Life of Research · Published 2 November 2026 · DOI 10.5281/zenodo.23886823

A Checklist Is Not a Simplification: Where the Published Gains of Decomposed LLM Evaluation Come From, Why a Controlled Null on Holistic Criteria Does Not Reach Instance-Specific or Verifiable Decompositions, and the Ablations a Decomposition Claim Owes

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Automatic evaluation with a language model as judge is often improved, in published work, by splitting the judgment into parts: sub-criteria, checklist questions, rubric items or atomic claims, each scored separately and then combined. A 2026 controlled study reports that on standard summarization and dialogue benchmarks this splitting gives no gain over a direct judge given the same human labels, and that an earlier method's reported advantage came from a regressor trained on those labels. Other work reports large gains from decomposition in instruction following, factuality and reinforcement-learning rewards. In this analysis the author reviews and synthesises published results to show that the two bodies of evidence are compatible, because "decomposition" names at least six interventions that travel together. A decomposed judge may receive supervision from human labels through its aggregator, produce finer-grained scores that change how ties are counted, receive information the direct judge lacks (a reference answer, retrieved evidence, or criteria written for the instance), make parts of the judgment checkable by an external verifier, become more consistent across evaluator models without becoming more valid, and change how calls and arithmetic are executed. The narrow claim the method was proposed for, that answering simpler questions makes the same judge better at the same judgment, is the one the controlled evidence supports least. In two of the three studies whose ablations separate information from decomposition, the information term is the larger: in one rubric study, removing the reference answer costs about four times as much correlation with human labelers as the skill-specific decomposition adds, and in a checklist study a fixed generic checklist does little while checklists written for each query by a stronger model do a great deal. Read at the resolution of its tables, the controlled null is itself dataset-dependent, with decomposition ahead by 0.03 to 0.07 Spearman on one benchmark for each of three judge models and behind by up to 0.13 on the other. The paper consolidates the measured values in a table, states an attribution procedure as an algorithm, and specifies the ablations a claim that decomposition helps should report.

Keywords: LLM-as-a-judge; NLG evaluation; task decomposition; checklists; rubrics; meta-evaluation; claim decomposition; evaluation methodology

1. Introduction

Language models are now the default instrument for scoring generated text. Reference-based metrics such as ROUGE correlate weakly with human judgments of summaries, and a judge model prompted with an evaluation criterion does markedly better (Liu et al., 2023a; Fabbri et al., 2021). A strong judge agrees with human pairwise preferences more than 80 percent of the time on conversational benchmarks, about as often as humans agree with each other (Zheng et al., 2023). The same literature documents the judge's failures: preference for whichever answer is shown first, so that a weaker model can be made to beat a stronger one on 66 of 80 queries by swapping positions (Wang et al., 2023); confusion between the quality criteria it is asked to score (Hu et al., 2024); trouble ranking high-quality candidates (Shen et al., 2023); and failure to notice deliberate quality drops in more than half of perturbed answers (Doddapaneni et al., 2024).

A large family of methods responds to these failures in the same way: by breaking the judgment apart. HD-Eval decomposes a criterion such as coherence into a hierarchy of sub-criteria and trains a regressor to combine their scores (Liu et al., 2024). CheckEval turns each criterion into yes-or-no questions and reports the fraction answered yes (Lee et al., 2025). TICK generates a checklist from each instruction (Cook et al., 2024). Branch-Solve-Merge plans sub-evaluations and merges them (Saha et al., 2024). DnA-Eval proposes weighted aspects and sums them with an external calculator (Li et al., 2024b). FActScore splits a generation into atomic facts and checks each against a knowledge source (Min et al., 2023). Rubrics, instance-specific criteria and checklists have spread from evaluation into training rewards (Gunjal et al., 2025; Viswanathan et al., 2025). A 2024 survey lists decomposition of evaluation standards into "specific, discrete criteria" among the effective strategies for LLM judges (Li et al., 2024a), and an earlier survey of LLM-based NLG evaluation organises the field into four ways of using the model: deriving metrics from it, prompting it, fine-tuning it, and pairing it with human raters (Gao et al., 2024).

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

In 2026 Steindl, Voskarides, Gasparin and Marcheggiani asked the question the family had mostly skipped: does the decomposition step itself help? They reimplemented HD-Eval and CheckEval with one judge model, built a direct-prediction judge that scores the target criterion without decomposition and gave it access to the same human labels, and compared the two on SummEval, TopicalChat and SEAHORSE (Clark et al., 2023). Their answer is that they found no evidence that decomposition "leads to performance gains over a fair baseline that does not use decomposition", and that HD-Eval's reported advantage came from the human labels its aggregator is trained on (Steindl et al., 2026). An HD-Eval variant whose sub-criteria were generated without being told which criterion they were for performed about as well as the targeted version, which the authors read as showing that the gains come from the learned aggregation rather than from decomposition.

That result sits uneasily beside the positive literature. TICK reports that answering generated checklist questions raises exact agreement with human pairwise preferences from 46.4 to 52.2 percent over direct scoring (Cook et al., 2024). FLASK reports that skill-specific rubrics correlate better with human labelers than a skill-agnostic rubric (Ye et al., 2023). Checklist-based rewards improve a 7B instruction model on every benchmark in one comparison where reward models do not (Viswanathan et al., 2025). Fact-level verification with search agrees with crowd annotators on 72 percent of 16,000 facts and wins 76 percent of a sample of disagreements (Wei et al., 2024a). If decomposition does nothing, these results need another explanation. If it does something, the controlled null needs one.

The question this paper asks is what a published comparison between a decomposed and a direct LLM judge measures, and under what conditions it supports the claim that decomposition, as distinct from the things that travel with it, improved the judgment. In this analysis the author reviews and synthesises published results to show that the two literatures are compatible once the word "decomposition" is unpacked. A decomposed evaluation pipeline differs from a direct one along at least six axes besides the act of splitting: it may be supervised by human labels through its aggregator, it emits scores at a finer resolution, it may carry information the direct judge was not given, it may convert part of the judgment into something an external verifier can check, it may be more consistent across judge models without tracking humans any better, and it changes how model calls and arithmetic are executed. Published gains attributed to decomposition fall, on the evidence collected here, mostly into the first five channels. The sixth cuts both ways. The channel the methods were motivated by, that a simpler question lets the same model make a better judgment from the same information, is the one with the least direct support.

The paper argues three things. First, the anchor's null is narrower than its abstract and more interesting than its headline. It is a result about generic, criterion-level decomposition of holistic quality ratings, scored against averaged human ratings on two small benchmarks, and read at the resolution of its own tables it is not a uniform null: decomposition holds a small lead on SummEval for all three judge models tested and trails by a wide margin on TopicalChat for two of them (Figure 1a). Second, the strongest published gains come from decompositions that change the information available to the judge, by writing criteria for the instance, by providing a reference answer, or by retrieving evidence, and in the two studies that ablate the information separately from the decomposition, FLASK (Figure 1b) and RocketEval, it accounts for more of the gain than the decomposition does. In a rubric-reward study (Figure 1c), instance-specific rubrics raise the reward's value even when the judge reads them and returns one holistic score. Third, decomposition has costs that grow with optimisation pressure and with the strength of the judge: it helps weaker verifiers and hurts stronger ones in fact-checking (Hu et al., 2025), its additive aggregation is what a policy exploits in rubric-based reinforcement learning (Liu et al., 2026), and items judged together interfere with each other (Yu et al., 2026).

The argument is not that decomposition is useless. Several of the channels are worth having on their own terms. Supervision through a calibrated aggregator is a legitimate way to align a judge, provided it is reported as supervision (Hashemi et al., 2024). Consistency across judge models is a real property that a benchmark maintainer may want even at constant validity (Lee et al., 2025). Making a judgment checkable by retrieval is the established route to measuring factual precision at scale (Min et al., 2023). The argument is about attribution. A claim of the form "decomposing the criterion improved the judge" needs controls that most published comparisons do not run, and the one comparison that ran them found the claim unsupported in its setting.

The paper proceeds as follows. Section 2 fixes terms. Section 3 reads the anchor at the resolution of its tables, and Section 4 sets out the positive record. Sections 5 to 10 take the six channels in turn, each with the published results that isolate it. Section 11 consolidates the values in Table 1, and Section 12 gives an attribution procedure as Algorithm 1. Sections 13 to 15 set out the strongest objections, the studies that would settle the open part, and what is not known.

2. Scope and terms

The paper concerns the use of a pretrained language model to assign a quality score or verdict to a generated text, where the score is meant to stand in for human judgment. It covers pointwise scoring, pairwise preference and claim-level verification, and it covers

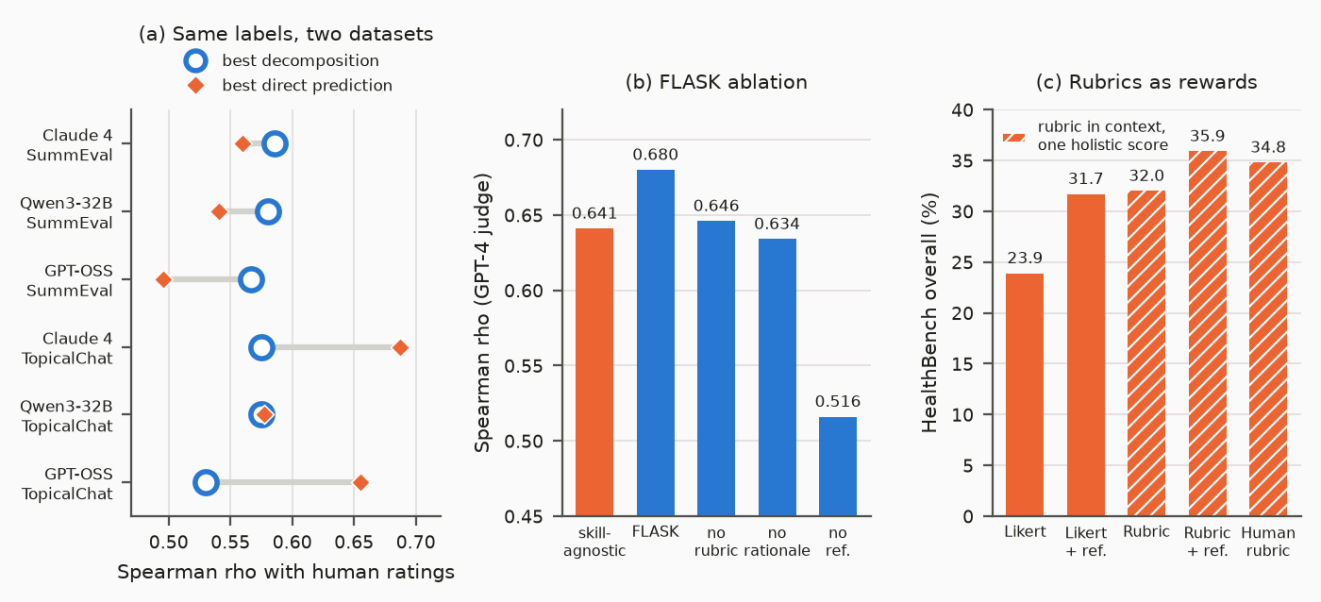


Figure 1. (a) Sample-wise Spearman correlation with human ratings for the best decomposition variant (HD-Eval, ICL or AOI decomposition, all with human labels and vertical aggregation) and the best direct-prediction variant with human labels, for three judge models on SummEval and TopicalChat; values from Steindl et al. 2026, Tables 1 and 6. (b) Spearman correlation between a GPT-4 judge and human labelers on 200 FLASK instances for a reference-guided skill-agnostic rubric, the skill-specific FLASK protocol, and FLASK with the score rubric, the rationale or the reference answer removed; values from Ye et al. 2023, Table 1. (c) HealthBench overall score of policies trained with a direct Likert reward, a reference-guided Likert reward, and a holistic judge given, as context, synthetic rubrics generated without and with reference answers or human-written rubrics (hatched); every reward in (c) is one holistic score; values from Gunjal et al. 2025, Table 1. Orange marks judgments without decomposition, blue with.

rubric- and checklist-based rewards where they bear on whether decomposed scores track quality. It does not cover decomposition as a reasoning strategy for solving tasks (Khot et al., 2022), except as background.

Holistic criterion. A quality dimension that human annotators rated as a single number, such as coherence, consistency, fluency or relevance in SummEval (Fabbri et al., 2021), or naturalness and engagingness in the TopicalChat annotations (Mehri and Eskenazi, 2020). Following Steindl et al. (2026), the paper calls these L1 criteria.

Decomposition. Any procedure that replaces one judgment of an L1 criterion with several judgments of narrower items, followed by a rule that combines them. The items may be generic sub-criteria shared across all instances (HD-Eval, CheckEval), questions written for each instance (TICK, InFoBench), or claims extracted from the text being evaluated (FActScore, SAFE). The combining rule may be an average, a weighted sum, a trained regressor or a further model call.

Direct prediction. A judge that scores the L1 criterion in one step. It may use chain-of-thought, in-context examples, a reference answer or a calibration step, as long as it does not score separate sub-items.

Channel. A difference between a decomposed and a direct pipeline, other than the splitting itself, that can change the measured result. Sections 5 to 10 name six. The splitting's own effect, which this paper calls the simplification effect, is what remains when all six are held fixed.

Validity and reliability. Validity here means agreement with the human ratings the benchmark treats as ground truth, measured by correlation or agreement rate. Reliability means agreement of the method with itself, across repeated runs or across judge models. The two are often reported together and are not the same property (Section 9).

Instance level and system level. An instance-level comparison scores each text and correlates with human scores per text. A system-level comparison averages over many texts per system and correlates the system rankings. The same method can look very different at the two levels (Section 6).

3. The anchor result, read at the resolution of its tables

Steindl et al. (2026) compare two decomposition methods, two extensions of their own, and a family of direct-prediction judges. The decomposition methods are HD-Eval, which decomposes each L1 criterion into L2 and L3 sub-criteria scored on an ordinal scale and combines them with a regressor trained on human L1 labels (Liu et al., 2024), and CheckEval, which decomposes each criterion into binary questions and reports the proportion answered yes without any human labels (Lee et al., 2025). The extensions are ICL decomposition, which shows the decomposer rated examples so that it can infer what annotators attended to, and AOI decomposition, which prompts for sub-

criteria that are atomic, observable from the text, and independent of each other. The direct judge's prompt, as the method is described, contains the criterion's name and definition, the rating scale, a reasoning instruction and five rated examples; the main table reports direct variants with and without additional in-context examples, and with and without a regressor trained on the same human labels HD-Eval uses, so that the two sides of the comparison can see the same supervision. Apart from the originally reported rows (HD-Eval, and CheckEval with Mistral-Large and GPT-4o), all methods in the main table use Claude 4 Sonnet at temperature zero, a 50/50 train-test split, and sample-wise Spearman correlation with human ratings, supplemented by the alternative annotator test (Calderon et al., 2025), which asks whether the judge is at least as good as a randomly chosen human annotator after a cost-benefit allowance and declares it a viable replacement when it wins against at least half of the annotators.

The averages in the anchor's Table 1 are close. On SummEval, the decomposition variants that use human labels reach Spearman values of 0.567 (HD-Eval reimplemented), 0.567 (ICL) and 0.586 (AOI); the direct variants with labels reach 0.560 and 0.545, and those without labels 0.553 and 0.552. On TopicalChat the order flips and the gap widens: decomposition with labels reaches 0.563 to 0.575, direct prediction with labels 0.672 to 0.687, and direct prediction without labels 0.701 to 0.716. These ranges, like Figure 1a, leave out the hybrid whose regressor also receives the direct L1 predictions as input (HD-Eval + L1), which reaches 0.560 on SummEval and 0.582 on TopicalChat, the highest TopicalChat value of any decomposition row and still below every direct variant. Every method that uses human labels reaches a winning rate of 1.0 on SummEval. The anchor's reimplementation of CheckEval with Claude 4 scores 0.411 and 0.428, well below the 0.549 and 0.645 its authors originally reported with Mistral-Large; Steindl et al. note that their conclusions hold even against the originally reported values, and the direct judges without labels are at or above the original CheckEval numbers on both benchmarks.

Two features of this table matter for what follows. The first is that the sign of the comparison depends on the dataset. The anchor's appendix repeats the comparison with Qwen3-32B and GPT-OSS-120B as judges. On SummEval, the best decomposition variant with labels leads the best direct variant with labels for all three judges, by 0.026 (Claude 4), 0.039 (Qwen3-32B) and 0.071 (GPT-OSS-120B). On TopicalChat it trails for all three, by 0.112, 0.002 and 0.125 (Figure 1a). The authors report this themselves: on SummEval the two decomposition variants they proposed "achieve the best scores" with the other models. The headline "no evidence of gains" therefore summarises a small positive effect on one benchmark and a larger negative one on the other, from single runs. The authors write: "We report results on a single run, but did not observe significant variations across runs" (Steindl et al., 2026). That addresses repeated runs of a judge at temperature zero; it is not an estimate of sampling variation over the test items. The TopicalChat annotations cover 360 responses (Mehri and Eskenazi, 2020), so a 50/50 split leaves about 180 test responses per criterion; SummEval covers 1,600 summaries (Fabbri et al., 2021). Neither the positive nor the negative differences come with an interval, and the published comparison cannot rank them. With horizontal aggregation, which the anchor reports in an appendix, the TopicalChat gap narrows but remains: 0.622 for AOI decomposition against 0.704 for a direct judge whose regressor is fitted the same way (Steindl et al., 2026, Table 3).

The second feature is that the anchor's own controls point at specific channels. The L1-agnostic variant asks the model for 25 general quality criteria without naming the target criterion, then trains the same regressor. With horizontal aggregation, in which every sub-score is a feature for every target, it reaches 0.579 on SummEval and 0.621 on TopicalChat, against 0.594 and 0.604 for targeted ICL decomposition and 0.608 and 0.622 for AOI (Steindl et al., 2026, Table 3). Sub-criteria that were never told what they were for do about as well as sub-criteria that were. This is the anchor's main evidence that decomposition does not simplify: if the gain came from breaking a criterion into its parts, parts chosen without knowledge of the criterion should do worse. The anchor's comparison of aggregation schemes (its Table 4) shows horizontal aggregation beating vertical aggregation for both of its decomposition variants on both benchmarks, which the authors attribute to the regressor having more features. And the rounding analysis shows that a regressor's floating-point output alone changes the metrics: rounding the outputs lowers correlation slightly and raises the alternative-test scores, because floats eliminate ties that integer judges cannot avoid (Section 6).

HD-Eval's original ablation is consistent with this reading, though it was not designed to test it. With half the training data and a neural-network aggregator, removing all three layers of the criteria hierarchy lowers SummEval Spearman from 0.535 to 0.513, and replacing the trained aggregator with a plain average of the sub-scores lowers it to 0.530 (Liu et al., 2024, Table 4). Both changes are small on Spearman; the larger drops are on Pearson, from 0.617 to 0.538 and 0.555. Pearson, unlike Spearman, rewards getting the linear scale right, which is exactly what a regressor fitted to human scores supplies. On TopicalChat the ablation is not monotone: removing layers 2 and 3 gives a Spearman of 0.543, while removing layers 1, 2 and 3 gives 0.590. HD-Eval's headline improvement over a GPT-4 judge, 15 percent on Pearson for summarization, is a comparison between a supervised pipeline and an unsupervised one (Liu et al., 2024).

The anchor's limitation section is explicit about scope. Its datasets are those of the methods it re-examines; it does not claim its finding extends to "complicated, multi-step reasoning tasks"; and it suggests future work in which humans rate the decomposed criteria as well as the originals (Steindl et al., 2026). Read at that scope, the result is strong. It shows that for criterion-level decomposition of holistic ratings on two standard benchmarks, the advantage previously credited to decomposition is not separable from supervision, and that a carefully prompted direct judge closes most or all of the gap. It does not reach instance-specific checklists, reference-guided rubrics, claim decomposition with retrieval, or decomposition used to construct rewards. Those are where the positive literature lives.

4. The positive record

The positive results fall into five groups, and each differs from the anchor's setting along at least one axis.

Criterion-level decomposition on the same benchmarks. CheckEval reports that across twelve judge models its checklist scores correlate with SummEval and TopicalChat ratings better than G-Eval's Likert scores, with mean Spearman across models on SummEval of 0.481 against 0.399 for G-Eval, and that its agreement across judge models rises sharply, from a Krippendorff's alpha of 0.09 to 0.48 over all models on SummEval (Lee et al., 2025). BinEval, a 2026 binary-question method, reports matching or beating UniEval and G-Eval on SummEval, TopicalChat and QAGS and producing score distributions closer to the human ones (Cho et al., 2026). UniEval itself recast multi-dimensional evaluation as Boolean question answering in 2022 and reported correlations 23 percent higher on summarization than the best earlier unified evaluators (Zhong et al., 2022). These are the results the anchor addresses directly, and the main difference from the anchor is the baseline: the comparison is against G-Eval, which uses no examples, rather than against a direct judge given rated examples or a trained calibrator.

Instance-specific checklists and rubrics. TICK generates a checklist from each instruction and has the judge answer every item; on an internal set of instructions with trained annotators, exact agreement with human pairwise preferences rises from 46.4 percent under direct 1-to-5 scoring to 52.2 percent, with an intermediate condition that puts the checklist in the prompt without requiring answers reaching 48.7 percent (Cook et al., 2024). InFoBench decomposes each instruction into requirements and reports higher consensus among annotators for the decomposed ratio than for direct scoring, especially on its harder set (Qin et al., 2024). BiGGen Bench is a holistic judge with an instance-specific rubric rather than a decomposition: each instance carries a reference answer and one scoring rubric, which "consists of evaluation criteria and descriptions for each score, ranging from 1 to 5", and the response is scored once against it. It reports that instance-specific criteria correlate with human judgments better than coarse criteria from MT-Bench or the domain-level skills of FLASK (Kim et al., 2024). RocketEval has GPT-4o write a checklist per query and a 2B model grade it, reaching a Spearman correlation of 0.965 with the Chatbot Arena ranking of the models tested on WildBench (Wei et al., 2025). WildBench builds checklists into its judge prompts (Lin et al., 2024).

Skill-level rubrics. FLASK decomposes instruction-following evaluation into twelve skills and scores each with a skill-specific rubric. With GPT-4 as judge on 200 instances rated by ten labelers, Spearman correlation with humans is 0.680, against 0.641 for a skill-agnostic rubric (Ye et al., 2023). Prometheus, a 13B judge trained on rubric-conditioned feedback, reaches a Pearson correlation of 0.897 with human evaluators on 45 customised rubrics, close to GPT-4's 0.882, when given a reference answer and a rubric (Kim et al., 2023).

Claim decomposition for factuality. FActScore splits a biography into atomic facts and checks each against Wikipedia; its automated estimator recovers human factual-precision scores with an error under 2 percent, and its authors report that retrieval makes a large difference (Min et al., 2023). SAFE replaces the fixed knowledge source with search and reports 72 percent agreement with crowd annotators at more than twenty times lower cost (Wei et al., 2024a). FineSurE checks summary sentences against the source and aligns them with key facts, reporting gains especially on completeness and conciseness (Song et al., 2024a).

Decomposed rewards. Rubrics as Rewards generates instance-specific rubrics from reference answers and uses them as rewards; the best variant, in which a judge reads the whole rubric and returns one score, improves HealthBench by up to 31 percent relative to a direct Likert reward (Gunjal et al., 2025). Reinforcement learning from checklist feedback extracts a checklist from each instruction, scores each item with a judge and, where possible, a verifier program, and improves Qwen2.5-7B-Instruct by 4 points of hard satisfaction rate on FollowBench, 6 points on InFoBench and 3 points of win rate on Arena-Hard (Viswanathan et al., 2025).

Human evaluation. The decomposition idea is older than LLM judges. In machine translation, professional annotators marking individual errors under the Multidimensional Quality Metrics framework ranked systems very differently from the shared task's own crowd-based ratings, and clearly preferred human translations to machine output (Freitag et al., 2021). LLM versions of the same error-span annotation achieved state-of-the-art system-ranking accuracy in a 2023 evaluation, with the authors' own caution about relying on a proprietary model (Kocmi and Federmann, 2023).

None of these studies is wrong on its own terms. The question is which of the differences between

decomposed and direct pipelines each one measures. Sections 5 to 10 take the differences one at a time.

5. Channel one: supervision through the aggregator

A decomposed pipeline produces several numbers per text, and something must turn them into one. If that something is trained on human ratings of the target criterion, the pipeline is a supervised model whose features happen to be LLM sub-scores. Its correlation with humans can rise because the features are better, or because the regressor has learned the annotators' scale, their weighting of aspects and their idiosyncrasies. A direct judge without a regressor has had no opportunity to learn any of these.

HD-Eval is explicit that its aggregator is trained "to map evaluation scores from decomposed criteria to human expert scores" with a mean-squared-error objective, and that a stronger regressor reduces how much labelled data is needed (Liu et al., 2024). The anchor's contribution is to give the direct judge the same opportunity: a regressor fitted on the single direct score, trained on the same human labels. With that control, the TopicalChat gap reverses and the SummEval gap shrinks to a few hundredths (Section 3). The L1-agnostic result then shows that the features need not be decompositions of the target criterion at all: 25 generic quality questions serve the regressor about as well (Steindl et al., 2026). On this evidence, what HD-Eval measured as a decomposition gain was largely a supervision gain.

Supervision through an aggregator is not a flaw. LLM-Rubric is designed around it. It asks an LLM nine rubric questions about a dialogue, collects the probability distribution over answers to each, and trains a small network with judge-specific and judge-independent parameters to predict individual human judges' ratings of overall satisfaction; it reports root-mean-square error below 0.5 on a four-point scale, about twice as good as the uncalibrated baseline (Hashemi et al., 2024). Its authors are clear that the calibration network, which needs human data, is part of the method, and the anchor classes it as calibration rather than decomposition for exactly that reason (Steindl et al., 2026). AutoCalibrate takes a different route to the same end: it drafts scoring criteria from in-context human-labelled examples and refines them, encoding human preference through the labels rather than through a regressor (Liu et al., 2023b). The same structure appears in reward modelling. ArmoRM predicts nineteen interpretable attribute scores and learns a gating network that mixes them into one reward from preference data (Wang et al., 2024). RocketEval also offers a supervised predictor that reweights checklist items to fit annotations, and its authors report that the supervised version gives higher list-level correlations than the unsupervised one, with "more significant improvements in list-level correlation" than in instance-level agreement (Wei et al., 2025). The advantage does not hold for every judge: Gemma-2-2B reaches 0.965 Spearman on WildBench with and without supervision, and at the instance level on MT-Bench its unsupervised agreement is 57.9 percent against 57.3 supervised, and Llama-3-8B's 63.8 against 62.9. The RocketEval figures cited in Sections 6 and 7 are the unsupervised ones.

The consequence for attribution is direct. Whenever a decomposed pipeline's aggregator has seen human labels and the direct baseline has not, the comparison confounds decomposition with supervision. The anchor's design removes the confound by giving both sides the labels. Very few other comparisons in the record do. CheckEval and TICK avoid it from the other side, since neither uses human labels in aggregation, and their comparisons therefore do not suffer from this channel. They are exposed to others.

6. Channel two: resolution, ties and the level of comparison

A judge that outputs an integer from 1 to 5 produces many ties on a benchmark of hundreds of texts. A decomposed pipeline that averages ten binary answers, or a regressor that outputs a real number, produces almost none. The meta-evaluation statistic then decides whether that difference matters, and most of the statistics in use are sensitive to it.

The problem was named in the paper that made LLM judges standard. G-Eval's authors observed that models "usually only output integer scores, even when the prompt explicitly requests decimal values", that this produces many ties, and that weighting each score by its token probability yields a continuous score that correlates better (Liu et al., 2023a). In machine-translation meta-evaluation, Deutsch, Foster and Freitag showed that the common variants of Kendall's tau handle ties inconsistently, that the handling can be exploited to improve a metric's apparent quality, and that a fair comparison between metrics that predict ties and metrics that do not needs either credit for correct ties or a calibration that introduces ties into the continuous metric (Deutsch et al., 2023). Their concern was LLM-based metrics and error-count metrics whose small discrete ranges produce many ties.

The anchor quantifies the effect for its setting. Ground truth on SummEval and TopicalChat is the mean of three annotators and so takes fractional values; a regressor can match a fractional target, an integer judge cannot; and floats remove ties, which the alternative test counts as wins and Spearman assigns average ranks. Rounding all outputs to integers lowers correlations slightly and raises alternative-test scores, and the authors round throughout so that small numerical artefacts cannot inflate conclusions (Steindl et al., 2026, Table 5). They identify the regressor's floating-point output as a further reason HD-Eval's aggregation step appeared to help.

The same mechanism operates in methods that report no regressor. CheckEval's score is a proportion of yes answers, a value on a grid much finer than a five-point scale when the checklist is long (Lee et al., 2025). DnA-Eval combines aspect scores by a weighted sum computed outside the model and compares the two sums to decide a pairwise winner; its headline is agreement with humans rising from 42.4 to 59.1 percent for GPT-4 on FairEval, a relative gain of 39.6 percent (Li et al., 2024b). Three of its four benchmarks contain pairs that humans rated as ties, and the paper reports agreement twice, with those pairs and without them. The headline pair of numbers is the one without them. With tied pairs included, the same comparison on FairEval is 46.3 against 51.3 percent, a gain of 5.0 points rather than 16.7. On MT-Bench the GPT-4 comparison changes sign with the tie convention: direct scoring leads with tied pairs included, 67.8 against 65.3, and trails without them, 74.4 against 78.4 (Li et al., 2024b, Table 2). This pattern fits a resolution effect. A direct judge that scores two responses on a five-point scale and declares a tie when the integers match will tie more often than a weighted sum of several aspect scores; on pairs where humans expressed a preference, every such tie counts as an error, while on pairs where humans tied, it can count as a success. The paper's ablation also shows that the external calculator beats asking the model to aggregate in its own output, which the authors read as a limitation of models at arithmetic (Li et al., 2024b). How much of the gain over direct scoring is a change in tie frequency, rather than better judgment, is not separated in the published comparison, but the size of the gain depends strongly on how ties are counted.

Resolution interacts with a second choice, the level at which agreement is measured. RocketEval's 0.965 Spearman correlation, obtained with Gemma-2-2B as judge, is a correlation between its ranking of test models on WildBench and their Chatbot Arena Elo ranking, where each model's score averages over many queries. The same 2B judge's instance-level agreement with human judgments on MT-Bench, the paper's other benchmark, is 57.9 percent, against 64.7 percent between humans (Wei et al., 2025). The two numbers come from different benchmarks and cannot be subtracted, but they show how differently one judge can read at the two levels. GEMBA-MQM's strongest claim is likewise about system-ranking accuracy (Kocmi and Federmann, 2023). Even the computation of an instance-level correlation is not standardised: Chiang and Lee describe two ways of computing it on these benchmarks, over the whole dataset or per source document, and report that recomputing G-Eval's correlations from its released data and code gave numbers "quite different" from those in the G-Eval paper (Chiang and Lee, 2023). A decomposition that makes each instance's score less noisy will help most at the instance level; one that shifts every system's scores by a similar amount will look strong at the system level and do little for any single verdict. The two claims are both legitimate. They are different claims.

7. Channel three: information the direct judge did not have

The largest channel, where it can be measured, is not about splitting at all. A decomposition is a piece of text, and writing it often adds information. A checklist generated from an instruction makes the instruction's implicit requirements explicit. A rubric generated from a reference answer carries the reference answer's content into the judge's prompt in another form. Atomic claims checked against Wikipedia or search results bring in a knowledge source. A direct judge that lacks the reference, the retrieved evidence or the strong model's reading of the instruction is answering a harder question, and a comparison against it measures the information as well as the decomposition.

FLASK's ablation separates these two. Its skill-agnostic baseline is reference-guided: the judge sees the question, a ground-truth answer and the response, and gives one score. The full FLASK protocol adds skill-specific rubrics and per-skill scoring, and it also selects for each instance the skills that apply to it, since it "first annotates fine-grained metadata for each instruction" (Ye et al., 2023). With GPT-4 as judge, the skill-agnostic baseline correlates with human labelers at Spearman 0.641 and FLASK at 0.680, a gain of 0.039 for the decomposition together with the per-instance skill selection. Removing the reference answer from FLASK drops it to 0.516, a loss of 0.164; removing the rationale drops it to 0.634, and removing the score rubric to 0.646 (Ye et al., 2023, Table 1; Figure 1b). The correlations are computed against per-skill human scores, and the holistic baseline is fitted to them by assumption: "For coarse-grained evaluation setting, we assume that a uniform score has been assigned for every skill for correlation calculation" (Ye et al., 2023). The ground truth is thus itself decomposed, which may favour the decomposed judge. Even so, in this study the reference answer is worth about four times what the skill decomposition is worth, and a decomposed judge without the reference is far worse than a holistic judge with it.

Rubrics as Rewards shows the information term for a reward signal with no decomposition at all. Every rubric condition in its Table 1 is the implicit variant, in which "All rubric criteria along with categorical weights are passed to an LLM-as-judge, delegating the aggregation to the model itself to produce a single scalar reward", and the judge is told to "Consider all rubrics holistically when determining your score" (Gunjal et al., 2025). Under the definitions of Section 2 that is direct prediction with an instance-specific rubric in context. Policies trained against a direct Likert judge reach 23.9 percent on HealthBench. Giving the Likert judge a reference answer raises this to 31.7. A holistic judge given a rubric

generated without the reference answer reaches 32.0; one given a rubric generated from the reference answer reaches 35.9; one given human-written rubrics reaches 34.8 (Gunjal et al., 2025, Table 1; Figure 1c). The authors draw the conclusion that expert grounding matters for rubric generation, and their numbers support a sharper one. Of the 12.0-point gain from the direct Likert reward to the reference-grounded rubric, 7.8 points are available to a judge that sees the reference and scores holistically. A rubric without the reference recovers about the same, 8.1 points, presumably because the stronger model that writes it supplies content of its own. Instance-specific rubric text on top of the reference adds the remaining 4.2 points over the reference-guided Likert judge, with the judge still returning one score: an effect of information or format, not of splitting the judgment. The same paper also scores the same instance-specific rubrics item by item and combines the verdicts with a fixed weighted sum (RaR-Explicit). That is the nearest existing case of the matched-information comparison proposed in Section 14, run for rewards, and the holistic reading wins: "Between the two rubric-guided methods, RaR-Implicit attains the strongest results overall; fixed weighted sums in RaR-Explicit offer more control but can be brittle" (Gunjal et al., 2025). These are policy scores after training, a downstream outcome rather than judge agreement, and one benchmark; they show the size of the information term, not its generality.

TICK's three conditions decompose its own gain in a similar way. Direct scoring agrees exactly with human preferences 46.4 percent of the time, putting the generated checklist in the prompt without requiring answers raises this to 48.7, and answering each item explicitly to 52.2 (Cook et al., 2024). About 40 percent of the gain is available from showing the judge the checklist as information. The rest comes with explicit item-by-item answers, which may reflect simplification, the finer score grid of a pass rate (Section 6), or both; TICK's design does not separate them.

The pattern repeats wherever the reference is varied. In FBI's perturbation study, evaluator models missed deliberate quality drops in more than half of cases on average, and reference-guided evaluation did better than both single-answer and pairwise evaluation. In single-answer scoring, GPT-4-turbo did better with simple prompts than with prompts that add explicit rubrics or named axes of evaluation, which the authors read as extra criteria possibly making an evaluation more thorough without necessarily improving detection of subtler errors; in pairwise comparison the result reversed, and "advanced strategies perform better than the basic strategies" (Doddapaneni et al., 2024). Prometheus is described by its authors as matching GPT-4 "when the appropriate reference materials (reference answer, score rubric) are accompanied" (Kim et al., 2023). BiGGen Bench is evidence for this channel rather than for decomposition: each of its instances carries a reference answer and one rubric of its own, scored once on a 1-to-5 scale, so its finding that instance-specific criteria track human judgments better than coarse or domain-level ones is a gain from information given to a holistic judge (Kim et al., 2024). In checklist-based rewards, the checklists are written from each instruction by Qwen2.5-72B-Instruct, ten times the size of the 7B policy being trained, and some items are verified by programs (Viswanathan et al., 2025). In each case a stronger model or a human has put information into the decomposition before the judge ever sees the response.

RocketEval contains the cleanest test of this channel inside a single positive paper. Its checklists are written by GPT-4o for each query, with five to ten questions per instance, and graded by small models, so as to give a lightweight judge "essential knowledge and critical focus areas" it could not construct for itself (Wei et al., 2025). To test whether the instance-level content matters, the authors added a baseline that uses six fixed questions as the checklist for every query. On instance-level agreement with MT-Bench human judgments, the fixed checklist gives Gemma-2-2B 43.9 percent against 39.2 for direct scoring, and the instance-level checklist 57.9; for Llama-3-8B the fixed checklist is worse than direct scoring, 46.9 against 51.4, while the instance-level checklist reaches 63.8. The authors' summary is that the fixed checklist "shows no notable performance enhancement and may even degrade performance" (Wei et al., 2025). Generic decomposition, the treatment the anchor tested, did little here too. Content written for the instance by a much stronger model did a great deal.

This is the channel the anchor's design closes most completely. Its decompositions are generic: sub-criteria of coherence are the same for every summary, and the L1-agnostic variant is generic even with respect to the criterion. No reference, no retrieval, no instance-specific content enters. In that setting the information channel is shut, and the remaining effect of decomposition is small and of unstable sign. That is consistent with the information channel being where most of the positive literature's gain lives. It does not prove it, because most instance-specific studies do not hold information fixed while removing decomposition. Of the three that come closest, two (FLASK and RocketEval's fixed-checklist control) find the information term larger than the decomposition term, and RocketEval finds generic decomposition alone worth little or nothing. The third, TICK, attributes about 40 percent of its gain to the checklist shown as information, leaving the rest to explicit item answers whose resolution and simplification effects it does not separate. Rubrics as Rewards does not enter this count, because all of its rubric conditions end in one holistic score; what it adds is a different comparison, the same instance-specific rubric read holistically or scored item by item, in which

the holistic reading gave the stronger policies.

8. Channel four: verifiability, and its costs

Some decompositions turn a judgment no external tool can make into many judgments one can. A biography's overall accuracy cannot be looked up; whether a named person was born in a named year can. Here decomposition is not a prompt engineering choice but a precondition for using a verifier at all, and the evidence for it is the strongest in the record. FActScore's estimator tracks human factual-precision scores within 2 percent and was used to score 6,500 generations from 13 models that would have cost about 26,000 dollars to have humans annotate (Min et al., 2023). SAFE agrees with crowd annotators on 72 percent of individual facts and, on a sample of 100 disagreements, is judged right 76 percent of the time (Wei et al., 2024a).

The same studies and their successors document what decomposition costs in this setting, and the costs are structural. Wanner et al. show that FActScore's value depends on which decomposition method produces the claims, so that a change of decomposer can be read as a change in the generator's factuality (Wanner et al., 2024). Jiang et al. show that the decompose-then-verify framework can be gamed by padding a text with obvious or repeated true claims, which inflate the supported fraction without adding information (Jiang et al., 2024). VeriScore's authors point out that FActScore and SAFE try to verify everything, including content such as analogies and hypotheticals that cannot be verified, and so penalise texts that contain it (Song et al., 2024b). Gunjal and Durrett argue that fully atomic claims lose the context needed to verify them and propose "molecular" facts that add the minimum context back (Gunjal and Durrett, 2024). A 2026 study finds that the decomposer can substitute its own belief for what the source text says, producing a claim that contradicts the passage it was meant to restate; the natural rate is low, between 0.2 and 0.4 percent of atomic claims, but the errors are stable across resamples rather than random (Yen, 2026).

Whether decomposition helps verification depends on the verifier. Hu, Long and Wang ran the same decompose-then-verify pipeline with verifiers of different strength and report that "decomposition generally benefits weaker verifiers, while it tends to negatively affect stronger verification systems", attributing the pattern to a trade-off between the accuracy gained on simpler sub-claims and the noise added by decomposition and retrieval (Hu et al., 2025). Akhter et al. separate the evidence side: with evidence aligned to each sub-claim, decomposition helps; with the usual setup, in which every sub-claim is checked against the same claim-level evidence, it fails to help and often hurts, and decomposition improves claim verification under oracle sub-claim labels but can be neutral or harmful when those labels are predicted (Akhter et al., 2026).

The verifiability channel therefore has a clear signature. Decomposition pays when it converts a judgment into units an external verifier is better at than the judge, when evidence can be aligned to those units, and when the verifier is weak enough that the simpler units are worth their noise. A strong verifier with whole-claim evidence gains little and may lose. None of this applies to the anchor's setting, where no verifier exists for "coherence" and its sub-criteria are judged by the same model as the original criterion.

9. Channel five: reliability is not validity

Several of the most widely cited decomposition gains are gains in agreement, but agreement among judges rather than with humans. CheckEval's headline is an increase of 0.45 in average agreement across evaluator models and a reduction in score variance; its abstract says CheckEval "strongly correlates with human judgments" but presents the agreement gain as the more important result ("More importantly"), and its introduction describes the correlation with humans as "competitive" (Lee et al., 2025). The agreement gains are large. Over all twelve judge models on SummEval, Krippendorff's alpha rises from 0.09 for G-Eval to 0.48 for CheckEval; within the group of large models it reaches 0.67, which the authors compare with human inter-annotator agreement on SummEval of about 0.7. The authors also rule out the most obvious artefact: binarising G-Eval's Likert scores raises its alpha only to about 0.11 on SummEval, far below CheckEval's 0.48 (Lee et al., 2025, Appendix C.3). Variance across judge models in Spearman correlation falls from 0.0100 to 0.0019.

These are reliability results and they are real. They do not establish validity. Two judges that answer the same narrow yes-or-no question will agree more often than two judges asked for a holistic 1-to-5 score, whether or not the question captures what the human annotators were rating. Highly consistent judges can still share an error, and the strongest evidence that a checklist score can rise without quality rising comes from rubric-based reinforcement learning (Section 10). The mean correlation gain CheckEval reports over G-Eval, from 0.399 to 0.481 on SummEval, is against an unsupervised zero-shot Likert baseline, and the anchor's reimplementation of CheckEval with Claude 4 scored below every direct-prediction variant in the anchor's own runs, 0.411 against 0.545 to 0.560 on SummEval (Steindl et al., 2026).

Human evaluation shows the same separation. InFoBench reports that annotators agree more with each other under its decomposed requirement ratio than under direct scoring (Qin et al., 2024), and TICK reports that giving human raters generated checklists raises their inter-annotator agreement on WildBench from 0.194 to 0.256 (Cook et al., 2024). Neither measures whether the

more consistent ratings are closer to a criterion of quality outside the ratings themselves. The MQM study is the one case in the record where a decomposed human protocol was compared against a holistic one under matched expertise. Freitag et al. collected scalar quality ratings from professional translators (pSQM) as well as MQM error annotations from a disjoint pool of professionals, and report that "pSQM has high system-level correlation with MQM"; it was the crowd ratings used in the shared task that disagreed with MQM (Freitag et al., 2021). The striking difference between MQM and the shared-task rankings was carried mostly by who rated, not by whether the rating was decomposed. The decomposed protocol did add something below the system level: segment-level correlation between pSQM and MQM was only about 0.5 Kendall tau, and two Chinese-to-English systems that pSQM placed 0.04 apart were 0.19 apart under MQM, which led the authors to recommend MQM where systems are close and pSQM as sufficient for a campaign like the shared task (Freitag et al., 2021). SummEval's own authors observed the same contrast between expert and crowd annotations of the summaries the anchor uses (Fabbri et al., 2021).

The checklist-usefulness study of Furuhashi et al. is the closest LLM analogue to this control. Across six ways of generating checklists and eight judge models (GPT-4o and seven open models from 7B to 32B parameters), selective use of checklists produced statistically significant differences in 20 of 48 pairwise-comparison cases, but for direct scoring "no statistically significant differences across any of the six checklist-generation policies evaluated with eight evaluation models" (Furuhashi et al., 2025). They also found that checklist items with low correlation to human scores often overlapped with items humans had written, which they read as evidence of inconsistency in the human ratings themselves.

The distinction matters because the two properties serve different users. A benchmark maintainer who reruns an evaluation with a new judge model wants reliability, and CheckEval delivers it. A researcher who reports that a method produces better summaries wants validity, and reliability does not supply it. A decomposition paper that reports both should say which one its headline is about.

10. Channel six: execution, aggregation and interference

The last channel concerns how a decomposed evaluation is run, and here the evidence cuts against decomposition as often as for it.

The first issue is whether items are judged together or apart. CheckEval asks multiple checklist questions in one call for efficiency, and reports a pilot control: "We compared single-question and multi-question inference in our pilot experiments and found no noticeable difference in performance" (Lee et al., 2025). Other studies find that joint execution is not innocent. Yu et al. show that when a judge evaluates several rubrics in one pass, its verdict on each depends on which others are present. On HealthBench, comparing each rubric's verdict in isolation with its verdict in a joint pass, Qwen3-32B reaches only 36 percent exact match at the sample level, while every model tested agrees with itself at least 98 percent of the time across random seeds, so the effect is systematic rather than sampling noise (Yu et al., 2026). RubricEval measures the accuracy consequence. With reasoning enabled, judging each rubric item in a separate call reaches balanced accuracy of 77.38 and 82.17 percent for its two judges, against 69.90 and 70.44 percent when all items are judged in one pass, a gap of 7 to 12 points; even GPT-4o reaches only 55.97 percent on its hard subset (Pan et al., 2026). DnA-Eval scores each aspect in a separate call for the same reason, citing anchoring among scores generated together (Li et al., 2024b). CheckEval's pilot is one setting in which the difference did not show; in the rubric settings of Yu et al. and Pan et al. it did. A decomposition run jointly is therefore not necessarily the same treatment as one run in isolation, and isolation multiplies the number of calls.

The second issue is who does the arithmetic. DnA-Eval reports that computing the weighted sum outside the model beats asking the model to combine its own aspect scores for nearly all models and benchmarks, and draws the conclusion that models are limited at "mathematical aggregation during the evaluation process" (Li et al., 2024b). That is a gain from moving a computation out of the model, available to any pipeline that produces intermediate numbers.

The third issue is what additive aggregation does under optimisation. A judge used once for evaluation faces no adversary; a judge used as a reward does. Liu et al. (2026) trained policies with checklist rewards that sum per-criterion credit and found that answer quality collapsed while the score rose, which they trace to the fixed price each criterion pays, so that cheap behaviours such as naming an item earn reward whether or not the item is satisfied. They state the diagnosis plainly: "We show that this additive aggregation is the weak point." Grouping the same criteria verbatim into conjunctive units recovers a third of the lost appropriateness, and scoring them holistically as a single list recovers nearly 40 percent. Mahmoud et al. (2026) study the same setting with a cross-family panel of reference judges and report that weak verifiers yield large proxy gains that do not transfer, that exploitation concentrates in partial satisfaction of compound criteria, and that even with a strong verifier, when the rubric leaves failure modes unspecified, "rubric-based verifiers prefer the RL checkpoint, while rubric-free judges prefer the base model." PReMISE frames rubrics as measurement specifications and reports that a vague rubric lets exploit

responses receive high scores 46.4 percent of the time, reduced to 36.0 percent by refinement (Roy et al., 2026). Bagaria et al. go further and show that classifiers trained on rubric text alone, with no access to the evaluated response, predict an LLM judge's verdicts above chance and sometimes above 0.8, and that when an answer is edited to reverse whether a rubric is satisfied, the judge's verdict reverses in only 37.7 percent of pairs (Bagaria et al., 2026).

These failures are not arguments against decomposition in evaluation as such. They are arguments that the same decomposition that makes a judgment more consistent also gives an optimiser a more legible target, and that the aggregation rule, not the items, is often what the optimiser exploits. Holistic judgment of the same criteria, in the reward setting, was the more robust of the two. The anchor's question, whether decomposition helps a fixed judge track fixed human ratings, is the benign case of a problem that becomes adversarial when the score is optimised.

A seventh difference runs across all six channels: the strength of the baseline. Several of the largest reported gains are against weak direct judges. CheckEval's comparison is with G-Eval, which uses auto-generated evaluation steps and no rated examples (Lee et al., 2025). Chiang and Lee found that G-Eval's auto chain-of-thought does not always help, and that simply asking the model to explain its rating improves correlation on both SummEval and TopicalChat (Chiang and Lee, 2023). BSM's gains of up to 26 points of agreement with LLaMA-2-70B-chat are against zero-shot prompting and self-consistency, while BSM applied to GPT-4 adds 3 points (Saha et al., 2024). DnA-Eval compares against direct scoring and chain-of-thought, and confirms the earlier finding that chain-of-thought alone does not reliably help (Li et al., 2024b). The anchor's direct judge, by contrast, combines a criterion definition, a reasoning instruction, rated examples and, in some variants, a calibration regressor. It is a direct judge built with the care usually spent on the decomposed one, and against it the decomposition gain mostly vanishes.

11. Consolidated evidence

Table 1 collects the measured values the argument rests on, one row per published result, with the channel the comparison leaves open or isolates. A channel listed as "open" is a difference between the compared conditions that the published design does not hold fixed; it is not a claim that the channel produced the result.

Three patterns stand out in Table 1. First, the two rows that come closest to varying only the decomposition, with supervision, reference and baseline held fixed, show small effects: 0.039 Spearman in FLASK's reference-guided comparison, where per-instance skill selection also changes, and a lead of a few hundredths on SummEval that reverses on TopicalChat in the anchor. Second, the rows that vary information show the largest effects in the table: 0.164 Spearman for FLASK's reference answer, 7.8 HealthBench points for a reference given to a holistic judge. Third, the rows that measure decomposition under execution or optimisation pressure are mostly negative, and several of the negative effects are comparable in size to the positive ones elsewhere.

12. An attribution procedure

Algorithm 1 states the checks a claim of the form "decomposing criterion c improved judge J on benchmark B" should pass before the gain is credited to decomposition. The steps are ordered so that the cheapest checks, which dissolve most published claims, come first.

Three points about the procedure. First, steps 1 to 3 are what the anchor did for supervision and what FLASK did for the reference answer. Most published comparisons fail step 1: the decomposed pipeline receives a checklist written by a stronger model, a reference answer, or retrieved evidence, and the direct baseline receives none of these. That does not make the decomposed pipeline worse, but its gain is not a decomposition gain until the direct pipeline gets the same inputs.

Second, steps 4 to 7 address the measurement channels. Rounding and tie-aware statistics follow the anchor and Deutsch et al. (2023). Reporting both levels follows from the gap between RocketEval's system-level and instance-level numbers (Wei et al., 2025). Isolation versus joint execution follows from Yu et al. (2026) and Pan et al. (2026). Intervals follow from the anchor's single-run design, whose dataset-dependent sign could be noise or could be real.

Third, steps 8 to 13 are the anchor's L1-agnostic control generalised. A decomposition gain that survives steps 1 to 7 could still come from having more LLM-generated features rather than from decomposing the criterion. The agnostic control asks whether items chosen without knowledge of the criterion do as well. If they do, the gain is a feature gain. Only a gain that survives both comparisons supports the motivating claim that splitting the criterion into its parts made the judgment easier.

Applied to the record, the procedure stops most claims early. HD-Eval's gain stops at step 3, as the anchor showed. TICK's stops at step 1 for about 40 percent of its size, since the checklist itself carries information the direct judge lacks; the rest needs steps 4 and 6. FLASK's comparison passes steps 1 and 2 for the reference and measures a gain of 0.039 for decomposition together with per-instance skill selection, at one seed on one judge, which needs step 7. CheckEval's correlation gain passes step 1, since neither side receives labels or references, but it was measured against a zero-shot baseline; against the anchor's stronger direct judge, the reimplemented CheckEval falls behind, so the gain does not survive a change of baseline. Its reliability gain

Table 1. Published comparisons between decomposed and direct LLM evaluation, with the channel each isolates or leaves open. Sp. is Spearman correlation with human ratings. All values are as reported by the cited source.

Source	Setting	Decomposed	Comparison	Reported value	Channel
Steindl et al. 2026	SummEval, Claude 4	AOI, labels	direct, labels	Sp. 0.586 vs 0.560	supervision held equal
Steindl et al. 2026	TopicalChat, Claude 4	ICL, labels	direct, labels	Sp. 0.575 vs 0.687	supervision held equal
Steindl et al. 2026	SummEval / TopicalChat	L1-agnostic	targeted AOI	Sp. 0.579 / 0.621 vs 0.608 / 0.622	simplification tested
Steindl et al. 2026	SEAHORSE, 5 criteria	AOI	direct	mean alpha 0.38 vs 0.43	supervision held equal
Liu et al. 2024	SummEval, 50% data	HD-Eval	no layers / mean aggregator	Sp. 0.535 vs 0.513 / 0.530	supervision open
Lee et al. 2025	SummEval, 12 judges	CheckEval	G-Eval	mean Sp. 0.481 vs 0.399	baseline open
Lee et al. 2025	SummEval, all judges	CheckEval	G-Eval / binarised	alpha 0.48 vs 0.09 / 0.11	reliability, not validity
Cook et al. 2024	internal pairwise	TICK	direct; checklist in prompt	52.2% vs 46.4%; 48.7%	information open
Ye et al. 2023	FLASK, GPT-4	skill rubrics	reference-guided holistic	Sp. 0.680 vs 0.641	decomposition plus instance skill selection
Ye et al. 2023	FLASK, GPT-4	skill rubrics	same, no reference	Sp. 0.680 vs 0.516	information isolated
Gunjal et al. 2025	HealthBench policy	rubric in context, one holistic score	Likert / Likert + ref.	35.9 vs 23.9 / 31.7	information isolated
Li et al. 2024b	FairEval, GPT-4	DnA-Eval	direct scoring	59.1% vs 42.4% (no ties); 51.3% vs 46.3% (ties)	resolution open
Li et al. 2024b	MT-Bench, GPT-4	DnA-Eval	direct scoring	78.4% vs 74.4% (no ties); 65.3% vs 67.8% (ties)	resolution open
Wei et al. 2025	WildBench, Gemma-2-2B	RocketEval	human-derived ranking	Sp. 0.965 (system level)	level
Wei et al. 2025	MT-Bench, Gemma-2-2B	instance checklist	direct; fixed checklist	57.9% vs 39.2%; 43.9%	information isolated
Furuhashi et al. 2025	InFoBench / LLMBar	selective checklists	none / all	0 of 48 / 20 of 48 significant	decomposition tested
Freitag et al. 2021	WMT 2020, humans	MQM	professional scalar	"high system-level correlation"	expertise isolated
Hu et al. 2025	fact-checking	claim decomposition	whole claim	helps weak, hurts strong	verifiability
Akhter et al. 2026	claim verification	sub-claims	whole claim	gains only with aligned evidence	verifiability
Wei et al. 2024a	16k facts	SAFE	crowd labels	72% agree; 76% of 100 wins	verifiability
Yu et al. 2026	HealthBench	joint rubrics	isolated rubrics	36% sample-level exact match (rubric-level agreement 0.887)	execution
Pan et al. 2026	RubricEval	one pass	per-item calls	69.90 / 70.44 vs 77.38 / 82.17	execution
Liu et al. 2026	rubric RL	additive items	holistic list	holistic recovers ∼40% of loss	aggregation
Mahmoud et al. 2026	rubric RL	rubric verifier	rubric-free judges	opposite preferences	aggregation

belongs under step 14. FActScore and SAFE are not decomposition claims in this sense at all: their gain is the verifier's, and the decomposition is the price of using one. The anchor passes steps 1 to 4 and 8, and meets the multi-judge part of step 7 with three judge models, but reports a single run per judge and no intervals over test items, which is why its null cannot distinguish a small positive SummEval effect from noise.

13. The strongest case against this reading

"The anchor tested two old methods on two old benchmarks. Modern decomposition is instance-specific and the anchor says nothing about it." This is correct, and Section 3 says so. The anchor's authors excluded instance-level checklists explicitly, citing RocketEval as an example (Steindl et al., 2026). The objection does not rescue instance-specific decomposition, though. It moves the question to whether instance-specific gains come from decomposition or from the information instance-specific text carries. Two of the three studies that come closest to separating these find the information term larger (Section 7). One study located here gives a judge instance-specific rubrics as context, asks for one holistic score, and compares that with item-by-item scoring of the same rubrics: Rubrics as Rewards, where the comparison is between training rewards rather than against human ratings, resolution is not matched, and the holistic judge gives the stronger policies (Gunjal et al., 2025). No study located here makes that comparison against human ratings at

Algorithm 1. Attributing a gain to decomposition in LLM evaluation

```
Input: judge J, criterion c, benchmark B with human ratings H,
       decomposed pipeline D, direct pipeline P
1  list what D receives that P does not: human labels, reference,
       retrieved evidence, instance-specific text, verifier calls
2  for each item in the list: give it to P, or remove it from D
3  if D's aggregator is trained on H: fit the same model on P's score
4  score D and P with outputs rounded to the scale of H, and with
       a tie-aware statistic; report both
5  report instance-level and system-level agreement separately
6  run D with items in isolated calls and in one joint call
7  repeat steps 2-6 over >= 3 seeds and >= 2 judges; report intervals
8  replace D's items with items generated without c (agnostic control)
9  delta_simplify <- score(D) - score(P) after steps 1-7
10 delta_target <- score(D) - score(agnostic control)
11 if delta_simplify within interval of 0: label "no decomposition gain"
12 elif delta_target within interval of 0: label "feature gain, not
       criterion decomposition"
13 else: label "decomposition gain", scoped to (J, c, B)
14 separately report agreement across judges as reliability, not validity
15 if the score will be optimised: test with a rubric-free judge after
       training, and report the gap
```

matched information and matched resolution; TICK's checklist-in-prompt condition comes nearest, for pairwise preference, without controlling resolution.

"**Reliability is the point. A judge that two labs can reproduce is worth more than one with slightly higher correlation.**" For some uses this is right, and Section 9 credits CheckEval with a large reliability gain. But reliability without validity can be bought cheaply: any narrowing of the question raises agreement. The reward-hacking results in Section 10 show what a consistent but incomplete instrument does under optimisation. A reliability claim should be reported as one, with validity reported beside it.

"**Decomposition makes evaluation interpretable, and interpretability is worth a small cost in correlation.**" This is a reasonable value judgment and outside what this paper evaluates. Interpretability is the most common justification in the methods papers (Lee et al., 2025; Cho et al., 2026; Li et al., 2024b). The relevant evidence is that interpretable items are not necessarily the items the judge's verdict depends on: rubric text alone predicts verdicts (Bagaria et al., 2026), and items judged jointly shift each other (Yu et al., 2026). An explanation that does not track the decision is less interpretable than it looks.

"**Human annotators decompose too. MQM shows experts reasoning about errors is the gold standard.**" MQM is a strong protocol, and its advantage over crowd ratings is real (Freitag et al., 2021). The same study, however, found professional holistic ratings largely matching MQM at the system level. The comparison that would credit decomposition to human raters is the expert-versus-expert one, and in it the decomposed protocol's advantage appeared mainly at the segment level and in separating closely matched systems.

"**The SummEval lead for decomposition is consistent across three judges. That is a positive result the paper is underplaying.**" It is consistent in sign across three judges, from 0.026 to 0.071, and the paper reports it as such (Figure 1a). It is also unaccompanied by any variance estimate, it comes mostly from the variants the anchor itself proposed (AOI and ICL), and it reverses on the second benchmark for two of the three judges. A consistent small lead on one benchmark and an inconsistent large deficit on another is not evidence that decomposition helps, and it is not evidence that it never does. It is a reason to run step 7 of Algorithm 1.

14. Studies that would settle the open part

Five studies, each feasible with public benchmarks and open judge models, would close most of what remains.

An information-matched instance-specific comparison. Take TICK's instance-specific checklists and give them to a direct judge as context, asking for one holistic score, against the same judge answering each item. BiGGen Bench already supplies the holistic arm, one instance-specific rubric scored once, and would need an item-by-item arm built from the same rubric. Hold the reference answer, if any, fixed across both arms. Rubrics as Rewards ran this comparison for training rewards, where the holistic arm won (Gunjal et al., 2025); it has not been run against human ratings. Round both to the human scale and report a tie-aware statistic. This isolates the decomposition term in the setting where the positive literature is strongest.

The anchor with intervals and a third benchmark. Rerun the anchor's main comparison with several seeds per judge and bootstrap intervals over test items, on SummEval, TopicalChat and one benchmark with more items per criterion. If the SummEval lead survives, there is a small decomposition effect for summary quality; if not, the null is uniform.

Human ratings of the sub-criteria. The anchor's own suggested future work: collect human ratings for the decomposed criteria as well as the originals (Steindl et al., 2026). This would show whether the LLM scores the

sub-criteria correctly and whether the human L1 rating is a function of human L2 ratings. If it is not, criterion decomposition is measuring a different construct, and a gain or loss on L1 is not about simplification.

Matched-expertise human protocols. Repeat the MQM-versus-pSQM comparison for a summarization or dialogue benchmark: the same pool of expert annotators, half using a decomposed protocol and half a holistic one, with instance-level agreement against a third expert pool. This would answer for human raters the question the anchor answers for LLMs.

Holistic versus decomposed rewards under a fixed information budget. Train policies against a reference-guided holistic judge and against reference-grounded rubrics at matched judge cost, and evaluate with a rubric-free panel. Gunjal et al. (2025) and Liu et al. (2026) together suggest the holistic judge captures most of the gain with less exploitability; a direct test would settle it.

15. What is not known

It is not known whether decomposing a holistic quality criterion into generic sub-criteria improves a well-prompted LLM judge at all. The one controlled study found a small lead on one benchmark and a large deficit on another, without intervals.

It is not known how much of the gain from instance-specific checklists is due to decomposition rather than to the information the checklist carries. Partial ablations put the information share at about 40 percent in one study and the decomposition share at about a quarter of the reference's contribution in another.

It is not known whether gains in agreement across LLM judges from checklist evaluation come with gains in agreement with humans that a direct judge with matched inputs could not achieve.

It is not known how much of any published gain from aggregated decomposed scores is due to tie elimination. Only the anchor reports rounded results, and only for its own methods.

It is not known whether human holistic ratings of summaries and dialogues are functions of the sub-criteria LLM decompositions produce. No study located here has collected human ratings at both levels for the same texts.

16. Limitations

This analysis reads published papers and inherits their designs. The anchor is a recent preprint accepted to EMNLP 2026 according to its arXiv record, read in its first version; later versions may add the variance estimates whose absence Section 3 notes. Several sources are 2026 preprints that have not been peer reviewed, including the rubric-interference, rubric-artefact, reward-hacking and decomposition-induced conflict studies, and the paper uses them for the existence and direction of effects rather than for their exact sizes. The channel decomposition in Sections 5 to 10 is a proposal for organising evidence. The channels are not independent: a reference-grounded rubric carries information, changes resolution and is executed item by item, so the ablations that separate information from decomposition are informative only for their own settings. The figure and table transcribe values from different metrics (Spearman correlation, agreement rates, balanced accuracy, policy benchmark scores) that should not be compared across rows. Finally, the paper does not evaluate interpretability, cost or latency, which are legitimate reasons to prefer decomposition that do not depend on its validity.

17. Conclusion

"Decomposition helps LLM evaluation" is a family of claims carrying at least six hidden channels. Supervision explains the method the controlled study set out to test. Resolution and the level of comparison explain part of the gap between headline correlations and the quality of individual verdicts. Information explains the largest gains wherever it has been separated: a reference answer, a stronger model's reading of the instruction, or retrieved evidence makes the judge better whether or not the judgment is split. Verifiability explains the factuality results, where decomposition is the price of using a verifier and is worth paying only when the verifier is weak and the evidence aligned. Reliability explains the large agreement gains among judges, which are real and are not validity. Execution and aggregation explain both a gain, when arithmetic moves outside the model, and the most serious losses, when items interfere or when an optimiser exploits an additive sum.

What remains when all six are held fixed is the effect the methods were motivated by, that simpler questions let the same model judge better from the same information. On the published evidence that effect is small, its sign depends on the benchmark, and it has been measured with intervals nowhere. The controlled null of 2026 is right about its setting and too narrow to settle the general question; the positive literature is right about its gains and mostly wrong about their source. The experiments that would decide between them are cheap, and Algorithm 1 lists them in the order they should be run.

References

Akhter, M. E., Ruggeri, F., Bilal, I. M., Procter, R., & Liakata, M. (2026). The Alignment Bottleneck in Decomposition-Based Claim Verification. arXiv:2602.10380. doi:10.48550/arXiv.2602.10380.

Bagaria, A., Sundaram, S. S., Krishnan, G. S., & Ravindran, B. (2026). Judging LLM-as-a-Judge: Concerning Rubric Artifacts in LLM-based Automated Text Generation Evaluation. arXiv:2609.02942. doi:10.48550/arXiv.2609.02942.

Calderon, N., Reichart, R., & Dror, R. (2025). The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. Proceedings of ACL 2025 (Volume 1: Long Papers), 16051-16081. arXiv:2501.10970.

doi:10.48550/arXiv.2501.10970.

Chiang, C.-H., & Lee, H. (2023). A Closer Look into Automatic Evaluation Using Large Language Models. arXiv:2310.05657. doi:10.48550/arXiv.2310.05657.

Cho, S., Chawla, K., Cai, P., Liu, Z., Zhu, C., Zhang, S.-X., & Sahu, S. (2026). Ask, Don't Judge: Binary Questions for Interpretable LLM Evaluation and Self-Improvement. arXiv:2606.27226. doi:10.48550/arXiv.2606.27226.

Clark, E., Rijhwani, S., Gehrmann, S., Maynez, J., Aharoni, R., Nikolaev, V., Sellam, T., Siddhant, A., et al. (2023). SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation. Proceedings of EMNLP 2023, 9397-9413. arXiv:2305.13194. doi:10.48550/arXiv.2305.13194.

Cook, J., Rocktaschel, T., Foerster, J., Aumiller, D., & Wang, A. (2024). TICKing All the Boxes: Generated Checklists Improve LLM Evaluation and Generation. Language Gamification Workshop, NeurIPS 2024. arXiv:2410.03608. doi:10.48550/arXiv.2410.03608.

Deutsch, D., Foster, G., & Freitag, M. (2023). Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration. arXiv:2305.14324. doi:10.48550/arXiv.2305.14324.

Doddapaneni, S., Khan, M. S. U. R., Verma, S., & Khapra, M. M. (2024). Finding Blind Spots in Evaluator LLMs with Interpretable Checklists. arXiv:2406.13439. doi:10.48550/arXiv.2406.13439.

Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation. Transactions of the Association for Computational Linguistics, 9, 391-409. arXiv:2007.12626. doi:10.48550/arXiv.2007.12626.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. arXiv:2104.14478. doi:10.48550/arXiv.2104.14478.

Furuhashi, M., Nakayama, K., Kodama, T., & Sugawara, S. (2025). Are Checklists Really Useful for Automatic Evaluation of Generative Tasks? arXiv:2508.15218. doi:10.48550/arXiv.2508.15218.

Gao, M., Hu, X., Ruan, J., Pu, X., & Wan, X. (2024). LLM-based NLG Evaluation: Current Status and Challenges. Computational Linguistics, 51(2), 661-687 (2025). arXiv:2402.01383. doi:10.48550/arXiv.2402.01383.

Gunjal, A., & Durrett, G. (2024). Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification. arXiv:2406.20079. doi:10.48550/arXiv.2406.20079.

Gunjal, A., Wang, A., Lau, E., Nath, V., He, Y., Liu, B., & Hendryx, S. (2025). Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains. arXiv:2507.17746. doi:10.48550/arXiv.2507.17746.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. Proceedings of ACL 2024 (Volume 1: Long Papers), 13806-13834. arXiv:2501.00274. doi:10.48550/arXiv.2501.00274.

Hu, Q., Long, Q., & Wang, W. (2025). Decomposition Dilemmas: Does Claim Decomposition Boost or Burden Fact-Checking Performance? NAACL 2025. arXiv:2411.02400. doi:10.48550/arXiv.2411.02400.

Hu, X., Gao, M., Hu, S., Zhang, Y., Chen, Y., Xu, T., & Wan, X. (2024). Are LLM-based Evaluators Confusing NLG Quality Criteria? arXiv:2402.12055. doi:10.48550/arXiv.2402.12055.

Jiang, Z., Zhang, J., Weir, N., Ebner, S., Wanner, M., Sanders, K., Khashabi, D., Liu, A., & Van Durme, B. (2024). Core: Robust Factual Precision with Informative Sub-Claim Identification. arXiv:2407.03572. doi:10.48550/arXiv.2407.03572.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2022). Decomposed Prompting: A Modular Approach for Solving Complex Tasks. arXiv:2210.02406. doi:10.48550/arXiv.2210.02406.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., et al. (2023). Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. arXiv:2310.08491. doi:10.48550/arXiv.2310.08491.

Kim, S., Suk, J., Cho, J. Y., Longpre, S., Kim, C., Yoon, D., Son, G., Cho, Y., et al. (2024). The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models. arXiv:2406.05761. doi:10.48550/arXiv.2406.05761.

Kocmi, T., & Federmann, C. (2023). GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. arXiv:2310.13988. doi:10.48550/arXiv.2310.13988.

Lee, Y., Kim, J., Kim, J., Cho, H., Kang, J., Kang, P., & Kim, N. (2025). CheckEval: A Reliable LLM-as-a-Judge Framework for Evaluating Text Generation Using Checklists. Proceedings of EMNLP 2025, 15771-15798. arXiv:2403.18771. doi:10.48550/arXiv.2403.18771.

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024a). LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579. doi:10.48550/arXiv.2412.05579.

Li, M., Liu, Z., Deng, S., Joty, S., Chen, N. F., & Kan, M.-Y. (2024b). DnA-Eval: Enhancing Large Language Model Evaluation through Decomposition and Aggregation. arXiv:2405.15329. doi:10.48550/arXiv.2405.15329.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Le Bras, R., & Choi, Y. (2024). WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. arXiv:2406.04770. doi:10.48550/arXiv.2406.04770.

Liu, M., He, J., Zhang, B., Li, F., Ma, W., Lin, R., Zhong, S., & Fang, Q. (2026). Scoring Higher, Answering Worse: Mitigating Reward Hacking in Rubric-Based RL via Protocol-Level Rubrics. arXiv:2609.38847. doi:10.48550/arXiv.2609.38847.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023a). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634. doi:10.48550/arXiv.2303.16634.

Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., & Zhang, Q. (2023b). Calibrating LLM-Based Evaluator. LREC-COLING 2024, 2638-2656. arXiv:2309.13308. doi:10.48550/arXiv.2309.13308.

Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., & Zhang, Q. (2024). HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition. Proceedings of ACL 2024 (Volume 1: Long Papers), 7641-7660. arXiv:2402.15754. doi:10.48550/arXiv.2402.15754.

Mahmoud, A., Rezaei, M., Wang, Z., Gunjal, A., Liu, B., & He, Y. (2026). Reward Hacking in Rubric-Based Reinforcement Learning. arXiv:2605.12474. doi:10.48550/arXiv.2605.12474.

Mehri, S., & Eskenazi, M. (2020). USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. arXiv:2005.00456. doi:10.48550/arXiv.2005.00456.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. Proceedings of EMNLP 2023, 12076-12100. arXiv:2305.14251. doi:10.48550/arXiv.2305.14251.

Pan, T., Lin, X., Yang, W., He, Q., Chen, S., Qi, L., Xu, W., Feng, H., et al. (2026). RubricEval: A Rubric-Level Meta-Evaluation Benchmark for LLM Judges in Instruction Following. arXiv:2603.25133. doi:10.48550/arXiv.2603.25133.

Qin, Y., Song, K., Hu, Y., Yao, W., Cho, S., Wang, X., Wu, X., Liu, F., et al. (2024). InFoBench: Evaluating Instruction Following Ability in Large Language Models. arXiv:2401.03601. doi:10.48550/arXiv.2401.03601.

Roy, S., Pujari, R., Kumarage, T., Peris, C., Gupta, R., Rumshisky, A., Natarajan, P., & Saligrama, V. (2026). PReMISE: Policy Rubrics as Measurement Specifications for LLM Judges. arXiv:2605.30803. doi:10.48550/arXiv.2605.30803.

Saha, S., Levy, O., Celikyilmaz, A., Bansal, M., Weston, J., & Li, X. (2024). Branch-Solve-Merge Improves Large Language Model Evaluation and Generation. Proceedings of NAACL 2024 (Volume 1: Long Papers), 8352-8370. arXiv:2310.15123. doi:10.48550/arXiv.2310.15123.

Shen, C., Cheng, L., Nguyen, X.-P., You, Y., & Bing, L. (2023). Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. arXiv:2305.13091. doi:10.48550/arXiv.2305.13091.

Song, H., Su, H., Shalyminov, I., Cai, J., & Mansour, S. (2024a). FineSurE: Fine-grained Summarization Evaluation using LLMs. arXiv:2407.00908. doi:10.48550/arXiv.2407.00908.

Song, Y., Kim, Y., & Iyyer, M. (2024b). VeriScore: Evaluating the Factuality of Verifiable Claims in Long-Form Text Generation. arXiv:2406.19276. doi:10.48550/arXiv.2406.19276.

Steindl, S., Voskarides, N., Gasparin, A., & Marcheggiani, D. (2026). Does task decomposition improve automatic NLG evaluation? EMNLP 2026. arXiv:2609.01139. doi:10.48550/arXiv.2609.01139.

Viswanathan, V., Sun, Y., Ma, S., Kong, X., Cao, M., Neubig, G., & Wu, T. (2025). Checklists Are Better Than Reward Models For Aligning Language Models. arXiv:2507.18624. doi:10.48550/arXiv.2507.18624.

Wang, H., Xiong, W., Xie, T., Zhao, H., & Zhang, T. (2024). Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. arXiv:2406.12845. doi:10.48550/arXiv.2406.12845.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., et al. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926. doi:10.48550/arXiv.2305.17926.

Wanner, M., Ebner, S., Jiang, Z., Dredze, M., & Van Durme, B. (2024). A Closer Look at Claim Decomposition. arXiv:2403.11903. doi:10.48550/arXiv.2403.11903.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., et al. (2024a). Long-form factuality in large language models. arXiv:2403.18802. doi:10.48550/arXiv.2403.18802.

Wei, T., Wen, W., Qiao, R., Sun, X., & Ma, J. (2025). RocketEval: Efficient Automated LLM Evaluation via Grading Checklist. ICLR 2025. arXiv:2503.05142. doi:10.48550/arXiv.2503.05142.

Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., Thorne, J., Kim, J., & Seo, M. (2023). FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv:2307.10928. doi:10.48550/arXiv.2307.10928.

Yen, Y.-F. (2026). Decomposition-Induced Context-Memory Conflict: When Fact-Checking Pipelines Contradict Their Own Source Text. arXiv:2608.10627. doi:10.48550/arXiv.2608.10627.

Yu, D., Zhang, T., Mou, Y., Zhang, Y., Ye, W., & Zhang, S. (2026). Mitigating Rubric Interference in LLM Judges via On-Policy Self-Distillation. arXiv:2608.14684. doi:10.48550/arXiv.2608.14684.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., & Han, J. (2022). Towards a Unified Multi-Dimensional Evaluator for Text Generation. arXiv:2210.07197. doi:10.48550/arXiv.2210.07197.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23886823