When Should Agents Evaluate? A Principled Framework for Human-AI Task Allocation in Language Model Evaluation

Sasha Mitts
Independent Researcher
United States
amitts18@gmail.com

Abstract

AI agents are increasingly deployed as evaluation tools for language models, promising scale and consistency that human evaluation cannot match. Yet emerging evidence reveals systematic failures: LLM judges exhibit pronounced biases in precisely the alignment and safety domains where reliable evaluation matters most, and their agreement with human judgment deteriorates on tasks requiring contextual or normative sensitivity. This paper argues that these failures reflect a deeper issue. Certain evaluation tasks derive their value from being situated in representative human psychology, and current agent architectures cannot replicate this property. I propose a principled framework for task allocation between human evaluators and AI agents, grounded in the distinction between *evaluation-as-measurement*, where agents excel, and *evaluation-as-understanding*, where human contextualization is epistemically necessary. I identify specific task characteristics that predict which mode applies and outline design principles for hybrid evaluation systems that preserve the epistemic contributions of each.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

LLM evaluation, agents-in-the-loop, human-AI collaboration, task allocation, meta-evaluation

ACM Reference Format:
Sasha Mitts. 2026. When Should Agents Evaluate? A Principled Framework for Human-AI Task Allocation in Language Model Evaluation. In *CHI '26 Workshop on HEAL*. ACM, New York, NY, USA, 4 pages.

1 Introduction

The use of AI agents as evaluation tools for language models has grown rapidly. LLM-as-judge approaches, in which a strong language model scores or compares outputs from other models, have become standard practice in both research and industry [12]. Multi-agent debate frameworks improve on single-judge reliability [4]. Agent-based auditing systems can automate red-teaming, behavioral evaluation, and alignment testing at scales that would be infeasible for human evaluators alone [1]. Agents offer speed, consistency, and scalability that human evaluation cannot match.

Yet emerging evidence suggests that agent evaluation exhibits systematic failures in precisely the domains where it matters most. Quantitative studies have identified at least twelve distinct bias types in LLM judges, and these biases are *more pronounced in alignment and safety evaluations* than in factual or objective tasks [8]. LLM judges exhibit self-preference bias, systematically overrating outputs from their own model family [11]. On challenging evaluation tasks requiring contextual or normative judgment, strong LLM judges perform only marginally better than random guessing [9]. Preference leakage, in which evaluator models are biased toward related generators, introduces contamination that is difficult to detect [7]. A systematic study across twenty NLP evaluation tasks found that LLM judges are not yet ready to replace human evaluators, with performance varying sharply by task type [2].

These findings are typically framed as problems to be solved through better prompting, debiasing, or multi-agent designs. I argue they reflect a structural limitation, and the contribution of this paper is a framework that explains *why* these failures cluster where they do and *when* human evaluation is epistemically necessary rather than merely preferred. Certain kinds of evaluation derive their epistemic value from being situated in human psychology, contextualized by lived experience, and responsive to the gap between stated criteria and operative values. Agent evaluators can scale the *measurement* dimension of evaluation, and they do so effectively. The *understanding* dimension requires something they cannot provide. This paper proposes a principled framework for distinguishing between these two modes and allocating evaluation tasks accordingly, speaking directly to the HEAL workshop's special theme on AI agents-in-the-loop.

2 The Epistemic Structure of Evaluation

2.1 Two Modes of Evaluation

I propose that evaluation of language model outputs operates in two epistemic modes, which I call *evaluation-as-measurement* and *evaluation-as-understanding*. This distinction has roots in psychometric theory: Cronbach and Meehl's foundational work on construct validity [5] established that measuring a construct requires more than assigning numbers against criteria—it requires understanding whether those numbers reflect the phenomenon they claim to capture.

Evaluation-as-measurement assesses outputs against pre-specified criteria with clear decision boundaries. Does the

CHI '26 Workshop, Yokohama, Japan
2026.

model produce syntactically valid code? Does the response contain factual inaccuracies that can be verified against a knowledge source? Is the output within an acceptable length range? These evaluations are well-suited to agent automation because the criteria are explicit, the judgment requires pattern matching or verification against known standards, and the evaluator's own psychology is irrelevant to the result. The value of the evaluation comes from its accuracy and coverage, both of which agents can provide at scale.

Evaluation-as-understanding assesses outputs in relation to situated human experience, expectations, and values. Does this response feel helpful in the context a user would actually encounter it? Does the model's tone match what a patient in a clinical setting would find reassuring or alienating? Is this explanation genuinely clarifying, or does it merely sound authoritative? These evaluations are epistemically different. The criteria are often not fully specifiable in advance; they emerge from the evaluator's contextualized response to the output. The evaluator's psychology is not noise to be eliminated but signal to be captured. The value of the evaluation comes from its representativeness of the human experience it aims to characterize.

This distinction cuts across task difficulty. An agent can solve a complex coding verification problem that a non-expert human cannot. What matters is what makes the evaluation *epistemically valid*: whether validity derives from accuracy against external criteria, or from fidelity to situated human experience.

In practice, measurement and understanding exist on a continuum rather than as a clean binary. Many evaluation tasks sit in a middle region where either mode could apply depending on context. The principled question for these cases is not whether a sufficiently capable agent *could* perform the evaluation, but whether the deployment context is stable enough for measurement to remain calibrated. AI systems change what users expect from technology; societal norms around AI-mediated interaction are actively forming; what counts as "helpful" or "appropriate" shifts as both capabilities and cultural attitudes evolve. When the context of use changes with enough velocity or unpredictability, evaluation-as-understanding becomes necessary to keep measurement-mode criteria aligned with the phenomena they aim to capture—even for tasks that could in principle be collapsed to measurement under static conditions. The choice between modes is therefore driven by the motivations of the system being built and what each mode affords: scale and consistency from measurement, or contextual fidelity and recalibration capacity from understanding. How to operationalize the velocity of contextual change—whether through empirical indicators like user feedback drift, norm shifts in deployment populations, or capability-driven expectation changes—remains an open question that future empirical work should address.

2.2 Why Agent Evaluators Struggle at Understanding

The empirical patterns in the LLM-as-judge literature become legible through this framework. Agent evaluators perform well on factual verification, code correctness, and other measurement tasks. They struggle on alignment, safety, and quality-of-experience evaluations [8, 9]. This is exactly what the framework predicts: alignment and safety evaluations are instances of evaluation-as-understanding, where the relevant criteria are contextual, culturally situated, and responsive to human psychology. When researchers observe that LLM judges exhibit more bias in alignment tasks than in factual tasks, they are observing the boundary between the two epistemic modes.

Self-preference bias [11] and preference leakage [7] further illustrate the point. These biases arise because agent evaluators respond to statistical familiarity (low perplexity) rather than to the situated quality of an output. An agent judge that rates familiar-sounding text higher is measuring statistical familiarity, which diverges systematically from experienced quality. Human evaluators carry biases too, but those biases operate within the same psychological space as end users. This makes human evaluation biases informative about real-world reception in ways that agent biases cannot be: a human evaluator who finds a response condescending is evidence about how users will receive it; an agent evaluator that flags "condescending tone" is applying a pattern classifier. This claim holds, however, only when evaluators are drawn from the population whose experience the evaluation aims to characterize. A crowdsourced annotator's reaction to a clinical AI assistant is no more representative of clinician experience than an agent's (see Section 3.3).

Prior work on human-centered AI evaluation has demonstrated this dynamic empirically. Mitts [10] found that expert annotator judgments diverged systematically from end-user evaluations: annotation-based evaluation indicated strong competitive wins, while user evaluations showed ties or losses. The evaluation framework that resolved this gap emerged from qualitative reflection with users and domain experts, surfacing quality dimensions that were not known in advance. This is evaluation-as-understanding producing results that evaluation-as-measurement could not.

2.3 The Atomic Decomposition Objection

A strong counterargument holds that any evaluation task can be decomposed into sufficiently atomic steps, each of which an agent can perform: read the output, propose a coding schema, apply the schema, revise it, assess explanatory power, and so on. If the individual cognitive operations are within agent capability, why can't agents handle evaluation-as-understanding through recursive decomposition?

The answer is that atomic decomposition changes what the evaluation *measures*. Consider evaluating whether a mental health chatbot's response would feel supportive to a user in crisis. An agent can check for empathetic language, assess tone against a rubric, and verify adherence to clinical

guidelines. Each atomic step is executable. The evaluation that results from composing these steps, however, measures conformity to a specification of supportiveness. Experienced supportiveness is a different phenomenon. The epistemic gap lies in what grounds validity: when a human evaluator from the target population reports feeling supported or dismissed, that response is constitutive evidence. When an agent applies a supportiveness rubric, it is measuring a proxy, however sophisticated.

This distinction matters practically. Agents performing atomic evaluation steps are valuable for identifying *candidate* issues, proposing analytic schemas, and ensuring coverage. These are measurement-mode contributions to what may be an understanding-mode evaluation. The framework I propose does not exclude agents from understanding-oriented evaluation pipelines. It specifies which steps within those pipelines derive their validity from human psychology and which do not.

3 A Framework for Task Allocation

The distinction between measurement and understanding is useful only if it can be operationalized. I propose four task characteristics that predict which epistemic mode applies, and therefore whether agent or human evaluation is appropriate.

3.1 Criterion Specifiability

If the evaluation criteria can be fully specified before evaluation begins (e.g., "does the output contain a SQL injection vulnerability?"), the task is measurement. If the criteria emerge or are refined through the act of evaluating (e.g., "is this response appropriate for a grieving user?"), the task is understanding. Agent evaluators excel at the former. The latter requires evaluators whose responses *constitute* evidence about the phenomenon being studied.

3.2 Context Dependence

If the quality of an output depends on features of the context that are difficult to fully specify in a prompt (cultural norms, emotional register, domain conventions, the user's history with the system), the evaluation requires human contextualization. Agents can be given context as input to a function. Humans inhabit context as a frame that shapes perception. This distinction matters most when context shifts what counts as a good output in ways that resist explicit parameterization.

3.3 Psychological Representativeness

If the evaluation is meant to predict or characterize how a target population will experience the output, the evaluator needs to be drawn from or representative of that population. This is standard practice in user research and survey methodology, but it has been largely abandoned in the move toward agent evaluation. An LLM judge can assess whether an output *conforms* to a specified norm. Whether a member of a particular community will *experience* it as aligned with their values is a separate question that requires a representative evaluator to answer. The distinction between stated preferences and operative values [6] is instructive: agents can evaluate against stated criteria, while the gap between stated criteria and lived experience is accessible only through human evaluation. Representativeness requirements create real cost and access tradeoffs, particularly in high-scale industrial settings where recruiting from specialized populations is expensive. The framework's value here is not in eliminating those tradeoffs but in making them visible and principled: when an evaluation substitutes convenience samples or agent judges for representative evaluators, the framework specifies what epistemic validity is being traded away. In some cases, ensuring representativeness may require broadening evaluation beyond pre-deployment testing to include structured assessment among the target population post-deployment, where the actual context of use can inform the evaluation.

3.4 Evaluative Reflexivity

Some evaluations change what is being evaluated. When users reflect on an AI output and articulate what they found helpful or harmful, they are simultaneously evaluating and constructing the criteria by which they evaluate. This reflexive process generates evaluation criteria that did not exist prior to the encounter. Agent evaluators operate on fixed criteria, even when those criteria are complex. The reflexive generation of new evaluative dimensions is one of the most valuable outputs of human-centered evaluation research [3, 10], and it is structurally inaccessible to evaluators that cannot be changed by what they encounter.

4 Implications for Hybrid System Design

These characteristics suggest concrete design principles for evaluation systems that integrate agents and humans.

Decompose evaluation pipelines into atomic steps and classify each. Many evaluation tasks contain both measurement and understanding components. Coding quality evaluation involves syntactic correctness (measurement) and readability (understanding). Safety evaluation involves policy compliance (measurement) and contextual appropriateness (understanding). Decomposing evaluation into atomic steps and classifying each allows targeted allocation. Agents can propose coding schemas, identify candidate issues, and verify factual claims; humans can assess experiential quality, cultural appropriateness, and emergent value dimensions.

Use agents for coverage, humans for calibration. Agents can evaluate every output in a dataset for measurement-mode criteria. Humans should evaluate strategically sampled subsets for understanding-mode criteria, and these human evaluations should calibrate and audit the agent evaluations rather than replicate them.

Preserve reflexivity in the evaluation loop. If evaluation is designed only around pre-specified criteria, it cannot surface the new dimensions that human evaluation often reveals. Hybrid systems should include structured opportunities

for human evaluators to generate new criteria through reflection on outputs they encounter, and these criteria should feed back into subsequent agent evaluation rounds.

Meta-evaluate agent evaluators against human evaluation on understanding tasks. Agent evaluation of measurement tasks can be validated against ground truth. Agent evaluation of understanding tasks should be validated against human evaluation, with explicit attention to where agent and human assessments diverge. Systematic divergence on understanding tasks should not be understood as a source of noise, but of critical information.

4.1 A Worked Example: Evaluating a Clinical AI Assistant

Consider evaluating a language model deployed as a clinical decision-support tool. A typical evaluation pipeline might ask: does the model provide accurate, helpful, and safe responses to clinician queries? The framework decomposes this into distinct epistemic steps.

Measurement steps (agent-appropriate): Verify that cited medical literature exists and supports the claims made. Check drug interaction warnings against a pharmacological database. Confirm that dosage recommendations fall within established clinical ranges. Assess whether the response addresses all components of the query. These steps have external ground truth and benefit from the coverage agents provide.

Understanding steps (human-necessary): Assess whether the response's confidence calibration matches what a clinician would find useful versus misleading. Evaluate whether the explanation's level of detail is appropriate for the clinical context (emergency triage versus routine follow-up). Determine whether the hedging language would lead a clinician to appropriately adjust their confidence or would instead introduce unhelpful ambiguity. These steps require evaluators who inhabit the clinical context and whose responses are representative of the target user population.

An agent can flag that a response contains hedging language (measurement). Only a clinician embedded in the relevant practice context can assess whether that hedging is calibrated appropriately for the situation (understanding). The framework's contribution is making this decomposition principled rather than ad hoc, grounding each allocation decision in the epistemic mode that applies.

5 Conclusion

AI agents belong in language model evaluation. Their ability to scale measurement and cover large output spaces is valuable and should be used aggressively. The contribution of this paper is a principled account of the limits on that use. Evaluation-as-understanding derives its value from situatedness in human psychology and context, and this property is structurally inaccessible to agent evaluators. A task allocation framework grounded in this distinction can guide the design of hybrid evaluation systems that use agents where they are epistemically sufficient and preserve human evaluation where it is epistemically necessary.

References

[1] Anthropic. 2025. Building and Evaluating Alignment Auditing Agents. (2025). Anthropic Alignment Research Blog.

[2] Anna Bavaresco, Raffaella Campagnano, Lorenzo Cavicchioli, et al. 2024. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403* (2024).

[3] A. Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: A Community-Centered Approach to Norm Elicitation for AI Alignment. *Scientific Reports* 14 (2024).

[4] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *Proceedings of the International Conference on Learning Representations*.

[5] Lee J Cronbach and Paul E Meehl. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin* 52, 4 (1955), 281–302.

[6] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[7] Dawei Li et al. 2025. Preference Leakage: A Contamination Problem in LLM-as-a-Judge. *arXiv preprint arXiv:2502.01534* (2025).

[8] Jiayi Li, Hanqi Sun, Yixuan Chen, et al. 2025. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. *arXiv preprint arXiv:2410.02736* (2025). ICLR 2025.

[9] Sijun Li et al. 2024. JudgeBench: A Benchmark for Evaluating LLM-based Judges. *arXiv preprint arXiv:2410.12784* (2024). ICLR 2025.

[10] Sasha Mitts. 2025. An Approach to Grounding AI Model Evaluations in Human-derived Criteria. In *Proceedings of the 2025 ACM Workshop on Human-AI Interaction for Augmented Reasoning (AIREASONING)*.

[11] Koki Wataoka, Tsubasa Takahashi, et al. 2024. Self-Preference Bias in LLM-as-a-Judge. In *NeurIPS 2024 Workshop on Safe Generative AI*.

[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36.