On Recursive Resolution: Scalable Ground Truth for Self-Improving AI Systems

KDD 2026
Jeju, Korea

A self-improving evaluation architecture that generates ground truth at 15.95× lower annotation cost

Disha · **Jason Whang** · **Manikandarajan Ramanathan** , Amazon, USA

1 INTRODUCTION

The Bootstrapping Paradox

Automated evaluation of compound AI systems faces a circular dependency: calibrating an LLM evaluator needs ground-truth labels, yet generating those labels at scale is exactly what automation is supposed to eliminate.

- Manual annotation does not scale: our RAG system needed **26 annotators × 10 days** to generate and evaluate 720 conversations.
- LLM-as-a-judge alone either assumes ground truth exists or accepts uncalibrated judgments, neither is production-ready.

We evaluate across **M = 13 quality metrics** on multi-turn conversations, decomposing the challenge into three sub-problems: ground truth generation, coverage (a test corpus matching real traffic), and decision (noisy failure rates → statistically rigorous pass/fail).

Three Contributions

i. **Iterative Consensus Synthesis** treats LLM–human agreement as a free signal: agreement becomes ground truth, and only disagreements escalate to a second annotator, achieving multi-annotator reliability at **15.95×** lower cost.

ii. **Tiered Wilson Interval Framework** transforms noisy binary judgments into confidence-bounded results, separating true regressions from sampling noise.

iii. **Recursive Resolution**: evaluation thresholds update based on remedy effectiveness, so the system improves its own detection criteria over time.

2 METHODS

Five-Layer Closed-Loop Architecture

L1 Evaluation (LLM-as-judge, 13 metrics) → **L2 Root-Cause Engine** (5 categories) → **L3 Remedy Engine** (targeted fix) → **L4 Rollout Manager** (shadow deploy + gates) → **L5 Continuous Learning** (feeds thresholds back to L1).

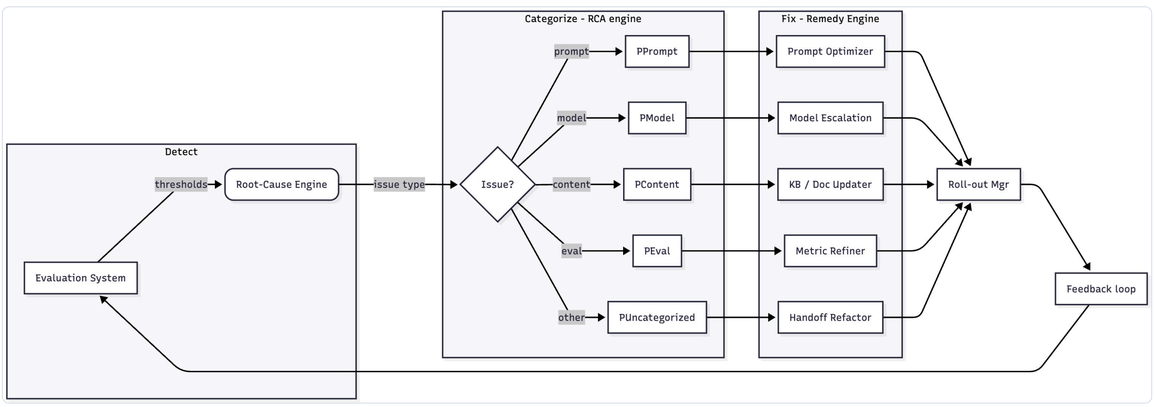

Fig 1. Detect → Categorize → Fix → Roll-out, with a feedback loop from Layer 5 back to Layer 1. When a remedy works, L5 updates the detection threshold, encoding the lesson that this failure mode is handled.

Two-Round Consensus Ground Truth

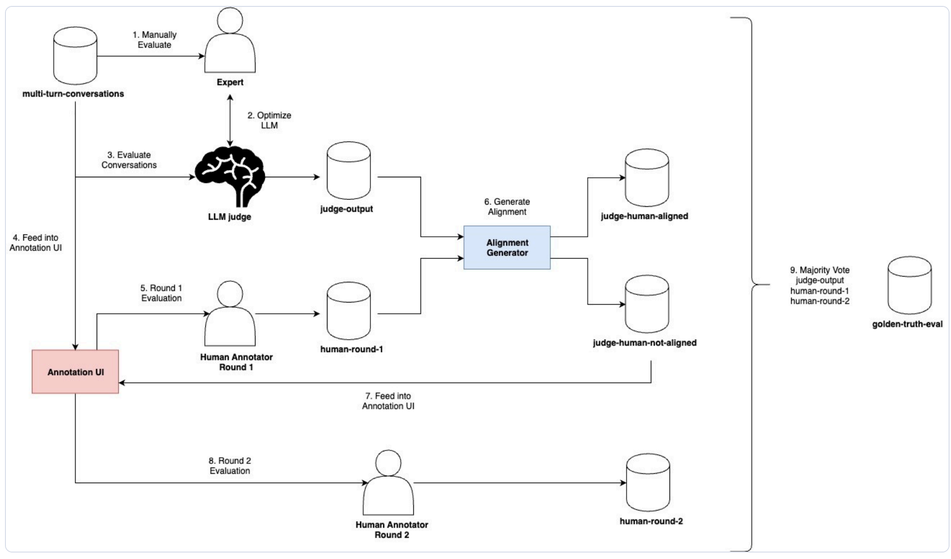

Fig 2. Full Round-1 agreement bypasses Round 2 entirely, preserving consensus quality only where it's needed, the disagreement cases.

Agreement structure gets the best of cheap-but-noisy single-annotator labels and reliable-but-expensive consensus. Only **13.2%** reached Round-1 complete agreement, so escalating just the disagreements drops expected cost to **1 + (1–p)** annotations per conversation.

$$G_i(c) = M_{1,i}(c) \text{ if judge \& annotator-1 agree (R1)} \qquad G_i(c) = \textbf{majority}(E_i, M_{1,i}, M_{2,i}) \text{ if disagreement} \rightarrow \text{R2}$$

where, for metric *i* on conversation *c*: $G_i(c)$ = final ground-truth label · E_i = LLM-judge verdict · $M_{1,i}$, $M_{2,i}$ = human annotator verdicts (Rounds 1 & 2).

Wilson Interval Decision Framework

A 5% observed failure rate on 100 cases could truly lie anywhere from 2%–10%. We use the Wilson score lower bound L_i to declare a regression only when we are 95% confident the true failure rate exceeds the established baseline θ_i ($\alpha = 0.05$, $z = 1.96$), distinguishing genuine quality drops from sampling noise.

$$L_i = \frac{\hat{p}_i + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Metrics are organized into three severity tiers, BLOCKER (w_B=10), P0 (w_{P0}=5), P1 (w_{P1}=2), and tier-level results aggregate into one decision:

- **FAIL**, any **BLOCKER** regresses, or ≥3 **P0** metrics regress
- **INVESTIGATE**, 1–2 P0 metrics regress
- **PASS**, otherwise

3 RESULTS

Efficiency Gains

15.95× — less annotation effort

307% — more test coverage

7,672→481 — annotation hrs / cycle

14,653 — turns evaluated (from 3,600)

The two-round mechanism cut per-cycle annotation from **7,672 → 481 hours** (325h R1 + 156h R2) while **increasing** coverage, enabling continuous evaluation.

Judge Reliability vs. Human Consensus

The LLM judge matches human inter-annotator agreement, strongest on objective safety metrics.

Metric (tier)	H_1	LLM	H_2
PII (BLOCKER)	100	99.9	100
Offensive (BLOCKER)	99.7	100	98.4
Knowledgeable (P0)	84.4	87.6	81.2
Relevant Resp. (P0)	86.3	84.4	84.0
Efficient (P1)	72.0	73.0	87.3

Alignment on subjective metrics (74–88%) falls within the human inter-annotator range, lower agreement reflects genuine ambiguity, not judge failure.

Statistically-Grounded Detection

We evaluated the system across **three locale-specific test suites**. The analysis below is for the third locale (Wilson, n=555, α=0.05): only regressions where the lower bound L_i exceeds baseline θ_i trigger action.

Metric	Tier	%Fail	L_i	θ_i	Status
Unknowledgeable	P0	55.1	50.9	27.2	**FAIL**
Unhelpful	P0	72.1	68.1	62.6	**FAIL**
Intent Underst.	P0	61.8	57.6	59.8	**Pass**
Non-relevant	P0	15.0	12.1	23.3	**Pass**
PII / Offensive	BLOCK	0.0	0.0	0.0	**Pass**

Intent Understanding fails raw threshold but **passes** once sampling noise is bounded, preventing a false alarm.

4 ANALYSIS

Recursive Resolution in Action, Two Case Studies

Case 1, Content Format Divergence

Judge flagged elevated **Unknowledgeable** failures. L2 traced **89%** to a locale whose help content lived in a separate collection (297 vs 431 docs).

Remedy: query both collections → coverage 297→427 docs. Shadow eval: Knowledgeable **58.3%→79.1%** (+20.8pts), p95 latency 1.2s→1.4s.

Recursion: L5 recorded the pattern and added a proactive **content-format audit**, heading off the same regression on the next locale **before** evaluation. Full test set: **42% fewer failed Q&A conversations.**

Case 2, Systematic Knowledge Gaps

Wilson framework confirmed two real P0 regressions (555 convs). L2 clustered **6 failure themes**; L3 triaged them into content gaps (fixable via KB) vs capability limits (engineering roadmap) → **Pass-with-Investigation**.

The triage converts raw evaluation signals into differentiated remediation strategies rather than a single verdict.

Six Systematic Failure Themes

Semantic clustering of the judge's reasoning surfaced six recurring themes (L3 triage in **green** / **red**):

- **Content gap**, Account access
- **Content gap**, Specific info (prices, dates)
- **Content gap**, Regional knowledge & edge cases
- **Content gap**, Temporal questions ("how long")
- **Capability**, Dynamic content / account status
- **Capability**, Context retention across turns

5 CONCLUSION

Conclusion & Key Takeaways

We presented iterative consensus synthesis, which breaks the ground-truth bootstrapping problem by treating LLM–human agreement as a free signal, cutting annotation effort **15.95×** while retaining multi-annotator reliability.

- LLM–human agreement is a free ground-truth signal, escalate only the disagreements.
- Wilson intervals separate real regressions from noise, grounding decisions statistically.
- Recursive feedback lets evaluation improve its own criteria. Scalable, self-improving evaluation of compound AI is achievable without sacrificing annotation quality.

Limitations & Future Work

Limitations. Ground-truth quality depends on annotator expertise; domain-specific training remains essential. The framework assumes metric independence, which may not hold; the third locale showed **50% overlap** between *Unhelpful* and *Unknowledgeable*. L3 remedies still require human approval; evaluation focused on English.

Future work. Move from detecting that an issue exists to which component is responsible via component-level (retriever / ranker / generator) attribution; extend to other modalities (voice / ASR error propagation); and pursue fully autonomous remediation.

6 REFERENCES

[1] Bai, Y. et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

[2] Card, D., Henderson, P. et al. 2020. With Little Power Comes Great Responsibility. *arXiv:2010.06595*.

[3] Khattab, O., Santhanam, K. et al. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv:2310.03714*.

[4] Kim, H., Hessel, J. et al. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. *arXiv:2212.10465*.

[5] Li, X., Zhang, T. et al. 2023. AlpacaEval: An Automatic Evaluator for Instruction-following Language Models. *github.com/tatsu-lab/alpaca_eval*.

[6] Liu, Y., Iter, D. et al. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv:2303.16634*.

[7] Madaan, A., Tandon, N. et al. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *NeurIPS 36*.

[8] Shah, P., Hakkani-Tür, D. et al. 2018. Bootstrapping a Neural Conversational Agent with Dialogue Self-Play, Crowdsourcing and On-Line RL. *NAACL-HLT*, 41–51.

[9] Shalev-Shwartz, S. 2012. Online Learning and Online Convex Optimization. *Found. & Trends in ML* 4, 107–194.

[10] Wang, Y., Kordi, Y. et al. 2023. Self-Instruct: Aligning LMs with Self-Generated Instructions. *ACL*, 13484–13508.

[11] Wilson, E.B. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 209–212.

[12] Zheng, L., Chiang, W.-L. et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS 36*, 46595–46623.