

A Statistical Framework for LLM-as-a-Judge Metrics

Reliability, Self-Preference and Metric Redundancy of LLM-as-a-Judge Metrics for Multi-Turn Agents

Jenny Lee

Master’s Research Paper Proposal, October 2026

1 Motivation

LLM-powered agents answer across turns, retrieve documents and call tools. Measuring the resulting performance is difficult, and people can disagree on its quality. LLM-as-a-Judge offers a practical way to score such outputs [1,2], but the score is itself a measurement. It can change when the judge, prompt wording or repeated call changes, and consistent judgments can still be biased [3,4].

I propose an open-source pipeline for examining how much confidence we can place in evaluation scores. The research will connect three questions: how reliably a metric scores an answer, whether some metrics largely repeat the information in others, and how many human labels are needed to estimate a human-evaluated mean with a chosen precision. Self-preference will be studied as a source of systematic differences between judges.

2 Data Resources

The main public resource is MTRAG [5]: 110 human-written conversations, averaging 7.7 turns, with 842 turn-level tasks over four document collections. Each task provides the earlier conversation and current question, with reference or retrieved passages depending on the setting. MTRAG-UN adds 666 tasks addressing unanswerable and underspecified questions [6]. For action conversations, an additional resource is τ^2 -bench [7], where tool calls can change an environment state.

The pipeline will take metric scores and their conversation, turn, agent, judge, prompt and run identifiers, together with the relevant evaluation inputs. In MTRAG, a candidate answers the current question using a supplied conversation history; the reference answer is used only where the metric requires it. In live agent conversations, the history may depend on the agent’s own earlier answers. I will keep these settings separate in the analysis and use an action-oriented resource for tool execution.

3 Metrics

The metrics below are examples, as are evaluation tools such as Ragas and DeepEval and methods such as G-Eval [8,1]. The pipeline will use the actual score definition and inputs of each metric, including whether a reference answer is needed.

Example metric	Meaning	Score type
Groundedness	Share of answer claims supported by the passages	Proportion
Answer relevance	Whether the answer addresses the current question in context	Depends on metric
Context coherence	Whether the answer is consistent with relevant earlier turns	Ordinal, e.g. 1–5
Appropriate abstention	Whether the answer appropriately acknowledges insufficient evidence	Binary
Task completion	Whether tool actions achieve the user’s goal	Binary

Possible open-weight judges include Qwen, Llama and Gemma. Metric definitions will specify how scores are produced: task completion may use an executable check, groundedness requires a rule for counting claims and handling answers with no claims, and abstention must distinguish justified refusal from unnecessary refusal.

4 Research Questions

The research paper has three parts, with Part A as the main focus:

- **A. Reliability and self-preference:** How reliable are scores across judges, prompts and repeated calls? What can a small panel tell us about a broader judge population, and how might model-family matching affect scores?
- **B. Metric redundancy:** Which metrics provide little additional information because their scores can be recovered from other metrics?
- **C. Human-label requirements:** How many human labels might be needed to estimate a human-evaluated mean with a chosen precision, given the labels already available?

5 Part A: Reliability Estimation and Self-Preference

5.1 Measurement Model and Reliability

The same answer can receive different scores when the judge or prompt changes, or when a judgment is repeated. I want to separate variation between answers from variation introduced by the evaluation process, using generalizability theory as a starting point [9,10].

For one metric and one agent, let Y_{ijpr} be the score for answer $i = (c, t)$, at turn t of conversation c , from judge j , prompt p and call r . A simplified starting model is

$$Y_{ijpr} = \mu + \alpha_c + \beta_{t:c} + \gamma_j + \omega_p + e_{ijpr}. \quad (1)$$

Here μ is the overall mean, $\alpha_c + \beta_{t:c}$ represents differences between answers, γ_j represents judge severity, and ω_p represents prompt effects. The remaining variation is represented by e_{ijpr} . I would examine whether this needs to be separated further into answer-judge interactions, prompt interactions and variation between repeated calls. Grouping turns within conversations reflects their shared context.

Writing τ^2 for the variance of the answer effects, the usual relative and absolute reliability coefficients take the form

$$E\rho^2 = \frac{\tau^2}{\tau^2 + \sigma_\delta^2}, \quad \Phi = \frac{\tau^2}{\tau^2 + \sigma_\Delta^2}. \quad (2)$$

For the score average being evaluated, σ_δ^2 measures error affecting comparisons between answers, while σ_Δ^2 also includes shifts in absolute score levels. For example, a judge may score consistently higher while preserving the ordering of answers. I want to study how these error variances depend on the numbers of judges, prompts and repeated calls. The answer effects describe expected scores under the chosen evaluation conditions; these coefficients do not directly measure agreement with humans.

5.2 Reliability with Limited Judge Panels

My main theoretical interest is how uncertain reliability estimates remain when only a few judges are available. Scoring more conversations helps us learn about the observed judges, but may leave substantial uncertainty about judges outside the panel.

I plan to begin with a simple variance-components model and study uncertainty in $\hat{\Phi}$ as the number of conversations increases while the panel stays small. I would then consider what changes when more judges are available. Extending the results to a broader judge population requires specifying that population and how the observed panel represents it.

The aim is to connect this analysis to confidence intervals for reliability and to decisions about adding judges, prompts or repeated calls. This would let the pipeline show both the estimated benefit of a judging configuration and the uncertainty in that estimate.

5.3 Relationship to Existing Methods

Messing studies uncertainty in evaluation pipelines using selected judges [3]. CyclicJudge considers evaluation allocation within a judge pool [11], while Choi et al. examine reliability through item response theory [12]. My focus is uncertainty in reliability estimates when few judges are available and answers come from multi-turn conversations.

I would begin with continuous scores, then consider how the analysis should change for ordered ratings, binary outcomes and proportions. The model and estimation method would depend on the score definition and available data.

5.4 Examining Family-Conditioned Self-Preference

A judge could score consistently while favoring outputs from its own model family. Building on work on family-conditioned preference [4], I want to compare scores from matching and non-matching judge-agent families while accounting for general differences between judges and outputs.

Let $s_{ja} = 1$ when judge j and agent a share a model family, and 0 otherwise. Writing θ_{ia} for the effect of agent a 's answer to task i , one possible extension is

$$Y_{iajpr} = \mu + \theta_{ia} + \gamma_j + \omega_p + \psi s_{ja} + u_{ja} + e_{iajpr}, \quad (3)$$

where ψ represents the family-matching effect and u_{ja} represents other judge-agent differences. Separating these effects requires assumptions about u_{ja} and enough matching and non-matching comparisons. I would investigate what the available judge-agent combinations allow us to conclude; a higher score from a matching judge alone would not establish self-preference.

6 Part B: Metric Redundancy and Incremental Information

I want to identify metrics whose scores largely repeat information already available from other metrics. This part would use metric scores themselves, without requiring human quality labels.

Pairwise correlations would provide an initial view of overlap. I would then examine whether several metrics together can predict another metric, drawing on reconstruction-based feature selection [13]. Regression is a possible starting point. For a continuous metric m , one way to summarize prediction error is

$$D_m = \frac{\sum_{i \in \mathcal{T}} \{x_{im} - \hat{g}_m(\mathbf{x}_{i,-m})\}^2}{\sum_{i \in \mathcal{T}} (x_{im} - \bar{x}_{m,\text{train}})^2}. \quad (4)$$

Here x_{im} is the score for metric m , $\mathbf{x}_{i,-m}$ contains the other scores, and $\hat{g}_m$ predicts m from them. Both the predictor and the baseline mean $\bar{x}_{m,\text{train}}$ are fitted without the evaluation conversations in $\mathcal{T}$. When the denominator is positive, D_m near zero indicates that the other metrics recover most of this metric's variation; a value near or above one indicates little improvement over the constant baseline.

I would consider suitable prediction measures for other score types and interpret the results alongside Part A. Poor prediction may reflect inconsistent scoring or a limitation of the predictor, as well as information that other metrics miss.

The pipeline would report overlap and suggest smaller metric sets, checking whether retained metrics can still recover omitted scores. I would also consider the purpose of each metric: statistical overlap alone may not justify removing one designed to detect an uncommon failure.

7 Part C: Human-Label Requirements for Target Precision

Judge scores may be available for many answers while human labels are scarce. I want to estimate how many labels might be needed to measure a human-evaluated mean score or win rate with a chosen precision. The pipeline should help plan both the total count and the additional labels needed.

I plan to explore prediction-powered inference (PPI), which combines judge predictions with paired human labels [15]. For mean estimation, it adjusts the average judge prediction by the average human–judge difference. Let Y be a human label, X the evaluation input, and $f(X)$ a judge prediction on the same scale. A simple starting case uses a fixed f , N judge-only observations and n separate paired observations, all sampled independently from the same target population:

$$\hat{\theta}_{\text{PPI}} = \underbrace{\bar{f}_N}_{\text{mean judge prediction}} + \underbrace{\overline{(Y - f)}_n}_{\text{human-based correction}}. \quad (5)$$

The two sources of uncertainty give

$$\text{Var}(\hat{\theta}_{\text{PPI}}) = \frac{v_f}{N} + \frac{v_e}{n}, \quad v_f = \text{Var}(f(X)), \quad v_e = \text{Var}(Y - f(X)). \quad (6)$$

Thus, lower variability in human–judge differences can reduce the label requirement. I want to use this relationship to study how the count changes with the quality of judge predictions and the desired precision.

For confidence level $1 - \alpha$ and interval half-width h , a normal-approximation starting calculation is

$$n_{\text{req}} \approx \left\lceil \frac{v_e}{(h/z_{1-\alpha/2})^2 - v_f/N} \right\rceil, \quad n_{\text{add}} = \max(0, n_{\text{req}} - n_0), \quad (7)$$

where $z_{1-\alpha/2}$ is the standard normal cutoff and n_0 is the number of usable existing labels. The denominator must be positive; otherwise, increasing human labels alone cannot meet the target with N fixed in this estimator.

Available representative pairs could provide an estimate of v_e . With no human labels, I would show counts under stated assumptions instead. A sensitivity analysis would repeat the calculation with lower and higher plausible values of v_e , showing how much the recommendation changes. Reliability Φ alone cannot determine this variance.

I would also examine uncertainty in these variance estimates and how the calculation changes when several labeled turns share a conversation. MTRAG’s public human ratings provide a possible resource for checking the calculations [14]. Applying them elsewhere would require assumptions or estimates appropriate to that population.

8 Expected Outcome: An Open-Source Analysis Pipeline

The main practical deliverable is an open-source pipeline accepting score tables from Ragas, DeepEval, G-Eval-based implementations or custom judges. The input would include scores and their conversation, turn, agent, judge, prompt and run identifiers.

Part A would report reliability, uncertainty and possible self-preference. Part B would describe metric overlap and suggest candidate subsets. Part C would estimate total and additional human-label requirements under stated assumptions. Together, these outputs would support decisions about judging configurations, metric choice and label collection. The same questions could also arise in essay scoring and code review, where the relevant assumptions would need to be considered.

References

- [1] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. *EMNLP*, 2023. arXiv:2303.16634.
- [2] L. Zheng et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS, Datasets and Benchmarks*, 2023. arXiv:2306.05685.
- [3] S. Messing. *Hidden Measurement Error in LLM Pipelines Distorts Annotation, Evaluation, and Benchmarking*. 2026. arXiv:2604.11581v6.
- [4] D. A. Awuni, L. E. K. Achenie, B. T. Partey, E. G. Owusu, and N.-N. D. Sowah. *Who Judges Matters: Measuring Family-Conditioned Preference in LLM-as-Judge Panels*. 2026. arXiv:2609.17857v1.
- [5] Y. Katsis et al. *MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems*. 2025. arXiv:2501.03468.
- [6] S. Rosenthal, Y. Katsis, V. Shah, et al. *MTRAG-UN: A Benchmark for Open Challenges in Multi-Turn RAG Conversations*. 2026. arXiv:2602.23184.
- [7] Sierra Research. τ^2 -bench. 2025. <https://github.com/sierra-research/tau2-bench>.
- [8] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. *EACL: System Demonstrations*, 2024. arXiv:2309.15217.
- [9] R. L. Brennan. *Generalizability Theory*. Springer, 2001.
- [10] R. L. Brennan and M. T. Kane. An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3):277–289, 1977.
- [11] Z. Zhu, O. Tieleman, A. Bukhtiyarov, and J. Chen. *CyclicJudge: Mitigating Judge Bias Efficiently in LLM-based Evaluation*. 2026. arXiv:2603.01865v3.
- [12] J. Choi, S. Park, C. Cho, H. Park, and B. Kim. *Diagnosing the Reliability of LLM-as-a-Judge via Item Response Theory*. 2026. arXiv:2602.00521v1.
- [13] J. Li, J. Tang, and H. Liu. Reconstruction-based unsupervised feature selection: An embedded approach. *Proceedings of IJCAI*, pp. 2159–2165, 2017. doi:10.24963/ijcai.2017/300.
- [14] IBM. *MTRAG: Human Data and Human Evaluation Reference Subset*. Official dataset documentation.
- [15] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnica. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. arXiv:2301.09633.