribute-first	0.626	0.867	86.70%	0.833	0.922	engaging (36.4%)
Dialogue	Gemini	CoT	0.515	1.1	71.70%	0.933	0.844	engaging (30.9%)
Dialogue	Gemini	Rubric	0.633	0.767	88.30%	0.733	0.888	engaging (34.0%)
Dialogue	Gemini	attribute-first	0.418	1.15	73.30%	0.95	0.883	engaging (33.6%)
Dialogue	Mistral	CoT	0.418	1.593	44.10%	1.593	0.687	uses_knowledge (29.6%)
Dialogue	Mistral	Rubric	0.377	1.136	71.20%	1.102	0.717	natural (34.2%)
Dialogue	Mistral	attribute-first	0.448	1.322	62.70%	1.322	0.720	engaging (41.6%)
Essay	Claude	CoT	0.522	1.299	60.80%	1.299	0.939	sentence_fluency (68.4%)
Essay	Claude	Rubric	0.577	0.845	85.60%	0.845	0.934	sentence_fluency (38.5%)
Essay	Claude	attribute-first	0.534	1.237	67.00%	1.237	0.856	sentence_fluency (51.4%)
Essay	GPT5	CoT	0.531	1.485	46.40%	1.485	0.675	word_choice (43.3%)
Essay	GPT5	Rubric	0.346	1.01	76.30%	0.969	0.784	word_choice (57.0%)
Essay	GPT5	attribute-first	0.566	1.495	45.40%	1.495	0.805	word_choice (50.4%)
Essay	Gemini	CoT	0.508	1.32	58.80%	1.32	0.882	word_choice (50.6%)
Essay	Gemini	Rubric	0.628	0.784	89.70%	0.784	0.842	sentence_fluency (24.9%)
Essay	Gemini	attribute-first	0.565	1.361	58.80%	1.361	0.884	sentence_fluency (40.0%)
Essay	Mistral	CoT	0.385	1.093	73.20%	1.052	0.806	organization (31.6%)
Essay	Mistral	Rubric	0.45	0.701	92.80%	0.515	0.805	sentence_fluency (25.4%)
Essay	Mistral	attribute-first	0.431	0.938	78.40%	0.938	0.812	sentence_fluency (44.7%)
SUMM	Claude	CoT	0.686	1.185	65.80%	0.639	0.931	coverage (64.9%)
SUMM	Claude	Rubric	0.704	1.198	62.90%	0.611	0.945	coverage (82.7%)
SUMM	Claude	attribute-first	0.679	1.144	67.20%	0.357	0.909	coverage (60.2%)
SUMM	GPT5	CoT	0.621	1.234	63.50%	0.273	0.909	coverage (66.1%)
SUMM	GPT5	Rubric	0.65	1.229	62.90%	0.38	0.869	coverage (87.4%)
SUMM	GPT5	attribute-first	0.632	1.215	64.90%	0.012	0.900	coverage (63.2%)
SUMM	Gemini	CoT	0.615	1.426	58.20%	0.623	0.919	coverage (72.4%)
SUMM	Gemini	Rubric	0.695	1.212	65.20%	0.572	0.891	coverage (79.4%)
SUMM	Gemini	attribute-first	0.614	1.36	59.90%	0.47	0.907	coverage (69.9%)
SUMM	Mistral	CoT	0.577	1.344	60.90%	-0.301	0.876	coverage (70.6%)
SUMM	Mistral	Rubric	0.473	1.495	51.50%	-0.09	0.822	coverage (90.1%)
SUMM	Mistral	attribute-first	0.473	1.629	53.50%	-1.107	0.844	coverage (49.9%)

Table 5: Targeted perturbation results. Negative specificity indicates that the targeted trait drops more than the average non-target traits.

Judge	Target trait	Target Δ [95% CI]	Non-target mean Δ	Specificity [95% CI]	Overall Δ
LLaMA-3.1-8B	conventions	-0.400 [-0.570, -0.240]	-0.146	-0.254 [-0.370, -0.136]	-0.21
LLaMA-3.1-8B	grammar	-0.292 [-0.416, -0.180]	-0.128	-0.164 [-0.252, -0.081]	-0.11236
LLaMA-3.1-8B	cohesion	-0.067 [-0.157, 0.011]	0.004	-0.072 [-0.155, 0.007]	0.011236
LLaMA-3.1-8B	vocabulary	0.000 [0.000, 0.000]	-0.004	0.004 [0.000, 0.013]	0
LLaMA-3.1-8B	phraseology	0.000 [-0.106, 0.096]	-0.030	0.030 [-0.036, 0.096]	-0.05319
Claude	cohesion	-0.131 [-0.188, -0.074]	-0.058	-0.072 [-0.121, -0.023]	-0.06711
Claude	conventions	-0.211 [-0.268, -0.154]	-0.099	-0.112 [-0.160, -0.064]	-0.13378
Claude	grammar	-0.154 [-0.204, -0.104]	-0.113	-0.041 [-0.084, 0.003]	-0.1505
Claude	phraseology	-0.071 [-0.117, -0.017]	-0.106	0.035 [-0.013, 0.082]	-0.10774
Claude	vocabulary	-0.071 [-0.114, -0.017]	-0.017	-0.054 [-0.096, -0.013]	-0.0303

Table 6: Attribution divergence with respect to human across models and tasks

Task	Model	Prompt	MeanAbsGap	MaxAbsGap
Dialogue	Claude	Attribute-first	9.58617	11.17741
Dialogue	Claude	CoT	10.51147	14.54501
Dialogue	Claude	Rubric	5.845067	11.69013
Dialogue	GPT	Attribute-first	6.582695	12.61899
Dialogue	GPT	CoT	8.008508	16.01702
Dialogue	GPT	Rubric	5.970185	6.587025
Dialogue	Gemini	Attribute-first	4.472784	7.105544
Dialogue	Gemini	CoT	3.810048	7.620096
Dialogue	Gemini	Rubric	6.435775	11.58014
Dialogue	Mistral	Attribute-first	7.546183	15.09237
Dialogue	Mistral	CoT	2.7277	5.450091
Dialogue	Mistral	Rubric	8.524175	17.04835
Essay	Claude	Attribute-first	19.58394	41.73973
Essay	Claude	CoT	24.15037	58.76696
Essay	Claude	Rubric	20.10106	41.45862
Essay	GPT	Attribute-first	16.90706	41.36806
Essay	GPT	CoT	14.18025	34.26387
Essay	GPT	Rubric	22.31621	47.95149
Essay	Gemini	Attribute-first	17.82395	30.37357
Essay	Gemini	CoT	18.97092	41.55763
Essay	Gemini	Rubric	14.05708	28.03101
Essay	Mistral	Attribute-first	19.96702	35.22933
Essay	Mistral	CoT	14.83902	36.74466
Essay	Mistral	Rubric	14.86442	32.36997
Summarization	Claude	Attribute-first	13.88609	20.82913
Summarization	Claude	CoT	10.79387	16.1908
Summarization	Claude	Rubric	3.736321	5.604481
Summarization	GPT	Attribute-first	11.91422	17.87133
Summarization	GPT	CoT	9.973803	14.9607
Summarization	GPT	Rubric	5.611594	8.417392
Summarization	Gemini	Attribute-first	7.400979	11.10147
Summarization	Gemini	CoT	5.770958	8.656437
Summarization	Gemini	Rubric	1.674108	2.511162
Summarization	Mistral	Attribute-first	20.79654	31.19481
Summarization	Mistral	CoT	9.335588	14.00338
Summarization	Mistral	Rubric	6.656842	9.985262

Judge	τ (LLM vs. human)	z	exact two-sided p
Claude	-0.067	-0.19	1.000
Gemini	$+0.200$	$+0.56$	0.719
GPT	$+0.200$	$+0.56$	0.719

Table 7: Exact permutation test for LLM–human trait-ranking agreement on ELLIPSE ($k = 6$ traits, 720 enumerated orderings). z is the observed τ in units of the exact null standard deviation (0.355).