Deference-Aware Evaluation for Human-in-the-Loop AI Systems: A Unified Quality Signal for AI Operating Under Human Oversight

Yuyu Shen [1]

Abstract

Deference-Aware (DA) Evaluation is a framework for evaluating AI systems operating under human oversight. In human-in-the-loop (HITL) systems, standard evaluation metrics treat confident wrong answers and uncertain ones identically, creating systematic pressure toward overconfident AI, precisely the failure mode most dangerous when consequences are severe. DA evaluation credits appropriate deferral as correct within a single unified quality signal, with confident wrong answers as the only true failures. Because DA metrics are direct extensions of standard metrics on the same full-population denominator, the gap between them reads as a single-number diagnostic that distinguishes hidden calibration, genuine confident errors, and well-calibrated models. We instantiate the framework across five systematic-review datasets (2,729 studies) and six frontier models in evidence screening. The model that ranks first under standard recall-first evaluation is among the weakest once confident errors are counted: it reaches high recall through confident over-inclusion, with the highest confident error rate and almost no deferral. Under DA evaluation the best-calibrated model instead ranks first on every dimension (highest DA-Recall and DA-Specificity, lowest confident error and confident false-negative rates) and this ordering is stable across the full confidence-threshold range. We propose a minimum reporting standard for HITL AI with direct implications for procurement policy, regulatory frameworks, and DA-aware training. The framework extends naturally to agentic and multi-agent settings where confidence-triggered fallback is the emerging deployment pattern.

[1]Hopperlace, London, United Kingdom. Correspondence to: Yuyu Shen <yuyu@hopperlace.ai>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

A central challenge for technical AI governance is translating ethical principles and policy goals into measurable technical standards (Reuel et al., 2024). Among the open problems Reuel et al. identify, one is particularly acute for AI systems operating under human oversight: "evaluations for decision-making systems in high-stakes settings will likely demand a higher level of confidence than other applications, but it is unclear how to determine the required level of rigor based on use case." This paper proposes Deference-Aware Evaluation as a direct answer to that problem.

The governance relevance is not incidental. Evaluation frameworks are not neutral measurement tools: they are incentive structures. A procurement standard, a regulatory benchmark, or an academic leaderboard that rewards overconfident AI will systematically select for AI systems that suppress uncertainty, regardless of developers' intentions. In high-stakes human-in-the-loop (HITL) deployments — clinical decision support, legal review, safety-critical inspection, evidence synthesis — this creates exactly the conditions the human factors literature warns against: AI systems that never express uncertainty give human reviewers no signal about which cases warrant additional scrutiny, triggering automation bias (Goddard et al., 2012; Parasuraman & Manzey, 2010) and negating the efficiency benefits that motivated AI adoption.

A note on positioning. The term "human-in-the-loop" is sometimes characterised as a classic or pre-agentic pattern, potentially superseded as AI systems become more autonomous. The current evidence points the other way: industry surveys project agentic AI in the majority of enterprise deployments within a few years (Gartner, Inc., 2026), and guidance from major AI developers explicitly frames agentic AI as requiring human oversight at critical decision points, via action constraints and escalation rules (Shavit et al., 2023). HITL is evolving from an approval-checkpoint pattern into a confidence-triggered fallback pattern, precisely the architecture DA evaluation is designed to assess. More fundamentally, even in fully agentic systems, the ability to be right when confident and to defer appropriately when uncertain is architecture-independent: the handoff target may be a human, a more specialised AI, or a structured

escalation path, but the underlying evaluation question is the same. This paper's normative argument, that uncertain consequential decisions belong to humans, applies to the subset of cases where accountability and moral weight demand it; the capability argument, that well-calibrated deferral improves outcomes, applies more broadly.

Standard metrics derive from information retrieval: recall (also known as sensitivity), precision, specificity, F1. They assume a closed system in which the AI is the final decision-maker, every case receives a decision, and every wrong answer is equally bad regardless of expressed confidence. This assumption is appropriate only for fully automated systems with no recourse. Whenever AI operates as part of a larger system (with a human reviewer, a more specialised AI, a consensus mechanism, or any escalation path), overconfidence is at best unhelpful and at worst actively harmful. It is most harmful in HITL contexts specifically, where a human reviewer is present precisely to handle the cases the AI is unsure about; an AI that never expresses uncertainty wastes the reviewer's expertise on the wrong cases and triggers automation bias on the rest. Existing frameworks either force the AI to commit on every case (forced-binary evaluation) or evaluate deferral capability as a separate score disconnected from task quality (abstention benchmarks, coverage/risk curves). Neither, to our knowledge, produces a unified signal of the AI's behaviour within its role.

Koessler & Schuett (2023) review risk assessment techniques applicable to AGI companies and explicitly exclude "techniques that aim to ensure that humans perform reliably on routine tasks," noting these "should be reconsidered" if employees are tasked with overseeing dangerous AI systems. That condition already applies across many high-stakes deployments. DA evaluation contributes toward filling this gap by providing an evaluation methodology for AI systems operating under human oversight. Orchestration design, deployment integration, and human-AI team protocols are complementary and necessary future work that this paper does not address.

Contributions. This paper makes three contributions to technical AI governance:

1. **A unified evaluation framework** (DA-Recall, DA-Specificity, and directional confident error rates) that scores appropriate deferrals and confident decisions within a single quality signal measured across the full population, removing the systematic pressure toward overconfident AI that standard metrics create.

2. **A diagnostic function**: because DA metrics are direct extensions of standard metrics on the same denominator, the gap between them reads as a single-number diagnostic that distinguishes hidden calibration, genuine confident errors, and well-calibrated models, prescribing different interventions for each.

3. **A minimum reporting standard**: metric dimensions that together give procurers, regulators, and evaluators the complete picture that standard metrics alone cannot provide, with direct implications for procurement policy, regulatory frameworks, and DA-aware training.

The framework is domain-agnostic. Empirical results in this paper use systematic review evidence screening as a clean proof case, but the framework applies wherever human oversight of AI decisions is meaningful.

2. Background and Related Work

Selective prediction and learning to defer. The formal study of AI abstention dates to Chow (1957), extended by El-Yaniv & Wiener (2010) into selective prediction and by Geifman & El-Yaniv (2017) to the reject option in deep neural networks. Madras et al. (2018) introduced Learning to Defer (L2D), modelling variable deferral cost as a function of human performance on deferred cases.

Evaluation of selective classifiers. Traub et al. (2024) represent the state of the art in evaluation methodology for selective classification, proposing AUGRC, a threshold-independent benchmark metric that scores selective classifiers more accurately across operating points. DA evaluation is orthogonal to AUGRC: where threshold-independent metrics ask "across all operating points, how well does this classifier perform?", DA evaluation asks "at the single threshold where this system is deployed, is the AI behaving safely within its HITL role?" Both are needed; neither replaces the other. Research benchmarks require threshold-independent comparison to determine architectural or training advantages. Deployment assessment requires operational evaluation at the chosen working point with deferred workload made explicit.

LLM abstention and uncertainty. Wen et al. (2025) survey abstention from three perspectives (query, model, and human values), identifying it as a mechanism for mitigating hallucinations and enhancing safety. Kirichenko et al. (2025) benchmark abstention recall across 20 datasets and 20 frontier models, finding that scaling models has little effect on abstention performance and that reasoning fine-tuning degrades it. These benchmarks target queries that are in principle unanswerable. HITL classification is a structurally different problem: every case is answerable in principle; the uncertainty is about whether the AI's confidence is sufficient to act on without human verification. The existing abstention literature treats abstention capability and task accuracy as disconnected scores, never integrated into a unified view of how the AI behaves within its HITL role.

Human factors in HITL systems. Automation bias, the

tendency to over-accept automated recommendations without independent verification, is well-documented (Goddard et al., 2012; Parasuraman & Manzey, 2010). Appropriate reliance depends on trust that is calibrated to actual system reliability (Lee & See, 2004), and complementary team performance, the team outperforming either party alone, requires the human to intervene on precisely the cases the AI gets wrong (Bansal et al., 2021); a system that never expresses uncertainty removes the signal that makes such selective intervention possible. Confidence miscalibration compounds this: modern neural networks tend to be systematically overconfident (Guo et al., 2017), and Li et al. (2024) show empirically that miscalibrated AI confidence impairs appropriate human reliance and reduces decision efficacy, and that users find miscalibration difficult to detect. The connection to evaluation methodology is direct: a framework that rewards confident AI outputs creates exactly the conditions for automation bias. Recent work on LLM assistants sharpens where the leverage lies. Bo et al. (2025) benchmark *interface-level* interventions layered onto an already-committed answer — a static reliance disclaimer, token-level confidence highlighting, and indirect (hint-based) presentation — and find they reduce over-reliance but largely fail to improve *appropriate* reliance, often dampening reliance even when the advice is correct, while users grow more confident after wrong reliance decisions. The implication is that caveating a delivered answer is a weak lever; the more promising point of intervention is upstream, in whether the system commits at all — the structural deferral DA evaluation rewards, rather than a post-hoc uncertainty annotation on an answer already given. A discrete deferral class also differs in where it directs attention: rather than diffusely discounting every answer, it concentrates human scrutiny on the flagged minority, though whether this preserves appropriate reliance in practice is an empirical question we leave to future work.

AI governance and risk assessment. Koessler & Schuett (2023) review risk assessment techniques for AGI companies and explicitly defer HITL oversight evaluation as future work. The RAISE framework (Thomas et al., 2025) (co-led by Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence) calls for validated performance metrics for AI evidence synthesis tools but does not specify what those metrics should be. DA evaluation provides the specification both communities have called for.

3. The Deference-Aware Evaluation Framework

3.1. Normative Premise

The framework rests on a position: in HITL contexts, uncertain decisions on consequential matters belong to humans. This is both a practical claim (the human is present precisely to handle cases the AI cannot reliably resolve) and a normative one. Two independent justifications apply.

Instrumental: In domains where AI operates under expert oversight, domain experts may currently outperform AI on uncertain cases. This justification may weaken as AI capabilities improve.

Normative: In high-stakes domains, uncertain decisions should be made by humans regardless of relative performance, because the decision carries moral weight and accountability that should not be delegated. This applies most clearly where consequences are irreversible, fundamental rights are at stake, or human accountability is a legal requirement: defence, clinical diagnosis, judicial decisions, critical infrastructure.

DA evaluation is compatible with both justifications. In either case, the operational implication is the same: an AI that correctly identifies its own uncertainty and defers has behaved correctly. The only failure is a confident wrong answer, a case the AI should have deferred but did not, bypassing human judgment.

This also clarifies behaviour when the deferred-to authority is not superior to the AI. Under the instrumental justification, the value of deferral tracks relative performance and weakens if the reviewer is the weaker party; under the normative justification, deferral remains correct regardless of relative performance, because the decision carries accountability that should not be delegated. Identifying the regime in which the reviewer underperforms the model (and whether deferral should then be re-routed to a more capable authority rather than suppressed) is an orchestration-design question that DA evaluation surfaces but does not resolve.

3.2. UNSURE as a First-Class Output

UNSURE is not a cosmetic addition but a claim about the epistemic structure of the task. In binary classification, the AI asserts: *I have processed this case and reached a decision.* With UNSURE, the AI asserts: *I have processed this case and determined that my decision is not reliable enough to act on without human review.* These are categorically different outputs with different downstream consequences. (In broader deployments where the receiving authority may be a more specialised AI or a structured escalation path, the same logic applies with the handoff target generalised; for this paper we stay focused on the human-review case.)

UNSURE is triggered when a model's confidence falls below a threshold θ. Confidence can be operationalised through various mechanisms (token probability, verbalised confidence scores, ensemble agreement), and DA evaluation is agnostic to the specific mechanism, provided the score is calibrated well enough for a threshold to be meaningful,

a non-trivial precondition given that model confidence is often miscalibrated (Guo et al., 2017). The threshold itself is a deployment policy decision: lower values maximise automation at the cost of more confident errors; higher values minimise confident errors at the cost of higher human workload. The evaluation framework remains constant regardless of θ — what changes is the reported deference rate and effective coverage.

3.3. Formal Definitions

Let TP, FP, TN, FN denote standard classification outcomes over cases where the AI expressed a confident decision. Let U^+ denote cases where the AI output UNSURE and ground truth is positive, and U^- where ground truth is negative. Total population $N = \text{TP} + \text{FP} + \text{TN} + \text{FN} + U^+ + U^-$.

Standard metrics:

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Deference-aware metrics:

$$\text{DA-Recall} = \frac{\text{TP} + U^+}{\text{TP} + \text{FN} + U^+} \tag{4}$$

$$\text{DA-Specificity} = \frac{\text{TN} + U^-}{\text{TN} + \text{FP} + U^-} \tag{5}$$

DA-Recall extends recall by crediting appropriate deferrals on true positives. DA-Specificity analogously credits deferrals on true negatives. Precision is not extended in the same way: precision measures the reliability of confident positive decisions, and UNSURE is not a positive assertion. Standard precision therefore applies unchanged under DA evaluation, computed over the subset of confident positive decisions.

DA-Recall and DA-Specificity are symmetric across the two ground truth classes: UNSURE is credited as correct whether the ground truth is positive or negative. This symmetry follows directly from the normative premise: if uncertain consequential decisions belong to humans as a matter of principle, that principle applies regardless of which direction the uncertainty falls. Crucially, both metrics share the same denominator structure as their standard counterparts, with U^+ or U^- added as a single term. This structural property — direct extension of standard metrics on the full-population denominator — is what enables the diagnostic function in Section 3.4 and the population-level interpretation in Section 4.

Primary failure signals. Confident errors are reported in both directions, as rates across the full population:

- *Confident false positive rate* $= \text{FP}/N$
- *Confident false negative rate* $= \text{FN}/N$
- *Confident error rate* $= (\text{FP} + \text{FN})/N$

Reporting both directions is essential because which direction is safety-critical depends on the domain. In screening, diagnosis, and fraud detection, confident false negatives (missing the case that matters) carry the highest cost. In defence, criminal justice, and content moderation, both directions can carry severe cost, and the framework should surface both. The framework surfaces both as first-class measurements; the domain determines which to weight most heavily in procurement and regulation.

Workload signals.

- *Deference rate* $= (U^+ + U^-)/N$
- *Effective coverage* $= 1 -$ deference rate

The metrics are deliberately simple: direct extensions of their standard counterparts, with errors reported as rates across the full population. This is a considered design choice. Threshold-independent benchmark metrics are the right tool for comparing systems across operating points in research; DA evaluation is the right tool for assessing deployed systems at the operational threshold. Simplicity keeps the framework interpretable by practitioners, auditable by regulators, and directly actionable without requiring threshold curve analysis. Rates rather than counts keep measurements comparable across evaluations of different size.

3.4. The Diagnostic Function

DA evaluation's primary purpose is to produce a unified quality signal for AI systems that operate under oversight: an evaluation that credits appropriate deferrals as correct behaviour, penalises confident wrong answers as the only true failures, and reports deferred workload explicitly. This removes the systematic pressure toward overconfident AI that standard metrics create. For this paper we focus on HITL, where the oversight authority is a human reviewer, but the framework applies to any context where deferral or escalation to an appropriate authority is a valid response.

Beyond producing this unified signal, DA evaluation enables diagnostic reasoning about model behaviour. Because DA metrics are direct extensions of standard metrics on the same full-population denominator, the side-by-side comparison of standard and DA metrics is a single-number diagnostic: the gap distinguishes three categorically different model states.

Large gap between standard and DA metrics. The model has hidden calibration: it recognises cases where its confidence

is low, but under forced-binary evaluation must commit anyway, defaulting to one side and being counted wrong on cases it could have flagged. DA metrics reveal this calibrated uncertainty; standard metrics hide it. Intervention: allow uncertainty expression in the deployment framework, or fine-tune on low-confidence borderline cases to push the model toward more confident correct decisions where the hedge is unnecessary. The target of fine-tuning is the borderline regime, not the deferred set specifically — the latter is an artefact of threshold choice.

Small gap with low absolute DA metrics and high confident error rate. The model is confidently wrong on cases where deferral would have been appropriate. The metrics barely move because the model is not deferring much: it is committing to incorrect decisions with high confidence. Intervention: retrain on misclassified cases, redesign prompts to improve calibration, or accept the model is unsuitable for confident deployment in this domain.

Small gap with high absolute DA metrics and low confident error rate. The model is genuinely well-calibrated. It is confidently right on most cases, defers appropriately on the few it is not sure about, and rarely commits confident errors. The metrics do not move much because there is little room to improve. This is the target state, not a failure mode.

The eventual goal of a well-calibrated HITL AI system is to answer correctly on as many cases as possible while appropriately deferring the rest: high DA-Recall, high DA-Specificity, low confident error rate. DA evaluation supports progress toward this goal in a way standard metrics cannot. Even at very high performance levels, DA evaluation remains valuable: a model at 98% standard recall with 1% appropriate deferral and few confident errors is substantively different from one at 98% standard recall with no deferral and several confident errors — especially in contexts where errors matter. Only DA evaluation with its directional confident error rates makes that difference visible.

4. Why Standard Frameworks Fail

We distinguish three evaluation frameworks for AI classifiers with a possible deferral output. Frameworks 1 and 2 appear in the current literature; Framework 3 is what we propose here.

Framework 1 (Forced Binary). The AI produces only a confident positive or negative prediction. Deferral is not a valid output. Standard recall, precision, specificity, and F1 apply. Appropriate for fully automated systems, but applying it to HITL AI creates systematic pressure to suppress uncertainty expression.

Framework 2 (Coverage/Risk). The AI can abstain or defer, but deferred cases are excluded from the denominator. The system is scored only on what it chose to answer. Coverage is reported as a separate dimension alongside risk.

Framework 3 (DA Evaluation). UNSURE is scored within the primary task metrics as correct, with metrics computed on the full population. Confident wrong answers are the only failures. Deferred workload is reported explicitly alongside.

The root distinction between DA evaluation and coverage/risk is the denominator. Coverage/risk measurements are conditional on the answered subset: "of what the AI chose to answer, what fraction was wrong." DA measurements are across the full population: "of all cases in the evaluation, what fraction did the AI get confidently wrong." This denominator difference propagates into three deployment-relevant consequences.

First, population-level interpretation of critical failure. The confident error rate under DA reads directly as a fraction of deployment workload: if the rate is 1.2%, the AI produces confident wrong answers on 1.2% of the cases it encounters. The critical-direction confident error rate (for example, confident false negative rate in screening) can be read at a glance as the fraction of real workload that produced the safety-critical failure. Under coverage/risk, the same information requires mentally composing coverage with conditional risk: a system reported as "77% coverage, 1.5% risk on answered" has an actual deployment failure rate of 77% $\times$ 1.5% = 1.16% — computable, but not the default reading, and the conditional-risk number in isolation tends to bias interpretation upward. Directional decomposition is achievable under coverage/risk if reported carefully, but the conditional-vs-population denominator difference remains, and it is what determines whether the number can be read directly as a deployment-workload fraction.

Second, diagnostic comparability with standard metrics. DA-Recall shares the same denominator structure as recall: true positives plus false negatives, now extended by U^+. A direct side-by-side comparison of recall and DA-Recall is a single-number diagnostic: the gap means specifically "this is the contribution of appropriately deferred true positives that forced-binary evaluation counted as errors." Coverage/risk can approximate this diagnostic through composition (a gap between standard recall and recall-on-answered carries similar signal), but the comparison is between metrics on different denominator structures (full population vs answered subset), not the same denominator with the deferred cases added. Readers familiar with standard metrics can interpret the DA comparison directly; the coverage/risk composition requires mentally tracking two numbers (coverage and conditional risk) to reconstruct the diagnostic. In regulatory and procurement contexts where standard recall and specificity are the established vocabulary, DA-Recall and DA-Specificity can be reported alongside them without requiring a different interpretation framework. Coverage/risk

produces metrics that look and behave differently from standard metrics.

Third, normative content. DA evaluation encodes a substantive claim that coverage/risk does not: that appropriate deferral is correct behaviour, not absence of behaviour. This claim is baked into the mathematics: deferred cases count toward the numerator of DA-Recall. Coverage/risk is agnostic on this point. This agnosticism is appropriate for threshold-independent research comparison — coverage/risk's original purpose — but not for deployment evaluation of HITL AI, where the question is whether the AI is behaving well within its role, and where the framework itself functions as an incentive structure.

5. Empirical Results

5.1. Setting

We instantiate DA evaluation in systematic review evidence screening at Level 1 (title and abstract). In this task, reviewers screen thousands of studies against pre-specified inclusion criteria to identify the subset relevant for a systematic review, the evidentiary basis for clinical guidelines and policy decisions affecting millions. Each study has a verified ground-truth label: INCLUDE (positive — relevant to the review) or EXCLUDE (negative — not relevant). AI models output INCLUDE, EXCLUDE, or UNSURE, with UNSURE routing the study to human review. This domain provides a clean proof case: binary ground truth with verified labels, established HITL quality standards (Cochrane mandates independent dual-reviewer screening (Higgins et al., 2024)), and a commonly reported failure mode in current AI screening tools — high recall but chronically low specificity under forced-binary evaluation. In L1 screening, the safety-critical confident error is the confident false negative (missed include).

We evaluate six frontier models across five published systematic reviews drawn from the SYNERGY benchmark (De Bruin et al., 2023): Donners et al. (2021), together with the Meijboom (2021), Nelson (2002), Oud (2018), and Sep (2021) reviews, comprising 2,729 studies per model in total, each study carrying a verified INCLUDE/EXCLUDE ground-truth label and PICO criteria extracted from the published protocol. Models: Claude Sonnet 4.6, DeepSeek v3, Kimi k2, Llama 3.3 70B, Mistral 3.1 24B, Gemma 3 27B. All accessed via OpenRouter for consistent infrastructure. Prompts include structured PICO criteria and instruct models to output INCLUDE/EXCLUDE/UNSURE with a confidence score and brief reasoning chain. Confidence threshold $\theta = 0.85$; a full threshold sensitivity sweep is reported in Appendix A. The prompt is stage-neutral — it does not bias toward L1 recall norms — allowing results to generalise across screening stages.

All metrics are computed against human inclusion/exclusion labels as ground truth. In 2.5% of studies (67 of 2,729) all six models committed confidently against the human label; we count these as confident errors throughout, so reported confident-error rates are upper bounds. Whether such unanimous divergence signals unreliable labels or correlated model bias, and how a deployed system should adjudicate it, is left to future work.

5.2. Ranking Inversion Under Standard vs. DA Metrics

Table 1 reports standard and DA metrics across the five reviews (2,729 studies per model). Under recall-first ranking following the L1 field norm, Gemma 3 27B ranks first (54.2% recall), with Kimi k2 (51.8%) and Claude Sonnet 4.6 (51.2%) close behind; the recall spread across the top four models is narrow. Under DA evaluation the ordering separates sharply. Claude ranks first on every DA dimension: 66.7% DA-Recall, 99.3% DA-Specificity, a 3.0% confident error rate, and the lowest confident error in both directions (0.7% false positive, 2.4% false negative), achieved with a 12.9% deference rate. The recall leader, Gemma, is the weakest model under DA evaluation: a 13.4% confident error rate, 89.0% DA-Specificity, and a 0.3% deference rate.

The inversion is explained by how each model reaches its recall. Gemma's standard recall (54.2%) equals its DA-Recall (54.2%): it almost never defers (0.3%), committing confidently on nearly every case, so crediting deferral changes nothing: its recall is bought through confident over-inclusion, visible as a 10.2% confident false-positive rate (against 3.2% false negative), the highest confident error rate in the table. Claude's standard recall (51.2%) rises to 66.7% DA-Recall, a 15.5-point gap: under forced-binary scoring the borderline cases Claude flags as uncertain are charged as errors; DA evaluation credits them as appropriate deferrals. Same denominator, two readings: standard metrics hide Claude's calibration, DA metrics reveal it.

This ordering is robust to the confidence threshold. Across $\theta \in \{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, Claude holds first place under DA evaluation at every threshold (Appendix A, Table 2); the mid-tier ordering shifts with θ (particularly above 0.90, where deferral rates diverge sharply) but the identification of the best-calibrated model does not. DA evaluation functions as a diagnostic of safety profile, not a single scalar leaderboard: which mid-tier model looks better depends on whether a domain weights confident false negatives or overall confident error, and the framework reports both.

5.3. Two Failure Modes, One Standard Score

Under recall-first ranking (the L1 field norm), Claude sits third (51.2%), behind Gemma (54.2%) and Kimi (51.8%), and the data reads as a modest Claude shortfall. Under

Table 1. Full results: five systematic reviews, 2,729 studies per model, six frontier models, $\theta = 0.85$. Rows ordered by standard recall. Rec = forced-binary recall. F1 = balanced F1. DA-R, DA-S = DA-Recall, DA-Specificity. Err = confident error rate. CFP, CFN = confident false positive and false negative rate (CFN safety-critical in L1). Def = deference rate. All values in percent; Err may differ from CFP+CFN by 0.1 due to rounding. Per-item outputs in Appendix A.

Model	Rec	F1	DA-R	DA-S	Err	CFP	CFN	Def
Gemma 27B	54.2	36.2	54.2	89.0	13.4	10.2	3.2	0.3
Kimi k2	51.8	51.8	52.1	96.4	6.7	3.4	3.4	2.1
Claude S. 4.6	51.2	**55.6**	**66.7**	**99.3**	**3.0**	**0.7**	**2.4**	12.9
Llama 70B	50.5	40.1	53.7	92.6	10.2	6.9	3.3	6.2
Mistral 24B	36.8	36.8	37.5	95.3	8.8	4.4	4.4	1.2
DeepSeek v3	28.4	35.5	30.7	97.7	7.0	2.2	4.9	5.0

balanced F1 (which weights precision and recall equally) the picture shifts: Claude leads at 55.6%, with the next-best model (Kimi) at 51.8% and the recall leader Gemma near the bottom at 36.2%. F1 surfaces what recall-first ranking hides: Gemma's precision is low (27.2%, driven by high confident false positive rates), which drags down any composite measure that weights precision.

But F1 alone does not distinguish *why* Claude's recall is lower than Gemma's. It could be hidden calibration (Claude recognises borderline cases and defaults to the safer side under forced-binary pressure) or genuine confident errors (Claude is wrongly excluding true includes with high confidence). Standard metrics treat both the same. DA evaluation separates them, and the separation is legible precisely because DA-Recall sits on the same denominator as recall:

- *Claude:* standard recall 51.2% → DA-Recall 66.7%. The 15.5 point jump reveals hidden calibration: Claude recognises which cases it cannot reliably resolve, and under forced-binary those cases become false negatives. Intervention: allow uncertainty expression in deployment, or fine-tune on low-confidence borderline cases to reduce unnecessary hedging where the model could safely commit.
- *Gemma, Llama:* standard recall ≈ DA-Recall (Gemma 54.2% = 54.2%, deference 0.3%; Llama 50.5% → 53.7%, deference 6.2%). The metrics barely move because the models commit confidently rather than defer, and the cost shows up as the two highest confident error rates in the table (13.4%, 10.2%). Intervention: retrain on misclassified cases, redesign prompts to improve calibration, or accept the model is unsuitable for confident deployment in this domain.

Standard F1 tells you these models underperform. DA evaluation tells you *why* and *what to do about it*, readable in a single side-by-side comparison with standard metrics.

5.4. Cost-Safety Trade-off Across the Confidence Threshold

The confidence threshold θ is a deployment policy lever, and DA evaluation makes its cost explicit. Raising θ routes more borderline cases to UNSURE, lowering confident errors at the price of higher reviewer workload. For Claude, the confident false-negative rate, the safety-critical error in L1 screening, falls from 2.4% at $\theta = 0.85$ to 0.3% at $\theta = 0.95$, as deference rises from 12.9% to 40.3% (Appendix A, Fig. 1, Fig. 2). Near-zero false-negative performance is thus achievable at scale, but as a threshold choice with a quantified workload cost rather than a free property of any single model.

Models differ in the efficiency of this trade. At $\theta = 0.95$, Llama 3.3 70B reaches the same 0.3% confident false-negative rate as Claude, but does so by deferring 51.6% of cases against Claude's 40.3%, a higher workload for the same safety floor. DA evaluation makes this frontier legible: the deployment question is not which model posts the lowest error at a single threshold, but which models meet the domain's confident-error requirement (the safety-critical false-negative rate above all) and, among those, which does so at the lowest deferred workload.

6. Policy Implications

6.1. Minimum Reporting Standard

We propose that HITL AI systems in high-stakes domains report all of the following dimensions:

1. Standard metrics (recall, precision, specificity, F1) for comparability with existing literature.
2. DA-Recall and DA-Specificity.
3. Confident error rate, reported with directional breakdown (confident false positive rate and confident false negative rate).
4. Deference rate and effective coverage.

This reporting standard gives evaluators, procurers, and regulators a complete picture that standard metrics alone cannot provide. It is compatible with existing reporting and oversight frameworks (including RAISE 2 for evidence synthesis AI (Thomas et al., 2025), the human-oversight requirements for high-risk systems set out in Article 14 of the EU AI Act (European Parliament and Council of the European Union, 2024), which explicitly call for overseers to interpret system output and avoid over-reliance, and the risk-management guidance of the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023)), and directly addresses a technical AI governance open problem identified by Reuel et al. (2024).

6.2. Procurement Policy

Organisations deploying HITL AI should require vendors to report all four dimensions. A procurement process selecting on recall alone will systematically favour overconfident models. Adding directional confident error rates surfaces the safety profile that recall hides, and the domain determines which direction to weight most heavily.

This has direct regulatory implications. In domains subject to audit (clinical AI, legal document review, financial compliance), the directional confident error rates provide the unit of failure that regulators can act on: not aggregate accuracy, but specific wrong answers made with high confidence, bypassing human review. Because these rates are measured on the full population, they read directly as "fraction of deployment workload that produced this failure," without requiring composition with coverage. Regulatory frameworks that require these rates can target the failure mode that matters most for accountability and liability in each domain.

6.3. DA-Aware Training

The reporting standard creates incentives for a different training objective: optimising models to achieve high DA-Recall and DA-Specificity with low directional confident error rates, rather than forcing trade-offs between recall and precision on cases where the model is genuinely uncertain. This means training models to express calibrated uncertainty on borderline cases rather than suppressing uncertainty to maximise one standard metric at the expense of another.

The human factors literature establishes why this matters beyond the numbers. AI systems that rarely express uncertainty give human reviewers no signal about which cases warrant additional scrutiny, creating the conditions for automation bias (Goddard et al., 2012; Parasuraman & Manzey, 2010). A procurement standard or regulatory benchmark that rewards calibrated uncertainty expression creates upstream pressure on training objectives, which is the prerequisite for appropriate human-AI team outcomes.

6.4. Broader Domain Applicability

The framework applies most directly where one condition holds as primary: a meaningful cost asymmetry between confident errors and appropriate deferrals. This is the defining condition: without it, there is no substantive reason to distinguish confident wrong answers from appropriate uncertainty, and standard metrics suffice. Where this asymmetry holds, two further conditions enable application.

A means of verifying correctness for evaluation purposes. Some form of ground truth, expert panel review, or retrospective verification that allows DA metrics to be computed at evaluation time.

An authority capable of handling deferred cases. Typically a human reviewer in the scenarios discussed in this paper, though more specialised AI, structured ensembles, or escalation paths can serve the same role in broader deployments (discussed further below). The normative and accountability arguments apply most clearly to human authority.

Five example domains satisfy these conditions with clear human authority:

Clinical decision support: Diagnostic AI operates under clinician oversight. A system that flags uncertain cases for additional review produces better human-AI team outcomes than one forced to commit confidently.

Legal and justice: AI tools for document review, case triage, and risk assessment operate under oversight of legal professionals who bear ethical and professional responsibility for outcomes. Flagging uncertain cases for attorney review, rather than committing confidently without an uncertainty signal, produces more defensible outcomes in adversarial proceedings and preserves the human judgment that legal accountability requires.

Safety-critical engineering: Automated inspection in aerospace, nuclear, and civil infrastructure, where a confident "safe" assessment on an uncertain component carries physical consequences.

Defence and security: Threat classification systems where both directions of confident error carry severe consequences: missing an actual threat or targeting a civilian. DA evaluation reports both error directions alongside the deference rate, making the trade-off legible and supporting human review on uncertain cases where committing confidently to either side carries moral weight.

Financial audit and compliance: Transaction monitoring systems where the cost of confident errors (missed fraud, false regulatory flags) substantially exceeds the cost of routing ambiguous cases to human reviewers.

This domain-agnosticism is deliberate and operates at the level of the metric, not the deployment policy. The quantities a domain must set, namely the confidence threshold θ (encoding risk tolerance), which directional confident error to weight most heavily, and the escalation pathway and accountable authority for deferred cases, are all domain-specific, and DA evaluation is designed to take them as inputs rather than abstract them away. Regulatory guidance reflects the same separation: the EU AI Act requires oversight measures commensurate with the risks, level of autonomy, and context of use of the system (European Parliament and Council of the European Union, 2024). DA evaluation supplies a common measurement vocabulary across domains; the weighting of that vocabulary remains a domain decision.

The principle underlying DA evaluation, crediting appropriate routing to the more suitable authority and penalising confident wrong answers, extends beyond human-AI settings. In multi-agent systems, a generalist AI that defers to a more specialised AI for domain-specific decisions is also behaving correctly under the DA principle. The accountability argument does not fully transfer to AI-to-AI deference, but the capability argument does. Extending DA evaluation to these broader settings — including the multi-agent architectures that increasingly define enterprise deployment — is natural future work.

7. Limitations

Empirical results use five verified reviews (2,729 studies) within a single domain (clinical/biomedical evidence screening). Whether model rankings, deference rates, and DA metric patterns generalise across further domains and languages remains an open empirical question. The confidence threshold ($\theta = 0.85$) was selected based on pilot analysis; Appendix A reports a sensitivity sweep across $\theta \in \{0.70, \ldots, 0.95\}$ confirming that the top-ranked model under DA evaluation is stable, though mid-tier rankings shift at high θ. Evaluation is scoped to Level 1 (title and abstract) screening; Level 2 full-text screening has not been evaluated, though the neutral prompt design means the framework extends naturally. DA evaluation as defined here assumes binary ground truth, a discrete UNSURE output, and a single authority for deferred cases. Extensions to a continuous deferral signal (e.g. routing priority proportional to distance below θ), soft labels, partial relevance, non-classification tasks (advice-giving, open-ended recommendations, conversational AI), multi-class, multi-authority, and AI-to-AI deference settings are natural directions but require additional formalisation. Finally, DA credits appropriate deferral as correct but does not model post-handoff outcomes; auditing reviewer performance on deferred cases and orchestration design are complementary future work.

8. Conclusion

Deference-Aware Evaluation is a framework for evaluating AI systems operating under human oversight that treats appropriate deferral as correct behaviour rather than as error or absence of behaviour. This produces a unified quality signal for HITL AI, with confident correct decisions, appropriate deferrals, and confident wrong answers accounted for in one coherent view, removing the systematic pressure toward overconfident AI that standard metrics create. The principle generalises beyond HITL to any system where appropriate deferral is a valid response. Beyond producing this signal, DA evaluation enables diagnostic reasoning: because DA metrics are direct extensions of standard metrics on the same full-population denominator, the side-by-side comparison of standard and DA metrics is a single-number diagnostic that distinguishes hidden calibration from genuine confident errors from well-calibrated models, prescribing different interventions for each.

The governance contribution is threefold. First, DA evaluation provides a concrete answer to a named open problem in technical AI governance (Reuel et al., 2024): it specifies the level of evaluation rigour required for high-stakes HITL deployment. Second, it fills the gap that organisational risk assessment frameworks (Koessler & Schuett, 2023) have explicitly deferred. Third, it provides a minimum reporting standard that can be adopted directly into procurement policy, regulatory benchmarks, and domain-specific guidance (including RAISE for evidence synthesis AI (Thomas et al., 2025)).

The framework is domain-agnostic. Future work includes extension to Level 2 full-text screening, cross-domain validation across clinical and legal domains, non-classification tasks, and multi-authority settings including AI-to-AI handoff. We invite adoption and extension of the framework as a minimum standard for HITL AI evaluation.

Impact Statement

This paper proposes an evaluation framework for AI systems operating under human oversight, motivated by the observation that standard evaluation metrics create systematic pressure toward overconfident AI in high-stakes deployments. The intended broader impact is positive: procurement standards, regulatory benchmarks, and training objectives that adopt DA evaluation should produce AI systems that express calibrated uncertainty, enabling more effective human oversight and reducing the risk of confident wrong decisions bypassing human review. The framework explicitly does not model post-handoff outcomes, and organisations deploying HITL AI should separately audit human reviewer performance on deferred cases. No new datasets were collected; empirical evaluation uses five published systematic review benchmarks.

LLM Use Disclosure

Large language models were used as follows. *Writing, editing, and formatting:* an LLM assisted with drafting, copy-editing, and LaTeX formatting of the manuscript. *Coding:* the evaluation harness used to query models, compute metrics, and run the threshold sweep was designed and directed by the author, with an LLM assisting in implementation; all code was reviewed by the author. *Research discussion:* an LLM was used as a sounding board. All research design, experimental setup, empirical results, the theoretical framework, and the normative arguments are the author's own, and the author takes full responsibility for the content.

The systems under evaluation (Section 5) are themselves LLMs; that experimental use is distinct from the authoring assistance described here.

References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 2021. doi: 10.1145/3411764.3445717.

Bo, J. Y., Wan, S., and Anderson, A. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, 2025. doi: 10.1145/3706598.3714097.

Chow, C.-K. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957.

De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., and Van de Schoot, R. SYNERGY — open machine learning dataset on study selection in systematic reviews. DataverseNL, 2023.

Donners, A. A. M. T., Rademaker, C. M. A., Bevers, L. A. H., et al. Pharmacokinetics and associated efficacy of emicizumab in humans: A systematic review. *Clinical Pharmacokinetics*, 60(11):1395–1406, 2021.

El-Yaniv, R. and Wiener, Y. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 (artificial intelligence act), article 14: Human oversight. Official Journal of the European Union, 2024. https://artificialintelligenceact.eu/article/14/.

Gartner, Inc. 2026 Gartner CIO and technology executive survey. Gartner, Inc., 2026. Reports that over 60% of organizations expect to deploy AI agents within two years (17% to date).

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Goddard, K., Roudsari, A., and Wyatt, J. C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (eds.). *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, version 6.5 edition, 2024. Version 6.5 (updated August 2024). Available from www.training.cochrane.org/handbook.

Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. arXiv:2506.09038.

Koessler, L. and Schuett, J. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*, 2023.

Lee, J. D. and See, K. A. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392.

Li, J., Yang, Y., Zhang, R., Liao, Q. V., Song, T., Xu, Z., and Lee, Y.-C. Understanding the effects of miscalibrated AI confidence on user trust, reliance, and decision efficacy. *arXiv preprint arXiv:2402.07632*, 2024.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 2018.

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology (NIST), 2023.

Parasuraman, R. and Manzey, D. H. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010.

Reuel, A., Bucknall, B., et al. Open problems in technical AI governance. *arXiv preprint arXiv:2407.14981*, 2024.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. Practices for governing agentic AI systems. White paper, OpenAI, December 2023. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Thomas, J., Flemyng, E., Noel-Storr, A., Moy, W., Marshall, I. J., Hajji, R., et al. Responsible use of AI in evidence SynthEsis (RAISE): recommendations for practice and guidance on building and evaluating AI evidence synthesis tools. Open Science Framework, 2025. https://doi.org/10.17605/OSF.IO/FWAUD.

Traub, J., Bungert, T. J., Lüth, C. T., Baumgartner, M., Maier-Hein, K. H., Maier-Hein, L., and Jäger, P. F. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.01032.

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

Appendix

A. Threshold Sensitivity

The confidence threshold θ is applied in post-processing to fixed model outputs; no model was re-queried, so the sweep isolates the threshold effect from sampling variance. Forced-binary recall, precision, and specificity are θ-independent by construction and were verified constant across the grid. The full run comprises six models over five reviews, 2,729 studies per model.

Table 2. DA-evaluation rank by confidence threshold (sort key: confident FN rate ascending, then confident error rate, then DA-Recall, then DA-Specificity). Claude holds first place at every threshold; mid-tier ordering is θ-sensitive, particularly above 0.90.

Model	.70	.75	.80	.85	.90	.95
Claude S. 4.6	1	1	1	1	1	1
Gemma 27B	2	2	2	2	2	3
Kimi k2	3	3	3	4	3	5
Llama 70B	4	4	4	3	4	2
Mistral 24B	5	5	5	5	6	6
DeepSeek v3	6	6	6	6	5	4

Table 3. Confident false-negative rate (%) by confidence threshold (safety-critical in L1 screening).

Model	.70	.75	.80	.85	.90	.95
Claude S. 4.6	2.97	2.64	2.53	2.35	1.54	0.26
DeepSeek v3	4.87	4.87	4.87	4.87	3.63	2.64
Kimi k2	3.37	3.37	3.37	3.37	3.19	3.11
Llama 70B	3.37	3.37	3.37	3.26	3.26	0.26
Mistral 24B	4.40	4.40	4.40	4.40	4.40	4.03
Gemma 27B	3.22	3.22	3.22	3.22	3.15	1.94

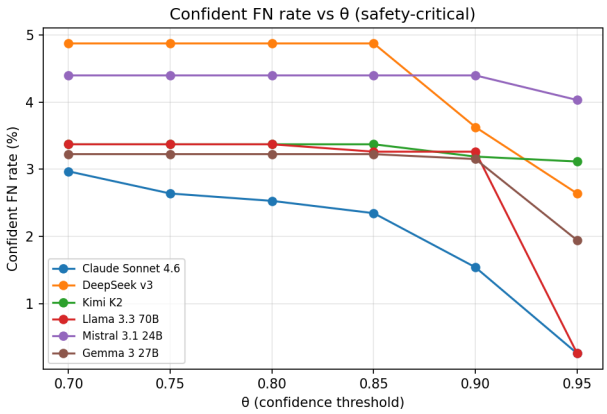

Figure 1. Confident false-negative rate vs. θ. Claude descends smoothly toward zero; over-committing models stay near-flat until very high θ.

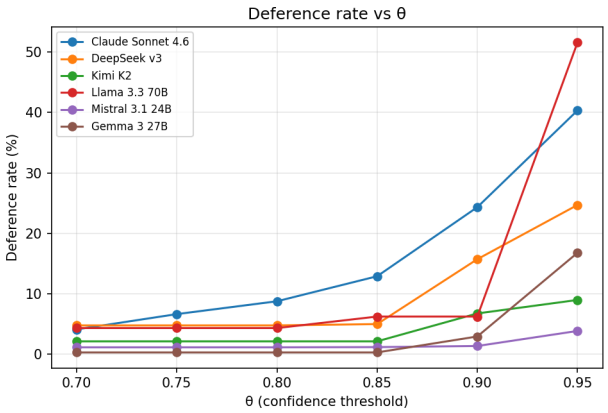

Figure 2. Deference rate vs. θ: the workload cost of the false-negative reductions in Fig. 1. At $\theta = 0.95$ Llama matches Claude's false-negative floor only by deferring more than half of all cases.