# Survey on LLM-as-a-Judge Reliability: A Review of Recent Advances

**Parsa Gholami*** **Aria Alyasin*** **Mohsen Shirazi***
**Omid Ghahroodi** **Hossein Shahabadi** **Danial Parnian**
*Department of Computer Engineering*
*Sharif University of Technology*

## Abstract

Large language models (LLMs) are increasingly used to evaluate open-ended model outputs, construct leaderboards, and select data or responses during training and inference. This shift creates a second-order measurement problem: an evaluator that is scalable and articulate may still be invalid for the intended construct, inconsistent, poorly calibrated, contaminated, or vulnerable to manipulation. This survey synthesizes recent LLM-as-a-Judge research through four complementary categories: frameworks and scope; biases and robustness; scalability and adversarial security; and mitigation and training. The reviewed evidence spans pairwise preference, scalar scoring, system ranking, multilingual and multimodal assessment, dynamic agent trajectories, safety evaluation, human-annotation replication, and test-time scaling. Across these settings, no single agreement score establishes either validity or reliability. Judge quality depends on the task, evaluation format, model family, prompt and parser, data provenance, and relationship between the judge and evaluated model. Recent methods improve individual components through bias detection, distributional alignment, selective escalation, structured evidence, and reasoning-grounded preference training, yet they do not remove the need for human validation. We conclude with a reliability-oriented deployment framework that combines construct specification, task-specific meta-evaluation, perturbation and contamination audits, repeated trials, calibrated abstention, and downstream utility tests.

## 1 Introduction

The rapid advancement of Large Language Models (LLMs) has made open-ended generation a central capability in modern AI systems. Evaluating such outputs is difficult because quality often depends on correctness, relevance, reasoning, style, safety, and task-specific constraints that cannot be captured by lexical-overlap metrics such as BLEU or ROUGE. Human evaluation remains the most direct way to measure these properties, but it is expensive, slow, sometimes inconsistent, and difficult to reproduce at leaderboard scale.

The "LLM-as-a-Judge" paradigm attempts to close this gap by asking a language model to apply a natural-language rubric and return a score, preference, critique, or ranking. Early results from MT-Bench and Chatbot Arena showed that strong judges can approximate human preferences in open-ended dialogue (Zheng et al., 2023). Since then, judges have moved beyond retrospective benchmarking: they now help construct preference data, rerank candidates, guide search, and refine outputs at test time. Consequently, a wrong verdict can affect not only a reported metric but also which models and data are selected for future development.

The central question is therefore not whether LLM judges are useful, but under what evidence they can be trusted. A high average agreement rate may coexist with position bias (Wang et al., 2024), run-to-run inconsistency (Haldar & Hockenmaier, 2025), safety-irrelevant artifact sensitivity (Chen & Goldfarb-Tarrant, 2025), transferable adversarial attacks (Raina et al., 2024), or preference leakage from related synthetic-data

*Equal contribution.

generators (Li et al., 2026). Moreover, agreement on static pairwise benchmarks need not predict whether a judge can improve a generator during test-time scaling (Zhou et al., 2025) or evaluate the intermediate actions of an autonomous agent (Zhuge et al., 2024).

This survey makes three contributions. First, it organizes recent LLM-as-a-Judge research into a four-category taxonomy that connects evaluation settings, failure modes, and mitigation strategies, grounded in a measurement-theoretic distinction between validity and reliability. Second, it compares studies according to the evidence they provide—construct validity, human agreement, stability, calibration, attack resistance, system ranking, and downstream utility—rather than treating all reported judge scores as interchangeable. Third, it derives a practical reliability protocol and identifies open problems in task-specific validation, uncertainty, contamination, adversarial security, and human escalation.

## 2 Background & Conceptual Foundations

### 2.1 What Is an LLM Judge?

Open-ended outputs often cannot be evaluated by exact matching: correctness, completeness, style, and safety may all matter. LLM-as-a-Judge addresses this setting by asking a language model to apply a natural-language rubric to outputs from a *generator*. The evaluated outputs are the *candidates*, and the evaluator is the *judge*. Human annotations provide an important comparison standard, but not infallible ground truth for subjective criteria.

Formally, let $x$ denote the task input, $Y = \{y_1, \ldots, y_m\}$ the set of candidate outputs, $r$ the rubric, and $e$ optional evidence such as a reference answer, retrieved document, image, test result, or execution trace. A judge with parameters $\theta$ returns

$$z \sim J_\theta(\cdot \mid r, x, Y, e, c), \tag{1}$$

where $c$ includes the prompt, candidate order, decoding settings, and output schema. The judgment $z$ may be a label, score, preference, ranking, distribution, or critique. The distributional notation matters because an unchanged item can receive different judgments across samples or configurations (Haldar & Hockenmaier, 2025; Salinas et al., 2025).

General-purpose judges are prompted directly, whereas specialized judges may be trained on human labels, teacher judgments, preferences, or rationales. Generative judges can return explanations; scalar reward models usually return only a value, although reasoning-grounded methods such as Con-J blur this boundary (Ye et al., 2025b). Judges may also be *reference-based*, using an answer or external evidence, or *reference-free*, relying primarily on the rubric and internal knowledge. These choices determine both available evidence and likely failure modes.

### 2.2 Evaluation Formats and Levels of Analysis

In *pointwise* assessment, a judge assigns one candidate a label or score; in *pairwise* assessment, it selects between two candidates or declares a tie; and in *listwise* assessment, it orders several candidates. Pairwise evaluation avoids interpreting an absolute scale but introduces order and tie-policy effects. Direct scoring supports detailed rubrics but requires stable scale semantics. Aggregating local decisions into system rankings adds further assumptions: accurate individual comparisons need not recover the correct global ordering (Gera et al., 2025).

The evaluation object can also range from a completed response to reasoning steps, search states, multimodal evidence, or an agent trajectory. Agent evaluation may require tool calls, workspace state, and partial accomplishments that are absent from the final response (Chen et al., 2024a; Zhuge et al., 2024). Validity at one level should not be assumed at another; Appendix B maps the evaluation design space represented in the corpus.

### 2.3 The End-to-End Evaluation Pipeline

An LLM judge call is only one component of an evaluation system. For clarity, we separate the pipeline into five stages:

1. **Construct specification:** define the target property, population, context, and intended use.
2. **Evidence construction:** choose candidates, rubrics, references, and contextual evidence.
3. **Judgment elicitation:** set the judge, prompt, format, decoding, rationale, and confidence requirements.
4. **Post-processing and aggregation:** parse outputs, handle ties or failures, and combine judgments into scores or rankings.
5. **Decision use:** report results or use them for selection, training, inference, revision, or escalation.

Reliability therefore belongs to the complete procedure, not only the backbone. Parsing and aggregation can alter results, and static human agreement may not predict performance when a judge guides search or refinement (Zhou et al., 2025).

### 2.4 Validity and Reliability as Distinct Measurement Properties

Chehbouni et al. (2025) provide the conceptual anchor for separating validity from reliability. A broad concept such as "helpfulness" must be defined for a context and operationalized through a rubric; each step can introduce error. A rubric that equates helpfulness with detail, for example, may reproducibly reward unnecessarily long answers.

*Validity* asks whether the judgment represents the intended construct for a specified task, population, and use. Evidence may concern rubric coverage, agreement with credible measures, insensitivity to irrelevant features, prediction of relevant outcomes, and the consequences of acting on the score. Human agreement supports convergence but cannot establish these other properties alone.

These forms of evidence answer different questions. Content evidence asks whether the rubric covers the construct; convergent evidence compares the judge with experts or crowds; discriminant evidence tests whether irrelevant cues affect the score; and predictive or consequential evidence evaluates the decisions produced by using it. A judge may satisfy one form while failing another, so validity is an argument assembled for a particular use rather than a single coefficient.

*Reliability* asks whether the measurement is stable across repeated trials, irrelevant order or prompt changes, and judges applying the same rubric. It is necessary but insufficient for validity: an evaluator may consistently prefer polished but incorrect responses, while high average accuracy can hide item-level instability. Table 1 summarizes the distinction and related deployment properties.

Claims that judges proxy humans, possess the needed expertise, scale, and reduce cost are empirical hypotheses rather than consequences of model size (Chehbouni et al., 2025). Here "reliability" is a broad operational umbrella, while *construct validity* and *test–retest reliability* retain their stricter measurement meanings.

### 2.5 Uncertainty, Disagreement, and Calibration

A verdict and its confidence are different outputs. Confidence may come from token probabilities, repeated samples, judge panels, or a calibrated model. *Calibration* relates that confidence to observed correctness or human agreement; raw probabilities need not be calibrated and may deteriorate under task or distribution shift (Xie et al., 2025; Jung et al., 2024).

Human disagreement may instead reflect subjectivity, an underspecified rubric, or genuine ambiguity. Distributional judges preserve this information rather than collapsing it to a majority label (Chen et al., 2025). Human disagreement, missing-evidence uncertainty, and judge instability remain distinct and may justify different actions: tie, abstention, another judge, or human escalation.

| Property | Central Question | Illustrative Failure | Typical Evidence |
|---|---|---|---|
| Validity | Does the score represent the intended construct and support its intended use? | A safety judge rewards apologetic wording rather than safer content. | Rubric analysis, expert comparison, controlled contrasts, downstream tests. |
| Reliability | Would the procedure return the same judgment when relevant evidence is unchanged? | Identical inputs receive different scores across runs or candidate orders. | Repeated trials, order swaps, prompt variants, inter-judge agreement. |
| Calibration | Does stated confidence correspond to empirical correctness or agreement? | Verdicts reported with 90% confidence agree with validated labels far less often. | Reliability diagrams, calibration error, selective-risk analysis. |
| Robustness | Does the judgment withstand irrelevant variation, distribution shift, and manipulation? | A stylistic artifact or short adversarial suffix changes the preferred answer. | Perturbation tests, out-of-distribution evaluation, adaptive attacks. |
| Utility | Does using the judge improve the downstream decision? | High benchmark agreement fails to improve reranking, search, or refinement. | Ranking fidelity, task gain, cost–coverage analysis. |

Table 1: Conceptual properties of an LLM-based evaluation system. No single metric establishes all five properties.

### 2.6 Bias, Robustness, and Sources of Failure

A *bias* is systematic dependence on a feature that should not determine the judgment; unreliability may instead be non-directional variation across calls. A biased evaluator can therefore be consistent. *Robustness* concerns stability under quality-preserving changes, from diagnostic order swaps to deliberate manipulation. Biases & Robustness covers controlled, non-adversarial failures, while Scalability & Adversarial Security covers strategic attacks and benchmark gaming.

Failures arise at the presentation layer (position, length, style, or authority cues), model layer (stochasticity, knowledge limits, family preferences, or superficial reasoning), implementation layer (prompts, schemas, and parsers), and data-lineage layer (related generators and judges) (Wang et al., 2024; Chen & Goldfarb-Tarrant, 2025; Haldar & Hockenmaier, 2025; Wang et al., 2025; Li et al., 2026). Deployment adds domain, language, modality, and agent-environment shift, while universal adversarial phrases explicitly target learned evaluator preferences (Raina et al., 2024).

The remainder of the survey asks whether a judge measures the intended construct, is stable, expresses meaningful uncertainty, withstands natural and adversarial variation, and improves the downstream decision. Section 3 organizes the evidence, and Appendix C presents a reliability measurement suite.

## 3 Review Methodology and Taxonomy

### 3.1 Scope and Selection Criteria

This survey examines LLM-based evaluators that apply natural-language criteria to generated outputs, model behavior, or agent trajectories. The scope includes prompted general-purpose judges, specialized fine-tuned evaluators, multimodal and multilingual judges, judge panels, and systems that use judgments for ranking, data construction, preference alignment, or test-time decision making. The evidence base emphasizes work made publicly available from 2023 through 2026, a period spanning the emergence of widely adopted dialogue benchmarks and subsequent research on evaluator bias, uncertainty, security, agentic assessment, and judge training. Adjacent research on human annotation, scalar reward modeling, and conventional automatic metrics is considered only when it directly informs the design or validation of LLM judges.

The literature was assembled through purposive, criterion-driven selection rather than an exhaustive systematic search. A study was included when it satisfied three conditions: (1) LLM-based judgment was a primary object of analysis or a central methodological component; (2) the work contributed evidence concerning validity, reliability, calibration, robustness, scalability, or downstream utility; and (3) the paper provided sufficient methodological, empirical, theoretical, or conceptual detail to support comparative analysis. Selection additionally sought coverage across evaluation formats, application domains, failure modes, and mitigation strategies so that the taxonomy would represent distinct research questions rather than variations of a single benchmark setting.

Studies were excluded when they used an LLM judge only as an unexamined reporting instrument, addressed generative-model quality without analyzing the evaluator, or duplicated an earlier version without a distinct contribution relevant to the survey. Both peer-reviewed publications and recent preprints were retained because methodological developments in this area often precede archival publication; publication status was recorded during study extraction rather than treated as a proxy for evidential strength. Conceptual and theoretical work was included alongside empirical studies when it clarified the assumptions, limits, or measurement properties of automated evaluation.

For each included work, the analysis records the evaluation object, judgment format, judge type, reference or evidence conditions, validation target, principal metrics, documented failure modes, mitigation strategy, and stated limitations. Because tasks, annotation protocols, judge models, and outcome measures differ substantially across studies, the survey adopts a structured qualitative synthesis rather than pooling effect sizes that are not directly comparable. Claims are interpreted within the setting in which they were established, and cross-study conclusions are reserved for patterns supported across multiple methodologies or domains.

The synthesis is organized by three research questions: **RQ1:** In which tasks and formats do LLM judges align with human or objective evaluation? **RQ2:** Which sources of bias, uncertainty, contamination, and attack undermine this alignment? **RQ3:** Which training, prompting, calibration, and human-in-the-loop strategies improve reliability, and what limitations remain?

### 3.2 Reliability Taxonomy

Rather than listing papers chronologically, we organize them by their primary reliability question. **Frameworks & Scope** defines evaluation settings across languages, modalities, rankings, test-time scaling, and agents. **Biases & Robustness** studies cognitive, response-level, multimodal, stochastic, and lineage-related failures under controlled variation. **Scalability & Adversarial Security** covers deployment cost, anchor selection, theoretical limits, selective escalation, benchmark gaming, and attacks. **Mitigation & Training** develops structured evidence, calibration, disagreement-aware alignment, and reasoning-grounded objectives.

Figure 1 summarizes the classification without treating the categories as isolated paper lists. Several works span multiple axes; for example, JETTS is a framework but also tests downstream utility, while distributional alignment is simultaneously an uncertainty model and a training method. Each empirical or methodological paper is placed according to its central contribution, and cross-category implications are discussed later. Chehbouni et al. (2025) is intentionally retained as a cross-cutting Background anchor rather than forced into one category.

The taxonomy answers the research questions at different levels: Frameworks & Scope defines where validity must be established; Biases & Robustness diagnoses controlled failures; Scalability & Adversarial Security tests operational and hostile conditions; and Mitigation & Training improves the judgment process.

### 3.3 Evolutionary Trajectory of Reliability Research

The taxonomy describes the literature's current structure, while Figure 2 shows its overlapping conceptual progression. Research moved from asking whether strong LLMs approximate human preference toward identifying when agreement fails, where it transfers, and what controls are required when judgments influence model development.

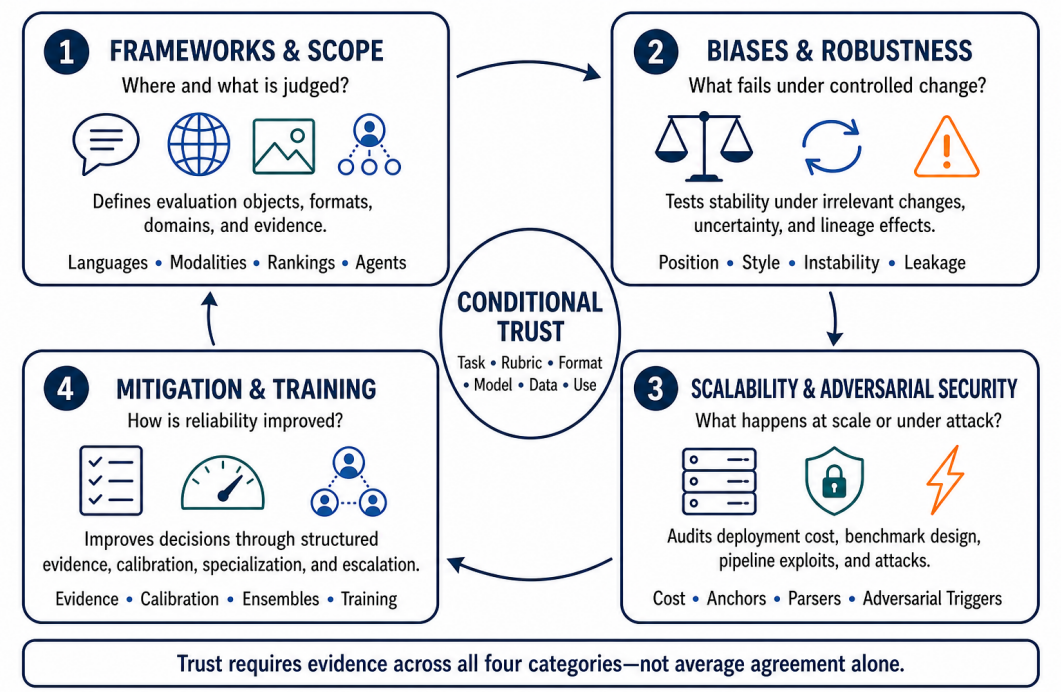


Figure 1: Conceptual taxonomy of LLM-as-a-Judge reliability. The four connected branches cover evaluation scope, controlled failures, scalable and secure deployment, and reliability improvement. The central conditions emphasize that trust must be validated for a particular configuration and use.

Three trends emerge: evaluation has expanded from response preferences to rankings and judge-guided inference; analysis has widened from one call to prompts, parsers, sampling, lineage, and escalation; and mitigation has shifted toward layered combinations of training, structured evidence, uncertainty, adversarial testing, and human review. The field now asks not merely whether LLMs can judge, but when their judgments support a particular decision.

## 4 Review of Key Papers

This section examines the literature according to the taxonomy above. Appendix A summarizes the composition of the reviewed corpus, while the discussion here remains focused on thematic synthesis.

### 4.1 Frameworks and Scope

Early frameworks established useful but limited forms of validity. MT-Bench and Chatbot Arena showed that strong LLMs can approximate human preferences in multi-turn dialogue and blind pairwise comparison (Zheng et al., 2023). JudgeBench then exposed weaker performance when correctness depends on subtle reasoning rather than style (Tan et al., 2025), while JuStRank showed that instance-level accuracy need not yield a faithful system ranking (Gera et al., 2025). Together, these studies separate response-level agreement from reasoning validity and ranking fidelity.

Later work broadened both the evaluation object and deployment setting. SOTOPIA uses goal-driven interaction to evaluate social intelligence (Zhou et al., 2024); Agent-as-a-Judge inspects tool use, environment state, and hierarchical requirements across complete trajectories (Zhuge et al., 2024); and MLLM-as-a-Judge extends evaluation to vision–language outputs, where scoring and batch ranking are less reliable than pairwise comparison (Chen et al., 2024a). M-Prometheus similarly shows that English-centered judge behavior does

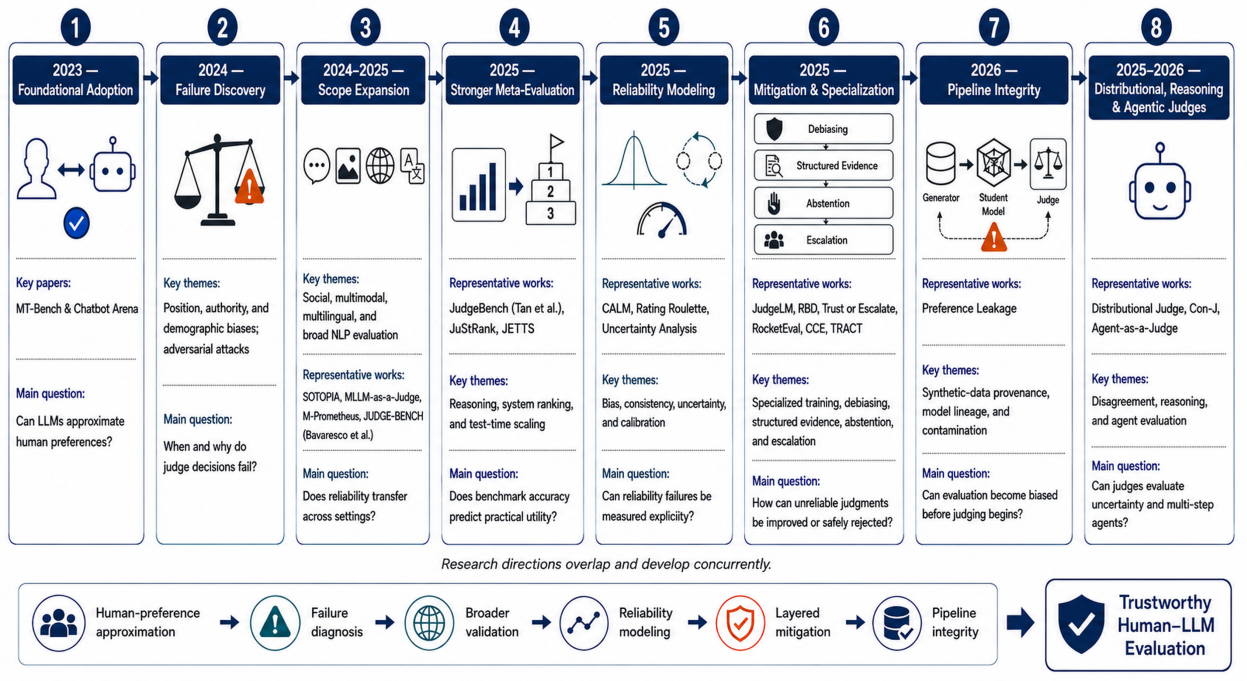


Figure 2: Evolutionary trajectory of LLM-as-a-Judge reliability research. The field progressed from approximating human preferences to diagnosing failures, validating judges across broader settings, modeling reliability explicitly, developing layered mitigation, and auditing pipeline-level contamination. The stages indicate dominant research emphases rather than a strictly chronological sequence.

not transfer automatically across languages and that multilingual training design matters (Pombal et al., 2025). These extensions make task, language, modality, and accessible evidence part of the validity claim.

The comparison among these settings is important. Dialogue evaluation can often rely on a visible response, whereas social and agentic evaluation depends on interaction history, hidden state, and partial progress. Multimodal and multilingual judging changes the evidence channel itself. Agreement obtained in English pairwise dialogue therefore supplies little direct evidence for trajectory scoring, absolute multimodal grading, or cross-lingual feedback. Each extension requires a new human or objective baseline aligned with what the judge can actually observe.

The empirical contrasts are substantial. On DevAI, an evaluator with environment access and hierarchical requirement checks reached 90% agreement with the consensus human baseline, compared with 70% for a conventional response judge (Zhuge et al., 2024). Multimodal judging performs comparatively well in pairwise evaluation but degrades under scoring and batch ranking (Chen et al., 2024a). M-Prometheus further finds that synthetic multilingual feedback can outperform straightforward translation of English judge data (Pombal et al., 2025). The relevant improvement comes from matching evidence and training to the evaluation setting, not simply replacing the backbone.

Broad meta-evaluation reinforces this conditional view. JUDGE-BENCH finds substantial variation across tasks, data sources, scales, and annotator expertise, with ordinary chain-of-thought prompting offering no general remedy (Bavaresco et al., 2025). JETTS asks a more consequential question: whether judges improve reranking, search, or critique-based refinement. Performance varies by use, and strong static preference accuracy does not ensure useful search guidance or refinement (Zhou et al., 2025). Judge quality must therefore be validated against both the target domain and the decision the score will support.

The breadth of these tests also matters. JUDGE-BENCH spans 20 NLP datasets and more than 70,000 individually annotated examples, while JETTS compares ten judges across reranking, step-level search, and refinement in math, code, and instruction following. Their shared result is not that judges always fail, but

that aggregate performance conceals where they succeed. Coverage across tasks must therefore be paired with analysis of downstream consequences.

### 4.2 Biases and Robustness

Two basic failures appear even before semantic bias is considered. Reversing candidate order can change pairwise verdicts, motivating evidence-first and balanced-position calibration (Wang et al., 2024). Repeating an identical call can also produce different scores, so a single deterministic-looking output does not establish test–retest reliability (Haldar & Hockenmaier, 2025). Order swaps and repeated trials diagnose different dependencies and should both be reported.

Controlled perturbation studies show that clean-set agreement can conceal reliance on irrelevant cues. Reference-free tests identify misinformation, gender, authority, and beauty biases, including susceptibility to fabricated references (Chen et al., 2024b); CALM expands this analysis to 12 cognitive and presentation biases (Ye et al., 2025a). In safety evaluation, apology, authority, halo, verbosity, and position cues can substantially change verdicts, while larger judges and model juries do not consistently eliminate the problem (Chen & Goldfarb-Tarrant, 2025). MM-JudgeBias extends the same logic across queries, images, and responses, revealing modality neglect and asymmetric evidence integration in multimodal judges (Lee et al., 2026). Robustness must therefore be tested with construct-preserving, domain- and modality-specific contrasts.

The observed effects are large enough to change practical conclusions. Apologetic phrasing alone shifts some comparative safety preferences by as much as 98%, yet clean human agreement does not predict this artifact resistance (Chen & Goldfarb-Tarrant, 2025). MM-JudgeBias evaluates more than 1,800 samples derived from 29 benchmarks and finds systematic problems across 26 multimodal judges (Lee et al., 2026). Such results justify reporting perturbation-specific robustness alongside ordinary accuracy rather than treating it as an optional diagnostic.

Greater reasoning capability is not a general solution. Large Reasoning Models are more robust on some fact-oriented tasks but remain susceptible to established biases and to *superficial reflection bias*, in which deliberative wording influences the decision without new evidence (Wang et al., 2025). External detection offers a complementary intervention: the Reasoning-based Bias Detector identifies several bias types before guiding revision and improves accuracy and consistency without modifying the underlying judge (Yang et al., 2025). Both results require caution because reasoning can expose artifacts as well as correct them.

These studies also distinguish validity failures from instability. Position, authority, and modality neglect can systematically direct the judge toward the wrong evidence even when its decisions are reproducible. Rating Roulette instead measures variation without requiring a directional bias. Safety artifacts demonstrate why both tests are needed: a model can agree with humans on the clean distribution yet change its verdict under irrelevant wording. Reporting only average agreement collapses these diagnostically different failures.

Other failures arise from uncertainty and pipeline dependence. Token-probability confidence varies with model family, scale, prompting, training, and distribution shift; ConfiLM shows that using this signal during training can improve out-of-distribution evaluation (Xie et al., 2025). Preference leakage operates earlier: judges favor students trained on synthetic data from related generators, even when model identities are hidden, and the effect increases with lineage proximity and synthetic-data share (Li et al., 2026). Response-level bias, stochastic instability, uncertainty, and data lineage are therefore separate audit targets rather than one reliability score.

The leakage finding changes what anonymization can guarantee. Hiding a model name may reduce explicit self-preference, but it cannot remove stylistic or distributional signatures already learned by a student from a related generator. Evaluation should therefore disclose generator–student–judge relationships and compare unrelated judge families, particularly when synthetic data constitute a large share of training.

Uncertainty is useful only when connected to an action and validated against errors. Repeated disagreement may flag volatility, whereas calibrated confidence can support abstention; neither reveals preference leakage, which requires provenance controls and cross-family comparisons. A robust pipeline must therefore combine

perturbation tests, repeated trials, calibration analysis, and lineage audits rather than substituting one diagnostic for the others.

### 4.3 Scalability and Adversarial Security

Scalability depends on the whole configuration, not model size alone. JudgeLM shows that smaller open judges can be trained to approximate teacher judgments while addressing known biases (Zhu et al., 2025). Multi-fidelity tuning further demonstrates that model, prompt, output format, temperature, and parser should be optimized jointly against accuracy and cost (Salinas et al., 2025). Anchor-based evaluation adds another design variable: extreme comparison anchors provide little information about closely matched systems, and standard sample sizes may be underpowered (Don-Yehiya et al., 2026). Efficient evaluation therefore requires configuration and benchmark-design validation, not merely a cheaper judge.

The studies quantify different sources of efficiency. JudgeLM reports more than 90% agreement with teacher judgments, whereas multi-fidelity search reduces the expense of exploring judge configurations relative to exhaustive evaluation. The anchor study evaluates 22 alternatives on Arena-Hard-v2.0 and finds that anchor choice can affect human-rank correlation as much as judge-model choice. These gains remain conditional on the teacher, search space, opponent pool, and target ranking.

Scale also increases the incentive and opportunity for manipulation. Adversarial null models can exploit templates and parsers without attempting the underlying task (Zheng et al., 2025), while short universal phrases optimized against a surrogate can transfer to unseen judges and inflate scores, particularly in absolute assessment (Raina et al., 2024). These attacks target different layers of the pipeline and motivate robust parsing, hidden or varied templates, input controls, and format-specific adversarial tests.

The distinction between the attacks is operationally useful. Parser exploitation can succeed even when the judge understands the task, whereas transferable phrases exploit the judge's learned preferences. Defending only the prompt or only the model therefore leaves another attack surface exposed. Cost optimization should include the expense of adversarial testing and monitoring, not just the price of clean evaluation calls.

Human labels nevertheless impose a theoretical and operational floor. If a judge is not more accurate than the evaluated model, optimal debiasing cannot reduce ground-truth labeling requirements by more than half (Dorner et al., 2025). Selective evaluation responds by calibrating when to decide, abstain, or escalate, and cascades can reserve expensive judges or humans for uncertain cases (Jung et al., 2024). The deployment question is thus not which judge is universally best, but which cases it may decide at an acceptable risk and cost.

This pairing clarifies the role of humans. The statistical limit rules out unlimited savings from a weak proxy, while selective guarantees use human labels to define an acceptable region of operation. Cost comparisons should therefore include calibration data, abstention coverage, audit frequency, and the consequences of judge errors, rather than only per-call price.

### 4.4 Mitigation and Training

One family of methods structures or calibrates evidence without retraining the judge backbone. RocketEval decomposes each item into a checklist, grades the facets separately, and reweights them against supervision (Wei et al., 2025). Crowd Comparative Evaluation supplies selected comparisons with additional responses to expose strengths and weaknesses that an isolated A/B judgment may miss (Zhang et al., 2025). SAJA instead maps multi-dimensional outputs from one structured-rubric call to a human-aligned score using a lightweight calibration head and a small labeled set, avoiding logit access and iterative prompt search (Kola et al., 2026). These methods trade additional structure or local supervision for better task alignment.

Their mechanisms should not be conflated. RocketEval changes the unit of judgment, Crowd Comparative Evaluation changes the contextual evidence, and SAJA changes the mapping from structured features to the final score. SAJA improves MT-Bench pairwise F1 from 78% to 86% and reports up to 4.8× cost savings over per-question approaches, illustrating how local calibration can improve both alignment and deployment efficiency without assuming universal transfer.

Training-based approaches modify what the judge learns to predict. TRACT combines reasoning supervision with a regression-aware loss so that numeric errors reflect their magnitude and reduces mismatch between training and inference rationales (Chiang et al., 2025). Distributional LLM-as-a-Judge learns the empirical distribution of human ratings rather than a collapsed label, retaining disagreement while improving robustness (Chen et al., 2025). Con-J uses contrastive judgments and preference optimization to align both verdicts and their rationales, outperforming a scalar reward model trained on the same preferences and reducing some dataset artifacts (Ye et al., 2025b).

The target representation determines what information survives training. TRACT preserves ordinal distance among scores, distributional alignment preserves disagreement among annotators, and Con-J preserves an explicit rationale tied to a preference decision. These improvements are complementary, but none ensures that a rationale is causally faithful or that an empirical annotation distribution remains calibrated after domain shift.

The two families impose different costs and assumptions. Evidence structuring and calibration can wrap closed models and are easier to update, but require extra context, calls, or local labels. Specialized training provides greater control over scores, distributions, and rationales, but can inherit teacher bias and may need retraining after task shift. Selection should therefore depend on the deployment constraint rather than on a universal mitigation ranking.

These interventions address different links between evidence, uncertainty, reasoning, and decision. Decomposition can improve coverage, calibration can adapt a general judge locally, distributional targets preserve disagreement, and regression- or preference-aware objectives better match output semantics. None establishes faithful reasoning or calibrated transfer by itself; intermediate evidence and uncertainty still require independent validation.

## 5 Comparison and Discussion

### 5.1 What the Evidence Establishes

Across the reviewed literature, four conclusions are well supported. First, validity and reliability are distinct and both are conditional. Stable scores can consistently operationalize the wrong construct, while high human correlation provides only one form of validity evidence (Chehbouni et al., 2025). Strong agreement in open-ended dialogue (Zheng et al., 2023) coexists with failures on subtle reasoning (Tan et al., 2025), variation across 20 NLP tasks (Bavaresco et al., 2025), and weak utility in some test-time-scaling settings (Zhou et al., 2025). Claims about a "reliable judge" must therefore name the construct, task, population, format, and decision being supported.

Second, operational reliability is multi-dimensional. Average accuracy does not reveal position sensitivity (Wang et al., 2024), repeated-trial instability (Haldar & Hockenmaier, 2025), safety-artifact dependence (Chen & Goldfarb-Tarrant, 2025), multimodal compositional bias (Lee et al., 2026), cognitive-style bias in reasoning models (Wang et al., 2025), confidence under distribution shift (Xie et al., 2025), or adversarial vulnerability (Raina et al., 2024; Zheng et al., 2025). Likewise, instance-level agreement does not necessarily imply correct system ranking (Gera et al., 2025), and preference-benchmark accuracy does not ensure effective reranking or search (Zhou et al., 2025). Consequential applications therefore require multiple complementary metrics.

Third, the evaluation format changes both the construct and the attack surface. Pairwise comparison avoids some scalar-calibration difficulties and is more robust than absolute scoring to universal suffix attacks (Raina et al., 2024), but it remains vulnerable to response order, tie policy, aggregation, and the choice of comparison anchor (Don-Yehiya et al., 2026). Direct assessment supports fine-grained rubrics and critiques but introduces ordinal or regression errors, motivating checklist decomposition (Wei et al., 2025), regression-aware learning (Chiang et al., 2025), and task-specific calibration (Kola et al., 2026). Distributional judging preserves annotator disagreement that a point label discards (Chen et al., 2025), while agent evaluation requires trajectory- and environment-level evidence that a response-only judge cannot observe (Zhuge et al., 2024).

Fourth, judge independence matters. Self-preference has long been treated as a response-level bias, but preference leakage shows that dependence can enter earlier through synthetic-data generation (Li et al., 2026). A judge may favor a related student even when evaluator prompts are neutral. Reliable experiments should therefore document the model lineage and provenance of training, reference, and evaluation data.

### 5.2 Mitigation Trade-offs

The mitigation literature divides into four layers. Prompt- and evidence-based methods, such as MEC/BPC, RocketEval, crowd comparative reasoning, and LRM self-reflection, can be applied without retraining the judge but increase inference cost and may introduce additional anchoring effects. External detectors such as RBD remain compatible with closed models, although the detector's reasoning requires independent validation. Lightweight calibration methods such as SAJA adapt structured judge outputs with a small human-labeled set while retaining a single judge call (Kola et al., 2026). Fine-tuned judges such as JudgeLM, M-Prometheus, ConfiLM, TRACT, Distributional LLM-as-a-Judge, and Con-J provide greater control, but inherit their annotations, empirical preference distributions, and base-model priors. Selective and cascaded evaluation operationalizes uncertainty through abstention or escalation (Jung et al., 2024), but exchanges coverage and cost for stronger guarantees.

These approaches are complementary. A practical system can use a task-tuned judge, structured evidence, repeated order-balanced trials, calibrated confidence, and a human escalation rule. A central design implication is that reliability should be treated as an end-to-end property of data, prompts, models, parsing, aggregation, and decisions rather than as a property of the backbone alone.

### 5.3 A Reliability-Oriented Deployment Protocol

The synthesis suggests a five-stage protocol. **(1) Define the construct:** specify the systematized concept, relevant quality dimensions, rubric, output format, and downstream decision. **(2) Validate locally:** collect task-specific content and convergent evidence, compare against expert or crowd judgments, report human–human agreement, and verify that the benchmark has enough power to separate the systems of interest. **(3) Stress-test:** repeat trials, swap candidate order, control length and style, vary pairwise anchors, test domain- and modality-relevant artifacts and bias perturbations, and run adversarial suffix and parser checks. Agentic settings should additionally vary trajectories, tool states, and partial completion. **(4) Audit dependence:** record generator, student, reference, and judge lineages and test cross-family judges for preference leakage. **(5) Calibrate action:** use task-specific human labels to calibrate judge outputs; report uncertainty or disagreement distributions, coverage, cost, and downstream utility; and abstain or escalate when evidence is insufficient. This protocol is stricter than selecting the model with the highest benchmark accuracy, but it matches the range of failures observed in the corpus. Figure 3 summarizes the resulting workflow, including its uncertainty-triggered human-audit and protocol-revision loop.

## 6 Open Problems and Future Directions

**Task-conditional meta-evaluation.** The breadth of JUDGE-BENCH and the operational design of JETTS should be combined: future benchmarks should measure both agreement on diverse tasks and the consequences of using a judge for selection, search, refinement, or training. Multilingual and multimodal work should additionally report results by language, culture, modality, and task rather than only macro averages (Pombal et al., 2025; Chen et al., 2024a). For agents, evaluation should cover intermediate actions, environment state, tool failures, and long-horizon requirement satisfaction beyond code-generation tasks (Zhuge et al., 2024).

**Uncertainty with guarantees under shift.** Token probabilities, verbalized confidence, self-consistency, and learned judgment distributions capture different phenomena. Future work should compare them under temporal and domain shift, distinguish annotator disagreement from model ignorance, and integrate distributional alignment with selective-risk guarantees (Xie et al., 2025; Chen et al., 2025; Jung et al., 2024). Deployable judges should support reliable detection of insufficient rubrics or knowledge rather than treating diffuse next-token probabilities as the sole indicator of uncertainty.

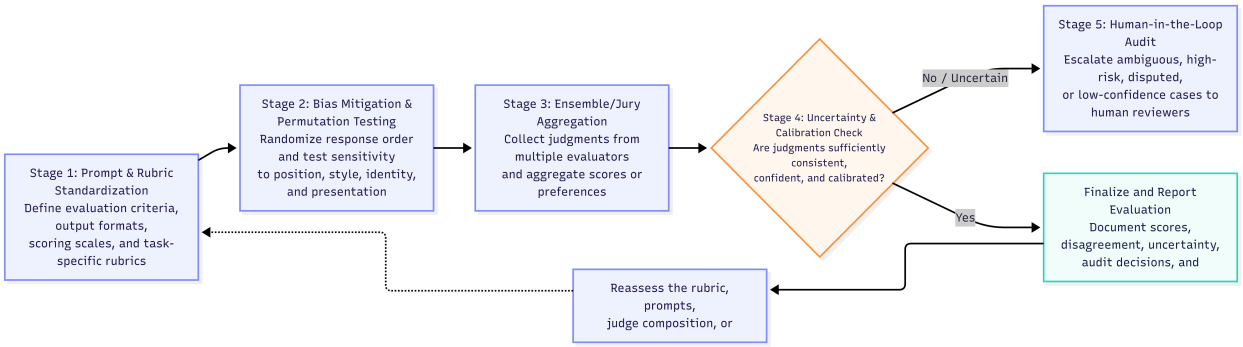


Figure 3: Recommended evaluation protocol for LLM-as-a-Judge systems. Standardized prompts and rubrics are followed by permutation-based bias testing, multi-judge aggregation, and uncertainty-aware calibration. Ambiguous, inconsistent, or high-risk judgments are escalated to human reviewers and may trigger revision of the evaluation design.

**Contamination-aware evaluation.** Preference leakage creates a need for model-lineage disclosure and cross-family evaluation. Open problems include black-box detection when training data are unavailable, separating legitimate stylistic similarity from leaked preference, and designing judge panels whose errors are genuinely diverse rather than inherited from a shared teacher (Li et al., 2026).

**Adaptive adversarial security.** Static perturbation sets may quickly become obsolete. Future benchmarks should include adaptive attackers, semantic-preservation checks, hidden prompt variants, parser fuzzing, and attacks against multi-judge or checklist systems (Raina et al., 2024; Zheng et al., 2025). Controlled artifacts should also reflect domain-specific spurious cues, as safety evaluators can fail on apology, authority, and halo language even without a strategic attacker (Chen & Goldfarb-Tarrant, 2025). Defenses must be evaluated against both natural artifact shift and attack transfer.

**Faithful and metric-aware reasoning.** Checklists, comparative context, and CoT-aware objectives improve outcomes, but generated rationales may still be post-hoc or vulnerable to superficial-reflection cues. Future work should test whether intermediate evidence causally supports the verdict, whether it is factually correct, and whether optimization for correlation or contrastive preference preserves calibration and fairness (Wei et al., 2025; Zhang et al., 2025; Chiang et al., 2025; Ye et al., 2025b; Wang et al., 2025).

**Human–LLM evaluation design.** Human evaluation should be allocated where it provides the most information: ambiguous items, distribution shifts, high disagreement, and cases near consequential thresholds. The field still needs principled methods for choosing the calibration set, estimating the required number of labels, and updating escalation policies as models and task distributions change (Dorner et al., 2025; Jung et al., 2024).

### 6.1 Limitations

This review uses purposive rather than exhaustive literature selection and may therefore omit relevant studies, particularly concurrent work in this rapidly evolving area. Substantial heterogeneity in tasks, judge models, annotation protocols, and metrics precludes meta-analytic aggregation and restricts cross-study comparisons to qualitative patterns. Several findings depend on proprietary or frequently updated models and on preprints whose methods or conclusions may change, limiting reproducibility and temporal stability. Evidence also remains uneven across languages, modalities, agentic settings, and high-stakes deployment contexts, so the resulting taxonomy should not be interpreted as uniformly validated across all applications.

## 7 Conclusion

This survey examined LLM-as-a-Judge evaluation through four categories: frameworks and scope; biases and robustness; scalability and adversarial security; and mitigation and training. The literature supports a

balanced conclusion. LLM judges are flexible and often economical evaluators, but agreement on a convenient benchmark establishes neither general construct validity nor test–retest reliability. Their decisions depend on format, task, stochasticity, disagreement and confidence calibration, model and data lineage, prompt and parser design, and adversarial conditions. Trustworthy use therefore requires a validated measurement pipeline: define the construct, compare against task-specific humans, audit bias and contamination, stress-test artifacts, attacks, and repeated trials, preserve uncertainty when disagreement is substantive, measure downstream utility, and escalate uncertain cases. Under these controls, LLM judges can complement human evaluation; without them, scale risks amplifying measurement error.

## References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2025.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. Neither valid nor reliable? investigating the use of LLMs as judges. *arXiv preprint arXiv:2508.18076*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yaochen Wang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024a.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? A study on judgement biases. *arXiv preprint arXiv:2402.10669*, 2024b.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. *arXiv preprint arXiv:2503.09347*, 2025.

Luyu Chen, Zeyu Zhang, Haoran Tan, Quanyu Dai, Hao Yang, Zhenhua Dong, and Xu Chen. Distributional LLM-as-a-judge. In *Advances in Neural Information Processing Systems*, 2025.

Cheng-Han Chiang, Hung-yi Lee, and Michal Lukasik. TRACT: Regression-aware fine-tuning meets chain-of-thought reasoning for LLM-as-a-judge. *arXiv preprint arXiv:2503.04381*, 2025.

Shachar Don-Yehiya, Asaf Yehudai, Leshem Choshen, and Omri Abend. Mediocrity is the key for LLM as a judge anchor selection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15491–15513. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.706. URL https://aclanthology.org/2026.acl-long.706/.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. *arXiv preprint arXiv:2410.13341*, 2025.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. JuStRank: Benchmarking LLM judges for system ranking. In *Proceedings of the Association for Computational Linguistics*, 2025.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. *arXiv preprint arXiv:2510.27106*, 2025.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*, 2024.

Sneha Kola, Pankaj Kumar Sharma, Soumyadeep Dey, Bamdev Mishra, and Mayur Datar. SAJA: A simple approach to judge alignment for LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 646–664. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.45. URL https://aclanthology.org/2026.acl-industry.45/.

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A benchmark for evaluating compositional biases in MLLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25336–25373. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1162. URL https://aclanthology.org/2026.acl-long.1162/.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. In *International Conference on Learning Representations*, 2026.

Jose Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre F. T. Martins. M-Prometheus: A suite of open multilingual LLM judges. In *Conference on Language Modeling*, 2025.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. *arXiv preprint arXiv:2402.14016*, 2024.

David Salinas, Omar Swelam, and Frank Hutter. Tuning LLM judge design decisions for 1/1000 of the cost. In *International Conference on Machine Learning*, 2025.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the Association for Computational Linguistics*, 2024.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. RocketEval: Efficient automated LLM evaluation via grading checklist. In *International Conference on Learning Representations*, 2025.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. In *International Conference on Learning Representations*, 2025.

Haoyan Yang, Runxue Bao, Cao Xiao, Jun Ma, Parminder Bhatia, Shangqian Gao, and Taha Kass-Hout. Any large language model can be a reliable judge: Debiasing with a reasoning-based bias detector. In *Advances in Neural Information Processing Systems*, 2025.

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

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. Learning LLM-as-a-judge for preference alignment. In *International Conference on Learning Representations*, 2025b.

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. Crowd comparative reasoning: Unlocking comprehensive evaluations for LLM-as-a-judge. In *Proceedings of the Association for Computational Linguistics*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Zi Lin, Zhuohan Li, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *International Conference on Learning Representations*, 2025.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Yonatan Bisk, Daniel Fried, Graham Neubig, Louis-Philippe Morency, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The JETTS benchmark of LLM-as-judges as test-time scaling evaluators. In *Proceedings of the International Conference on Machine Learning*, 2025.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, 2025.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

## A Corpus Coverage

The final corpus contains 33 works published or released between 2023 and 2026: one from 2023, seven from 2024, 22 from 2025, and three from 2026. As recorded at the time of review, 22 appeared in conference proceedings and 11 were retained as recent preprints. This concentration reflects the rapid growth of reliability-focused LLM-as-a-Judge research rather than an attempt to treat publication status as a measure of evidential strength.

For a mutually exclusive overview, each work is assigned by its primary contribution: nine examine *Frameworks and Scope*, ten address *Biases and Robustness*, seven study *Scalability and Adversarial Security*, and six develop *Mitigation and Training*. One measurement-theoretic paper is retained as a cross-cutting conceptual foundation. These assignments prevent double counting in the corpus summary; the main synthesis still discusses work across categories when its evidence bears on multiple reliability questions.

## B Evaluation Design Space

Figure 4 separates three choices that are often conflated: the object being judged, the form of the verdict, and the protocol used to establish reliability. These dimensions interact, but none determines the others; the same response can be scored, compared, ranked, or critiqued, and each choice requires different reliability evidence.

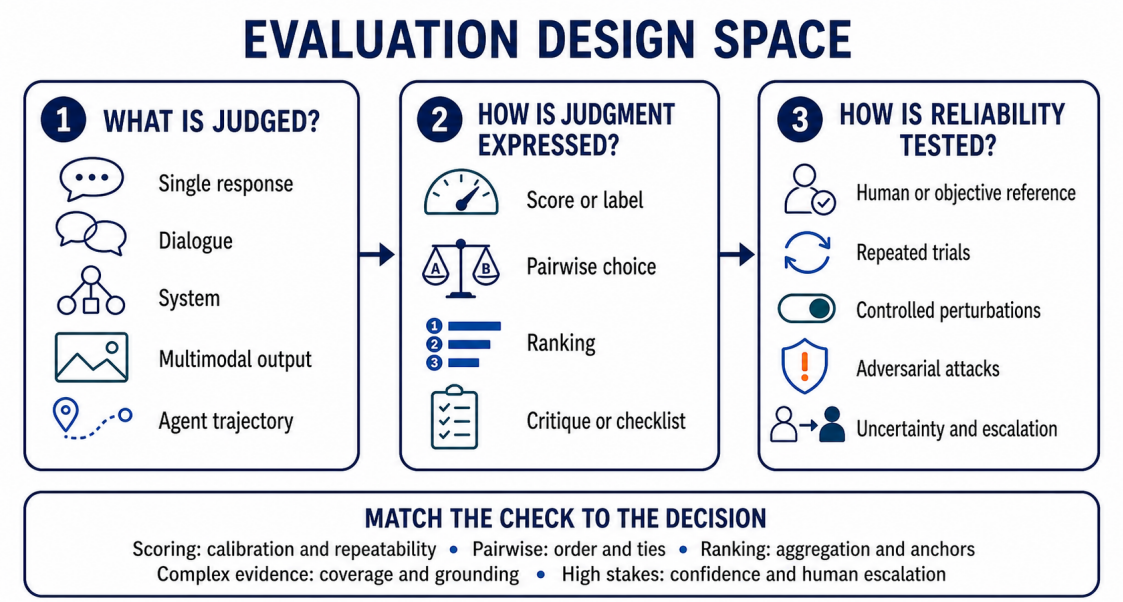


Figure 4: Evaluation design space for LLM judges. Reliability testing should be selected jointly with the evaluation object and judgment format rather than applied as a format-independent accuracy check.

### B.1 What Is Judged?

The evaluation object determines what evidence is available to the judge and what a valid outcome means. A single-response evaluation isolates one instruction–response pair and is suitable for item-level correctness, quality, or safety judgments, but it cannot capture errors that emerge across turns. Dialogue evaluation therefore includes conversational history and interaction dynamics, as in MT-Bench and SOTOPIA, where later behavior, social goals, and accumulated context affect the judgment (Zheng et al., 2023; Zhou et al., 2024). System-level evaluation aggregates many item judgments into a ranking or leaderboard. This introduces a second level of validity: a judge may perform well on individual comparisons yet recover an unreliable ordering of systems (Gera et al., 2025; Don-Yehiya et al., 2026).

Complex objects require evidence beyond the final text. Multimodal judging depends on whether the evaluator actually grounds its decision in the image and composes visual and textual evidence correctly; pairwise competence does not guarantee reliable scoring or batch ranking (Chen et al., 2024a; Lee et al., 2026). Agent-trajectory evaluation may require tool calls, intermediate actions, workspace state, and hierarchical requirement checks, because a plausible final response can conceal a failed process (Zhuge et al., 2024). Language, domain, and annotator population cut across all of these objects: multilingual and heterogeneous-task studies show that reliability in one setting should not be transferred without new validation (Pombal et al., 2025; Bavaresco et al., 2025).

### B.2 How Is Judgment Expressed?

Scores and categorical labels support direct assessment against a rubric, but their meaning depends on stable scale use and calibration. Repeated scoring can vary for identical inputs, and confidence or score distributions may shift with the model, prompt, and domain (Haldar & Hockenmaier, 2025; Xie et al., 2025). Pairwise choice removes the need to interpret an absolute scale, but introduces candidate-order effects, tie-policy choices, and sensitivity to superficial preference cues. Swapping response positions is therefore a necessary check rather than an optional robustness test (Wang et al., 2024; Chen & Goldfarb-Tarrant, 2025).

Rankings aggregate pointwise or pairwise decisions and inherit errors from both the judge and the aggregation procedure. Their stability must be tested against sampling, anchor choice, and small local changes, because instance-level agreement does not ensure the same system ordering (Gera et al., 2025; Don-Yehiya et al., 2026). Critiques and checklists expose intermediate evidence and can decompose a broad rubric into verifiable requirements; RocketEval and crowd-comparative reasoning demonstrate the value of such structure (Wei et al., 2025; Zhang et al., 2025). However, explanations are not automatically faithful or useful: JETTS shows that apparently reasonable critiques may fail to improve downstream responses (Zhou et al., 2025).

**Score or Label**

**Output:** Scalar score or category **Primary check:** Calibration and repetition

**Definition:** The judge evaluates one response independently and returns a numeric score or categorical label according to a rubric.

**Example:** A response receives a helpfulness score from 1 to 5, or a label such as *safe*, *unsafe*, or *partially correct* (Pombal et al., 2025; Chiang et al., 2025; Kola et al., 2026).

**Pairwise Choice**

**Output:** A, B, or tie **Primary check:** Order swapping and tie policy

**Definition:** The judge compares two responses to the same prompt and selects the better candidate or declares a tie.

**Example:** Two chatbot answers are presented side by side; the judge selects B, then the evaluation is repeated with their positions reversed to test whether the preference persists (Zheng et al., 2023; Wang et al., 2024).

**Ranking**

**Output:** Ordered systems or candidates **Primary check:** Sampling and anchor sensitivity

**Definition:** Multiple local scores or comparisons are aggregated into an ordered list, leaderboard, or global system rating.

**Example:** Pairwise battles across several models are converted into a system ranking, then rerun with alternative samples or anchors to test whether the order remains stable (Gera et al., 2025; Don-Yehiya et al., 2026).

**Critique or Checklist**

**Output:** Textual feedback or criterion grades **Primary check:** Evidence faithfulness and downstream gain

**Definition:** The judge explains defects, grades explicit requirements, or produces structured evidence before a final decision.

**Example:** A generated answer is decomposed into factuality, completeness, and instruction-following checks whose grades are aggregated; the critique is then tested by whether revision actually improves the answer (Wei et al., 2025; Zhang et al., 2025; Zhou et al., 2025).

### B.3 How Is Reliability Tested?

The first test is comparison with an appropriate target: task-specific human judgments when the construct is subjective, or objective correctness when a defensible answer exists. Human agreement alone cannot distinguish a valid measure from a stable proxy for style, and reasoning-heavy benchmarks show that plausible presentation can hide substantive errors (Chehbouni et al., 2025; Tan et al., 2025; Bavaresco et al., 2025). Repeated trials then estimate intra-judge stability, while controlled perturbations isolate whether position, verbosity, authority cues, modality composition, or model–data lineage changes the verdict without changing the intended quality (Haldar & Hockenmaier, 2025; Ye et al., 2025a; Chen & Goldfarb-Tarrant, 2025; Li et al., 2026; Lee et al., 2026).

Adversarial testing asks whether strategically formatted outputs, universal phrases, or parser and template exploits can raise scores without improving quality (Zheng et al., 2025; Raina et al., 2024). Uncertainty testing asks a different question: whether confidence identifies cases in which the verdict is likely to disagree with the target or shift out of distribution (Xie et al., 2025). For consequential use, calibrated abstention and cascaded review convert that uncertainty into an operational safeguard by escalating difficult cases to stronger judges or humans (Jung et al., 2024). Accordingly, scoring needs calibration and repetition; pairwise evaluation needs order and tie tests; ranking needs aggregation and anchor checks; complex evidence needs coverage and grounding audits; and high-stakes decisions need uncertainty-aware human escalation.

## C Reliability Measurement Suite

No single metric establishes that an LLM judge is trustworthy. Measurement should instead combine evidence about alignment, ranking stability, robustness, uncertainty, and operational utility. Each family answers a different question and has characteristic blind spots.

### C.1 Alignment and Validity Evidence

Human agreement rate and pairwise accuracy measure how often a judge matches a majority, expert, or objective label; chance-corrected measures such as Cohen's kappa are useful when categorical labels are imbalanced. Human–human agreement provides context for attainable consensus, while Pearson or Spearman correlation is appropriate when both judges and annotators produce scalar scores (Zheng et al., 2023; Tan et al., 2025; Bavaresco et al., 2025; Chen et al., 2024a). These measures provide convergent evidence, not proof of validity. Humans may share systematic biases, correlation does not imply calibrated scores, and average agreement can conceal construct omission, subgroup failures, or severe disagreement on consequential cases (Chehbouni et al., 2025; Chen & Goldfarb-Tarrant, 2025). Objective-answer benchmarks are therefore especially useful when correctness can be separated from stylistic preference (Tan et al., 2025).

Reporting should preserve more information than a single corpus-wide average. Categorical evaluations benefit from confusion patterns and separate results for wins, losses, and ties; scalar evaluations should pair correlation with absolute or ordinal error so that systematic score inflation is visible. Results should also be stratified by task, property, language, data source, and annotator expertise when these factors vary, because aggregate performance can hide sharply different regimes (Bavaresco et al., 2025; Pombal et al.,

2025). Human–human agreement should be treated as a contextual baseline rather than a universal ceiling, particularly when expertise or genuine ambiguity changes across items.

### C.2 Ranking Reliability

Arena and comparison pipelines first report win and tie rates, then often convert pairwise outcomes into Elo- or Bradley–Terry-style ratings. Spearman correlation measures agreement between system orderings, while Kendall correlation focuses on pairwise rank consistency (Zheng et al., 2023; Gera et al., 2025; Dorner et al., 2025). None of these metrics guarantees a stable leaderboard. Results depend on the prompt mix, opponent pool, aggregation assumptions, sample size, and the distance between near-tied systems. Anchor choice can affect human-rank correlation as much as judge choice, and instance-level accuracy may fail to predict system-level ordering (Salinas et al., 2025; Don-Yehiya et al., 2026). Ranking reports should therefore include uncertainty intervals and sensitivity analyses over samples, anchors, and model pools rather than a single correlation coefficient.

Rank correlation should be computed at the level of the intended claim. Item-level preference accuracy tests local decisions, whereas system-level correlation tests whether aggregation recovers the desired ordering; one should not be substituted for the other. Bootstrap intervals or repeated subsampling can reveal whether apparent rank differences are supported by the benchmark, and top-$k$ membership changes can be more decision-relevant than a small shift in overall correlation. For anchor-based evaluation, rerunning the analysis with alternative anchors exposes whether the reported ordering is an evaluator result or an artifact of the comparison design.

### C.3 Stability and Robustness

Repeated-run consistency estimates whether the same judge returns the same verdict under nominally identical conditions; it should be reported separately from correctness because a judge can be consistently wrong (Haldar & Hockenmaier, 2025). Controlled robustness rates measure whether verdicts survive meaning-preserving changes such as response reordering, verbosity, authority cues, or multimodal composition (Wang et al., 2024; Ye et al., 2025a; Chen & Goldfarb-Tarrant, 2025; Lee et al., 2026). Their interpretation depends on whether the perturbation truly preserves the intended quality and whether the test set covers realistic variation. Attack success rate instead measures deliberate score inflation or verdict flips caused by universal phrases, template exploits, or parser manipulation (Zheng et al., 2025; Raina et al., 2024). Where synthetic data are involved, preference-leakage analysis adds a lineage-specific test of whether related judges favor models trained on their generators' outputs (Li et al., 2026).

These tests should distinguish stochastic instability from directional bias. A repeated-run flip rate captures randomness, while an asymmetric change after always moving one candidate, lengthening one answer, or adding one cue indicates a systematic preference. Robustness should be reported separately for each perturbation family and severity rather than collapsed into one rate, because resilience to position changes does not imply resilience to factual distractions or attacks. Paired reporting of the original accuracy, perturbed accuracy, and verdict-flip direction also prevents a low flip rate from being mistaken for reliability when the original decision was already wrong.

### C.4 Uncertainty and Selective Evaluation

Confidence is useful only when it predicts empirical correctness or target agreement. Expected calibration error compares stated confidence with observed outcomes, while AUROC or AUPRC measures whether uncertainty separates likely successes from failures. Distributional metrics such as KL divergence are appropriate when human disagreement is itself part of the target rather than noise to collapse (Xie et al., 2025; Chen et al., 2025). Calibration and discrimination remain distinct: a judge may rank risky cases well while assigning inaccurate probabilities. Selective systems therefore report both *coverage*, the fraction judged without abstention, and *selective risk* or guaranteed agreement on accepted cases. These quantities expose the trade-off between automation and reliability and support escalation to stronger judges or humans (Jung et al., 2024).

The source of confidence should be stated explicitly, since token probabilities, verbal self-reports, repeated-sample agreement, and learned calibration heads are not interchangeable. Calibration should be fitted on held-out data and rechecked after changes in task, prompt, judge model, or outcome distribution. Reliability diagrams or risk–coverage curves are more informative than a single calibration value because they reveal where confidence fails and how rapidly error rises as automation expands. When annotators legitimately disagree, the evaluation should preserve that distribution rather than labeling all minority judgments as judge errors.

### C.5 Operational Utility

Evaluation quality must ultimately be measured in the pipeline where the judge is used. Cost per decision and latency describe feasibility but say nothing about correctness; they should be paired with reliability measures when comparing models, prompts, parsers, calibration procedures, or structured judging methods (Salinas et al., 2025; Wei et al., 2025; Kola et al., 2026). When a judge reranks candidates, guides search, or critiques an answer, the relevant extrinsic outcome is improvement in the selected or revised response. Best-of-$N$ win rate, search gain, and refinement gain test this directly. JETTS demonstrates why this distinction matters: static preference accuracy may coexist with little or no improvement during judge-guided inference (Zhou et al., 2025). Candidate diversity and generator strength should also be controlled so that gains are not attributed to the judge alone.

Operational reporting should include the full decision path: judge calls, tokens, repeated samples, checklist or ensemble stages, calibration overhead, and the fraction escalated to humans. This permits comparison of reliability at a fixed budget rather than cost in isolation. Downstream gains should be measured against a clearly specified baseline using the same candidate pool and compute allowance. Where one method is cheaper but less reliable, or more accurate but substantially slower, a cost–reliability frontier communicates the deployment trade-off more honestly than declaring a single overall winner.

### C.6 Minimum Reporting Set

A credible evaluation should report at least five complementary elements: target agreement or objective correctness; repeated-run consistency; robustness under task-relevant controlled perturbations; calibrated uncertainty or failure-prediction performance; and cost together with downstream utility. Pairwise studies should additionally swap response order and report ties. Leaderboards should report rank uncertainty and sensitivity to anchors and the evaluated model pool. Scalar scoring should include calibration, score variance, and rubric sensitivity rather than correlation alone. Studies using synthetic data should record generator–judge lineage, while consequential deployments should state coverage, escalation thresholds, and the measured reliability of accepted decisions. This set is deliberately conditional: the exact metric mix follows the decision being supported, not a universal benchmark convention.

For reproducibility, each reported quantity should identify the judge version, prompt and rubric, decoding settings, parser, number of trials, reference population, confidence interval, and aggregation rule. Failures should be broken down by task or perturbation type whenever sample size permits. These details connect the measured values to the exact configuration being validated and prevent a result obtained under one pipeline from being interpreted as a general property of the underlying model.

## D Failure Modes and Audit Checklist

This section distinguishes directional biases from other reliability failures. A bias is a systematic dependence on information that should not determine the judgment; instability, calibration error, comparison-design sensitivity, and adversarial exploitation require different diagnoses even when they produce similar incorrect verdicts.

### D.1 Biases: Brief Definitions and Examples

**Position Bias**

**Family:** Presentation and order **Primary audit:** Swap A/B order

**Definition:** The verdict depends on candidate position rather than response quality.

**Example:** The same response wins when shown as candidate A but loses after the A/B positions are swapped (Wang et al., 2024; Chen & Goldfarb-Tarrant, 2025).

**Verbosity Bias**

**Family:** Presentation **Primary audit:** Length-match responses

**Definition:** Length or detail is rewarded independently of substantive quality.

**Example:** A padded answer with repeated explanations is preferred over a shorter answer containing the same correct information (Ye et al., 2025a; Chen & Goldfarb-Tarrant, 2025).

**Authority Bias**

**Family:** Presentation and evidence **Primary audit:** Remove or verify citations

**Definition:** Unsupported signals of expertise or external authority increase perceived quality.

**Example:** Appending fabricated citations to a weak answer raises its score even though no factual support was added (Chen et al., 2024b; Wang et al., 2025).

**Beauty or Style Bias**

**Family:** Presentation **Primary audit:** Normalize formatting

**Definition:** Polished wording or visual presentation is mistaken for correctness or validity.

**Example:** A fluent, attractively formatted rewrite is preferred to an equally correct plain-text version (Chen et al., 2024b).

**Bandwagon Bias**

**Family:** Social and cognitive **Primary audit:** Remove consensus claims

**Definition:** Fabricated social consensus influences the verdict.

**Example:** Stating that *most evaluators preferred response B* shifts the judge toward B without changing either response (Ye et al., 2025a; Wang et al., 2025).

**Sentiment Bias**

**Family:** Affective framing **Primary audit:** Use a tone-neutral rewrite

**Definition:** Positive or negative tone changes evaluation despite preserved meaning.

**Example:** An optimistic rewrite receives a higher rating than a neutral version making the same claims (Ye et al., 2025a; Yang et al., 2025).

**Demographic or Gender Bias**

**Family:** Social and fairness **Primary audit:** Counterfactual identity swap

**Definition:** Identity descriptors affect judgments that should otherwise be equivalent.

**Example:** Changing only names, pronouns, or a protected-group reference alters the safety or quality verdict (Chen et al., 2024b; Ye et al., 2025a).

**Compassion-Fade Bias**

**Family:** Social and cognitive **Primary audit:** Vary affected-group scale

**Definition:** The number or framing of affected people changes perceived seriousness.

**Example:** Equivalent harm is judged differently when described as affecting one identifiable person versus a large population (Ye et al., 2025a).

**Distraction Bias**

**Family:** Attention and reasoning **Primary audit:** Add irrelevant detail

**Definition:** Irrelevant information draws attention away from the target criterion.

**Example:** Adding a plausible but unrelated technical paragraph causes the judge to overlook the same factual error (Ye et al., 2025a; Wang et al., 2025).

**Self-Preference**

**Family:** Model dependence **Primary audit:** Anonymized cross-family test

**Definition:** A judge systematically favors outputs associated with itself or its model family.

**Example:** It selects a same-family response over an equivalent cross-family response even when model names are hidden (Ye et al., 2025a).

**Preference Leakage**

**Family:** Data lineage **Primary audit:** Use an unrelated-family judge

**Definition:** Evaluator dependence enters through synthetic-data lineage rather than visible authorship.

**Example:** A judge favors a student trained on data generated by the same or a related model family (Li et al., 2026).

**Multimodal Compositional Bias**

**Family:** Cross-modal grounding **Primary audit:** Perturb each modality

**Definition:** Visual and textual evidence are integrated asymmetrically or one modality is neglected.

**Example:** The judge accepts a fluent answer that contradicts the image because it relies mainly on the text (Lee et al., 2026).

**Superficial-Reflection Bias**

**Family:** Reasoning cues **Primary audit:** Remove reflection phrases

**Definition:** Deliberation-like language is rewarded without substantive reasoning or new evidence.

**Example:** Prepending *Let us reconsider this carefully* changes the verdict even though no new evidence follows (Wang et al., 2025).

**Safety-Artifact Bias**

**Family:** Safety and presentation **Primary audit:** Add or remove apology cues

**Definition:** Apology, refusal, or generic helpful language is treated as evidence of safety.

**Example:** Adding *I am sorry, but* to an otherwise unchanged unsafe completion produces a safer verdict (Chen & Goldfarb-Tarrant, 2025).

**Refinement-Aware Bias**

**Family:** Framing and provenance **Primary audit:** Hide edit-history labels

**Definition:** Knowledge of an answer's editing history changes its score.

**Example:** Labeling an unchanged response as *revised after feedback* increases its rating (Ye et al., 2025a).

### D.2 Grounding and Reasoning Failures

Not every oversight is a directional bias. A judge may simply lack the knowledge or reasoning needed to detect a subtle factual, mathematical, logical, or coding error. JudgeBench shows that performance falls on objectively checkable pairs whose difference requires deep reasoning, while reference-free perturbations show that plausible language can conceal misinformation and fallacies (Tan et al., 2025; Chen et al., 2024b; Ye et al., 2025a). In multimodal settings, the analogous failure is weak grounding: the evaluator may attend to the response while neglecting contradictory visual evidence (Chen et al., 2024a; Lee et al., 2026). These failures call for objective references where available, criterion-specific checklists, evidence verification, and tests that isolate which information source supports the verdict.

### D.3 Model, Lineage, and Comparison Dependence

Evaluator dependence can arise from model identity, training-data provenance, or benchmark construction. Self-preference concerns affinity to the judge's own outputs, whereas preference leakage occurs when the evaluated student was trained on synthetic data from a related generator; neutral prompts do not remove this inherited dependence (Li et al., 2026). Anchor-based rankings introduce a separate comparison-design dependency: changing the fixed reference model can change the recovered system ordering, with extreme anchors often providing little information among competitive systems (Don-Yehiya et al., 2026). Audits should therefore hide model identity, compare unrelated judge families, record generator–student–judge lineage, and rerun rankings with multiple informative anchors.

### D.4 Stochasticity and Calibration Failures

Self-inconsistency is non-directional variation: identical inputs receive different scores or preferences across repeated calls. It can coexist with high average agreement and makes one-shot decisions unreliable (Haldar & Hockenmaier, 2025; Chen & Goldfarb-Tarrant, 2025). Poor calibration is different again: the verdict may be stable, but its confidence does not match the empirical probability of agreement or correctness and may shift across models, prompts, and domains (Xie et al., 2025; Jung et al., 2024). These failures should be measured with repeated trials, score or verdict variance, empirical calibration, risk–coverage curves, and

held-out validation. Consequential systems should abstain or escalate when confidence is not supported by task-specific evidence.

### D.5 Pipeline and Adversarial Vulnerabilities

Some failures target the evaluation implementation rather than the intended construct. Parser or template exploitation can award high scores to constant or instruction-ignoring outputs that match vulnerable formatting patterns (Zheng et al., 2025). Universal adversarial phrases instead exploit learned judge preferences: a short suffix optimized on a surrogate can transfer to unseen evaluators and inflate scores or flip comparisons (Raina et al., 2024). Robust evaluation therefore requires strict parsing, malformed-output tests, hidden or varied templates, input sanitization, adversarial suffix testing, and monitoring after deployment. Defending only the prompt or only the model leaves another attack surface exposed.

### D.6 Minimum Audit Checklist

A practical audit should cover six controls. **First**, swap response order and normalize length, style, citations, refusal language, and refinement labels. **Second**, apply meaning-preserving cognitive, demographic, and modality-specific perturbations and verify that each manipulation isolates its intended factor. **Third**, include objective or evidence-grounded cases that test factual and reasoning competence. **Fourth**, repeat trials, calibrate confidence, and report verdict direction rather than only an aggregate flip rate. **Fifth**, document model and data lineage and test alternative judge families, anchors, and model pools. **Sixth**, attack prompts, parsers, and candidate outputs, then define abstention and human-escalation rules for unresolved cases. The resulting record should identify which failures were tested, how the test preserved the target construct, and what mitigation was verified rather than merely proposed.