Pluralistic Human Evaluation as a Mission-Readiness Signal for Deployed AI

Before deploying AI into mission-critical decision-support contexts, organizations need confidence beyond single-score benchmarks. Current practice collapses heterogeneous human judgment into single guiding metric (loss function), losing the pluralistic uncertainty that distinguishes "this AI handled the case acceptably across diverse reviewers" from "this AI averaged well while badly failing one critical sub-population."

We have built infrastructure for live, multi-reviewer, multi-context evaluation of deployed AI, capturing pluralistic verdicts as a continuous data stream rather than collapsing them to a fixed dataset. The current corpus contains over 25,000 blind reviews from over 500 verified human evaluators across 58 commercial AI models and multiple domains spanning marketing content, patient communication, cultural fluency, etc.

This talk reports findings from running this evaluation infrastructure at production scale and frames a methodology for mission-readiness assessment that respects pluralistic uncertainty:

1. Human performance in pluralistic evaluation. What structural choices in reviewer selection, task formulation, and aggregation optimize signal quality. We share what we have learned about preserving diverse human judgment as a first-class signal rather than averaging it away.
2. Decision-support implications. How pluralistic production verdicts inform AI-deployment decisions. We illustrate with patterns where a model's mean pass rate appears acceptable while category-specific or reviewer-subgroup disagreement reveals deployment-blocking failure modes.
3. Framework primitives. Reviewer-disagreement-as-signal, multi-context coverage, and continuous re-evaluation rather than one-shot acceptance - primitives translatable to other mission-readiness assessment contexts.

The talk does not present a commercial product. It presents an open-source Python SDK with data access (Apache 2.0 / CC-BY-SA 4.0) that mission-

readiness practitioners can examine. We focus on what the substrate has taught us about deployed-AI confidence in subjective, high-context use cases.

We close with two questions for the audience:

1. How do current mission-readiness frameworks accommodate the structural heterogeneity of human judgment on deployed AI?
2. Where does a single confidence number break down for the use cases your team faces?

We argue the answer informs a non-trivial redesign of how deployed AI is assessed before mission insertion.