Layer-Appropriate Augmentation for Human-AI Risk Assessment: Integrating Case-Based Reasoning with Fuzzy Calibration

Gareth McCONOMY[a,1], Raymond BOND[a,b], David GLASS[a], Jun LIU[a] and Holly TONER[b]
[a] *School of Computing, Ulster University, Belfast, Northern Ireland, UK*
[b] *Kainos plc, Belfast, Northern Ireland, UK*
ORCiD ID: Gareth McConomy https://orcid.org/ 0009-0002-0931-7664

Abstract. AI risk assessment demands both structured reasoning and calibrated confidence, capabilities that frontier large language models (LLMs) possess imperfectly. Prior work has independently shown that case-based reasoning (CBR) scaffolding can improve LLM risk identification and that ordered weighted averaging (OWA) aggregation can correct systematic miscalibration. However, these interventions operate at different layers of the assessment pipeline and have not been tested in combination. We present the first integrated evaluation, combining CBR structured reasoning (reasoning layer) with OWA fuzzy calibration (output layer) across three frontier LLMs assessing 160 case-pattern pairs in ten real estate AI deployment scenarios. CBR scaffolding significantly improved assessment accuracy for one of three models (GPT-4o mean absolute error reduction of 0.062, $p<0.001$), while Grok 4 and Gemini 2.5 Pro showed no significant benefit, demonstrating that scaffolding effectiveness is model-dependent. OWA calibration functioned as a tuneable governance alignment tool whose effectiveness depends on calibrating the orness parameter to the specific expert and assessment context. The combined CBR and OWA pipeline outperformed calibrated holistic assessment (MAE improvement 0.034, $p<0.001$), demonstrating that structured reasoning and output calibration provide complementary benefits. We frame these findings as design principles for hybrid human-AI risk assessment: structure the reasoning, calibrate to the expert, validate per model, and retain human authority over governance parameters.

Keywords. human-AI collaboration, AI risk assessment, case-based reasoning, fuzzy aggregation, OWA operators, calibration, trustworthy AI

1. Introduction

Organisations deploying artificial intelligence systems face a growing governance challenge. Regulatory frameworks including the EU AI Act [1] and the NIST AI Risk Management Framework [2] now require structured risk assessment before deployment,

[1] Corresponding Author: Gareth McConomy, School of Computing, Ulster University, Belfast BT15 1ED, Northern Ireland, UK; e-mail: mcconomy-g@ulster.ac.uk.

yet the domain expertise needed to conduct these assessments is scarce relative to the volume of systems being deployed.

Large language models offer a potential path through this bottleneck. They can reason about deployment contexts and produce assessments at scale. However, two systematic weaknesses undermine their reliability: shallow reasoning that identifies surface-level risk factors without tracing causal mechanisms, and poorly calibrated confidence producing scores that systematically diverge from expert judgment, with the direction and magnitude varying across models and contexts [3, 4].

This paper is the fourth in a research programme investigating how structured knowledge and calibration methods can make LLM-assisted risk assessment trustworthy. Earlier work constructed a knowledge graph of 87 causal AI risk patterns and showed it improved LLM causal reasoning about deployment risks [5]; investigated calibration directly, finding that OWA fuzzy aggregation corrects bias at the output layer through a governance-interpretable parameter while fuzzy decomposition at the reasoning layer degrades performance, establishing that different pipeline layers require different interventions [6]; and tested CBR scaffolding as an alternative reasoning-layer intervention, with significant improvements in risk identification for some models but not others [7]. The current study integrates the CBR and OWA strands for the first time. All four studies share the same expert benchmark, knowledge graph, and deployment cases, enabling direct comparison.

Across 2,880 experimental trials we find that CBR scaffolding significantly improves assessment accuracy for one of three models, that the benefit is model- and task-dependent, that the direction of calibration bias depends on assessment composition (explaining an apparent reversal from the OWA study), and that the combined pipeline outperforms calibrated holistic assessment. We frame these as design principles for hybrid human-AI risk assessment in which the domain expert remains the essential calibration anchor.

The framework is explicitly hybrid: LLMs provide scalable pattern recognition and causal reasoning, while the human expert provides the calibration anchor. The causal risk patterns from the knowledge graph [5] provide the shared ontology that makes this collaboration possible: because both expert and LLM work from the same explicit causal structures, the expert can meaningfully evaluate the AI's reasoning rather than face an opaque score. The integration is complementary, which the experimental evaluation demonstrates.

2. Background

2.1. The AI risk assessment challenge

AI risk assessment requires identifying how deployment context interacts with known failure modes to produce harm. The assessor must determine whether the enabling conditions for a risk pattern are present, trace how the causal mechanism would operate, and judge how contextual factors amplify or mitigate the risk. This demands both broad knowledge of AI failure modes and deep understanding of the deployment domain. LLMs possess the former in abundance but lack reliable calibration: they produce risk scores that systematically diverge from expert judgment, with the direction and magnitude varying across models and contexts [3, 4].

Calibration matters especially for governance: when risk scores inform deployment decisions or audit documentation, systematic bias creates real harm, scores too low let risky systems pass review, scores too high impose unnecessary compliance costs and delay beneficial deployments [8]. Practitioners need calibrated assessments reflecting a specific organisational risk posture, not raw model outputs. This motivates the two-layer approach: structure the reasoning to improve accuracy, then calibrate the output to align with expert judgment.

2.2. Case-based reasoning for structured assessment

Case-based reasoning solves new problems by adapting solutions from similar past cases [9]. The classical CBR cycle comprises four phases: RETRIEVE similar cases, REUSE their solutions, REVISE the adapted solution, and RETAIN the result. When applied as scaffolding for LLM reasoning, CBR forces the model to decompose its assessment into explicit, auditable phases rather than producing a single holistic judgment. In the preceding study in this programme [7], CBR scaffolding improved risk pattern identification for some models but not others on a ranking task, establishing model-dependency as a key moderating variable. This study extends that investigation from ranking to quantitative applicability scoring.

2.3. OWA aggregation for calibration

Ordered weighted averaging, introduced by Yager [10], provides a parametric family of aggregation operators between the minimum and maximum functions. Given values to aggregate, OWA sorts them in descending order and applies a weight vector determined by a single parameter called orness, ranging from 0 (pessimistic, taking the minimum) to 1 (optimistic, taking the maximum), with 0.5 producing the arithmetic mean. When applied to LLM risk assessment, OWA aggregates multiple scores into a calibrated estimate where orness has a natural governance interpretation: lower values reflect a more precautionary posture. In the preceding calibration study [6], OWA with orness 0.6 reduced MAE by 7.7% compared to the best individual model when models exhibited overconfidence. That study also tested fuzzy decomposition at the reasoning layer and found it degraded performance by 34.8%, establishing the layer asymmetry that motivates the current integration: CBR (not fuzzy methods) at the reasoning layer, OWA at the output layer.

2.4. Research questions

Three questions motivate this study. First, does CBR scaffolding improve LLM risk applicability scoring, extending previous findings from pattern ranking to quantitative assessment (RQ1)? Second, how does OWA calibration interact with CBR structured outputs (RQ2)? Third, what design principles emerge for combining reasoning and calibration layers (RQ3)?

3. Method

3.1. The AI risk knowledge graph

Central to this study is a knowledge graph of 87 causal AI risk patterns, synthesised through literature review of academic research documenting AI failures, regulatory guidance, and technical reports across healthcare, criminal justice, financial services, and real estate domains [5]. Unlike taxonomic risk registers that categorise risks by label (for example, bias, or privacy), the knowledge graph encodes each risk as a causal structure. This distinction matters because risk assessment is fundamentally about mechanism, not categorisation. Knowing that bias is a risk category tells an assessor very little about whether or how it would manifest in a specific deployment. Knowing the causal pathway through which proxy variables encode protected characteristics, and the contextual conditions that enable that pathway, is actionable. Existing risk taxonomies such as those in the NIST AI RMF [2] and the EU AI Act [1] identify categories of concern but leave the causal reasoning to the assessor. The knowledge graph fills this gap by encoding the reasoning structure itself.

Each pattern comprises four components (Figure 1). Triggers are observable indicators that suggest a pattern may be relevant. Enabling conditions are the contextual prerequisites that must be present for the risk to materialise. The causal chain specifies the mechanism through which harm unfolds as an ordered sequence from initial trigger through intermediate effects to realised harm. Harms describe the specific adverse outcomes including their type and severity.

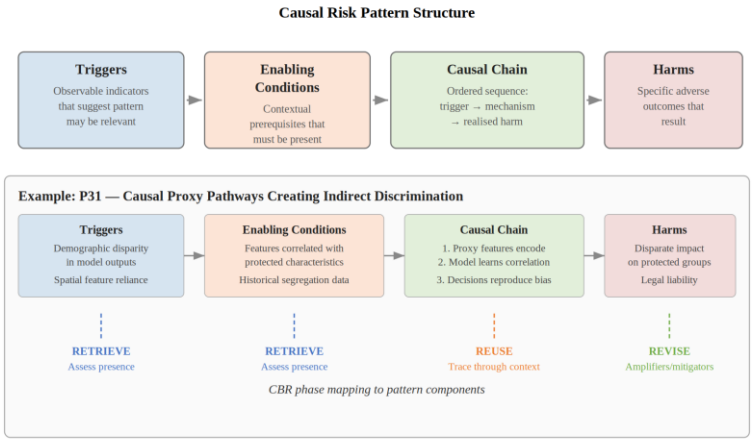


Figure 1. Causal risk pattern structure with example (P31: proxy discrimination) and mapping to CBR phases. Each pattern component maps directly to a CBR phase, enabling structured assessment of the mechanism of harm.

This causal structure is what makes the knowledge graph suitable for CBR scaffolding. The mapping is direct: RETRIEVE assesses each enabling condition against the deployment case, REUSE traces the causal chain step by step through the specific context, and REVISE considers whether contextual factors amplify or mitigate the mechanism (Figure 1). For Pattern P31 (causal proxy pathways) against a tenant screening system, RETRIEVE checks whether features correlated with protected characteristics are present, REUSE traces how neighbourhood deprivation scores encode

historical segregation and how screening decisions reproduce exclusion, and REVISE considers whether fairness constraints mitigate or market concentration amplifies the mechanism. A holistic assessment might flag discrimination as a concern without engaging with the specific pathway.

3.2. Deployment cases and risk patterns

We constructed ten realistic AI deployment scenarios in the real estate domain, each presenting a decision maker facing genuine uncertainty about whether to continue, modify, or discontinue an AI system. The cases span algorithmic rent pricing, tenant screening, automated property valuation, predictive maintenance, AI lease analysis, contract negotiation optimisation, insurance underwriting, development site selection, property recommendations, and building permit review. Each was approximately 1,000 words including system description, performance data, stakeholder perspectives, and decision context.

From the 87 patterns in the knowledge graph, 16 were selected by the domain expert as most applicable across the ten cases (Table 1), spanning fairness, governance, operational risk, market dynamics, oversight, and systemic vulnerability. This breadth ensures the experiment tests CBR scaffolding across qualitatively different causal mechanisms rather than a narrow subset.

Table 1. The 16 causal risk patterns used in this study. Each encodes triggers, enabling conditions, a causal chain, and harms as described in Section 3.1.

ID	Pattern name	Risk category
P2	Failed human-automation handoffs in safety-critical systems	Operational
P3	Automation-induced skill degradation under reduced practice	Workforce
P5	Emergent user behaviour and data drift causing production failures	Technical
P6	Market concentration leading to correlated error and systemic failure	Systemic
P11	Jurisdictional arbitrage enabling differential deployment and risk export	Governance
P12	Compliance theatre: safety claims diverging from operational practice	Governance
P23	Absent binding insurance preconditions for AI governance controls	Governance
P24	Tacit algorithmic collusion through shared optimisation	Market
P31	Causal proxy pathways creating indirect discrimination	Fairness
P36	Dynamic feedback loops amplifying allocation disparities	Fairness
P39	Process-based monitoring detecting deception with adaptation limits	Oversight
P45	Procurement validation gap	Governance
P64	Monoculture vulnerability and cross-system attack propagation	Security
P70	Proxy-driven disparate impact in high-stakes decisions	Fairness
P76	Endogenous feedback amplification in deployed decision systems	Systemic
P86	Fluency-induced overreliance on incorrect AI outputs	Human factors

3.3. Expert benchmark

Risk applicability is a matter of expert judgment: there is no objective ground truth against which an LLM's assessment can be checked. Following established practice in expert elicitation, where a single trained expert is preferable to a pooled aggregate that washes out professional reasoning, we use one well-characterised domain expert as a calibration target rather than a proxy for truth. The contribution is the layered method, structure the reasoning, calibrate, validate per model, retain human authority, not the specific numeric scores produced against this target.

The expert (over twenty years of real estate advisory experience, currently a PhD student in human-centric AI) assessed each pattern's applicability to each case on a continuous 0.0–1.0 scale, producing 160 benchmark scores spanning the full range: 34 low (<0.35), 59 moderate (0.35–0.64), 67 high (≥0.65). The framework is designed to be recalibrated to any expert through the OWA orness parameter.

3.4. Models and experimental conditions

We tested three frontier LLMs: Grok 4 (xAI), Gemini 2.5 Pro (Google), and GPT-4o (OpenAI). Claude (Anthropic) was excluded because it was used in developing the experimental infrastructure, which would compromise independence.

Table 2. Experimental conditions. C2 and C4 are computed from C1 and C3 respectively.

Condition	**Reasoning layer**	**Output layer**	**Source**
C1: Vanilla	Holistic	Raw per-model	New runs
C2: Vanilla + OWA	Holistic	OWA (α=0.6)	From C1
C3: CBR	Retrieve-Reuse-Revise	Raw per-model	New runs
C4: CBR + OWA	Retrieve-Reuse-Revise	OWA (α=0.6)	From C3

The experiment comprised four conditions in a two-by-two factorial design crossing reasoning structure with output calibration (Table 2). C1 and C3 each required 1,440 API calls (3 models × 160 pairs × 3 replications), totalling 2,880. C2 and C4 were computed from raw scores using OWA with exponential weights at orness α=0.6, the value identified by [6] as minimum-MAE in a 0.1–0.9 sweep against the same expert benchmark (a "Goldilocks zone" between over-precaution at α≤0.3 and under-correction at α≥0.7). We retain α=0.6 here to enable direct comparison with [6] and to test whether an orness calibrated against one pattern composition transfers to a different one; §4.3 examines what happens when it does not. All calls used temperature 0.3. All prompts, the 16 risk patterns, the ten cases, the expert benchmark scores, and the analysis code are available at https://github.com/ggmcconomy/hhai2026-cbr-owa.

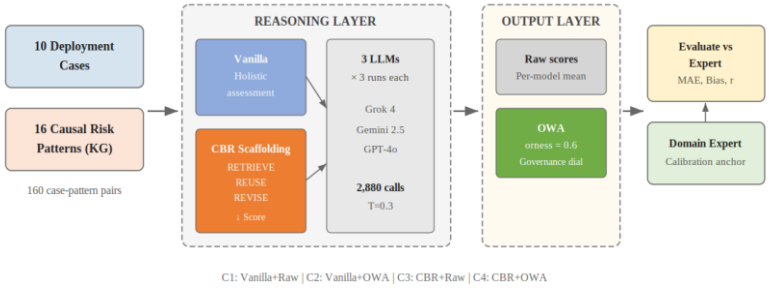


Figure 2. Experimental pipeline. Deployment cases and causal risk patterns from the knowledge graph are combined into 160 case-pattern pairs, assessed by three LLMs under two reasoning conditions (vanilla and CBR), with output calibration via OWA. The domain expert provides the benchmark and sets the governance parameter.

3.5. Prompt design

The vanilla prompt instructed the model to holistically assess the applicability of a risk pattern to a deployment case, considering enabling conditions, causal mechanisms, and contextual factors, returning a JSON object with reasoning and an applicability score between 0.0 and 1.0. The CBR prompt replaced this with three explicit phases: RETRIEVE (assess each enabling condition as present, absent, or uncertain with evidence), REUSE (trace the causal chain through the deployment context), and REVISE (identify amplifying and mitigating factors). Both prompts included identical calibration guidance instructing models to use the full score range.

Figure 3 illustrates the qualitative difference. Both prompts receive the same pattern documentation; the vanilla prompt allows holistic engagement (typically a single narrative paragraph), while CBR requires the model to address each component in turn, each enabling condition assessed as present, absent, or uncertain with case evidence; the causal chain traced through the deployment context; amplifying and mitigating factors named explicitly. The structured prompt produces both a more accurate score and an auditable reasoning chain a human reviewer can inspect and challenge.

Illustrative Output Comparison: P31 (Proxy Discrimination) × Tenant Screening Case

VANILLA (Holistic Assessment)

Input: Full pattern (triggers, conditions, chain, harms)

reasoning:

"The tenant screening system uses neighbourhood deprivation indices and credit scoring data that could serve as proxies for protected characteristics. The case mentions a 23% approval rate disparity between areas with different demographic profiles. Several enabling conditions appear present, including reliance on spatial features and historical data reflecting segregation patterns. However, the system includes a fairness audit. Overall, there is substantial risk of indirect discrimination."

applicability_score: 0.72

- Mentions conditions but does not assess each
- References causal pathway without tracing it
- No systematic present/absent determination
- Reviewer cannot verify which evidence was used

Expert benchmark: 0.95

Error: |0.72 − 0.95| = 0.23

CBR (Structured Reasoning)

Input: Same pattern + RETRIEVE → REUSE → REVISE structure

RETRIEVE — Enabling Conditions:

- Spatial proxy features: PRESENT (deprivation index)
- Historical segregation data: PRESENT (redlining era)
- High-stakes individual decisions: PRESENT (housing)
- ○ Lack of fairness testing: ABSENT (audit exists)

Structural match: 0.75 (3/4 conditions present)

REUSE — Causal Chain:

1. Deprivation index encodes racial segregation →
2. Model learns area → risk correlation →
3. Screening reproduces exclusion patterns →
4. Disparate impact on protected groups

Mechanism strength: STRONG

REVISE — Contextual Adjustment:

Amplifying: 12,000 annual decisions, no appeals

Mitigating: Fairness audit (but post-hoc only)

applicability_score: 0.88

- **Each condition assessed with evidence**
- Causal mechanism traced step-by-step
- Reviewer can verify each determination

Expert benchmark: 0.95

Error: |0.88 − 0.95| = 0.07

Figure 3. Illustrative comparison of vanilla (left) and CBR (right) outputs for the same case-pattern pair. CBR produces a structured, auditable reasoning chain with individually assessed enabling conditions and a step-by-step causal trace. Scores and reasoning are illustrative of typical output patterns observed across the experiment.

3.6. Evaluation

We evaluated performance using mean absolute error (MAE, measuring average deviation from the expert benchmark), bias (systematic directional error, where negative values indicate the model scores below the expert), and Pearson correlation (linear association with expert scores). All comparisons use bootstrap confidence intervals with 10,000 iterations.

4. Results

Of 160 case-pattern pairs, 142 had complete data across all conditions (both vanilla and CBR, all three models, all three replications). The 18 missing pairs reflect response truncation failures from Gemini that could not be parsed. All reported analyses use this common set of 142 pairs.

4.1. Main findings

Table 3 reports MAE, bias, and Pearson correlation for all four conditions, with vanilla and CBR results shown per model and OWA conditions shown as the aggregated single value. We unpack the findings in turn: CBR's effect on assessment accuracy (§4.2), OWA's role as a calibration dial rather than a universal correction (§4.3), and the integration effect when both layers are combined (§4.4).

Table 3. Assessment accuracy across conditions. MAE = mean absolute error vs expert benchmark (lower is better). Bias = mean signed error (negative indicates model scores below expert). Correlation = Pearson r with expert benchmark.

Condition	MAE	Bias	Correlation	n
C1 Vanilla (Grok 4)	0.207	−0.172	0.739	142
C1 Vanilla (Gemini)	0.217	−0.063	0.696	142
C1 Vanilla (GPT-4o)	0.234	−0.230	0.764	142
C2 Vanilla + OWA	0.253	−0.248	0.746	142
C3 CBR (Grok 4)	0.201	−0.175	0.719	142
C3 CBR (Gemini)	0.225	−0.094	0.638	142
C3 CBR (GPT-4o)	0.171	−0.031	0.605	142
C4 CBR + OWA	0.218	−0.199	0.710	142

4.2. CBR improves assessment accuracy

CBR scaffolding produced a significant improvement for one of three models. GPT-4o showed a large effect, with MAE improving from 0.234 under holistic assessment to 0.171 under CBR (difference 0.062, 95% CI [0.034, 0.090], $p<0.001$). Grok 4 showed a small, non-significant improvement from 0.207 to 0.201 (difference 0.006, $p=0.37$).

Gemini showed no significant change (difference −0.009, p=0.51). The model-dependency of this result is a key finding: structured reasoning scaffolding helps some models substantially but not others, and the benefit cannot be predicted from a model's baseline performance.

CBR reduced negative bias substantially for GPT-4o (−0.230 to −0.031), making it the most closely aligned condition. Grok 4's bias was unchanged (−0.172 to −0.175), suggesting it was not engaging with the structured prompts in the same way. This contrasts with prior work where the same model benefited substantially from CBR on a pattern ranking task [7]: CBR may help a model identify which patterns are relevant (a retrieval task) without improving how applicable a pattern is scored (a calibration task). Models already performing well holistically on a given task may not benefit from the additional structure.

Correlation with the expert benchmark decreased under CBR for all models, most notably GPT-4o (0.764 to 0.605). This suggests CBR improves absolute calibration (scores closer to expert values) while slightly compressing the score range, reducing rank order agreement. For governance purposes absolute calibration is typically more important than rank ordering, since decisions depend on whether a score exceeds a threshold rather than on relative rankings.

4.3. OWA as governance alignment

OWA aggregation at orness 0.6 worsened performance when applied to vanilla outputs. All three models produced scores below the expert benchmark (negative bias ranging from −0.063 to −0.230), and OWA's precautionary weighting, which places more weight on lower ranked scores, amplified this gap. The resulting C2 condition had the highest MAE (0.253) and most negative bias (−0.248) of any condition. The effect was significant for all models (Grok 4: p<0.001, Gemini: p=0.015, GPT-4o: p<0.001).

This is consistent with the prior finding that OWA is effective calibration [6] while appearing to contradict it. In that study models showed overconfidence and precautionary weighting corrected in the right direction; here models show underconfidence and the same weighting amplifies the error. Task composition explains it: 16 patterns selected as most applicable produces a higher mean benchmark (0.56 vs 0.52) with 42% of scores in the high-risk range vs 28%. Models anchor toward the middle, overshooting low expert scores and undershooting high ones. The implication is that orness must be empirically calibrated for each assessment context including pattern-set composition, a strength of the framework, since OWA accommodates different risk postures rather than imposing a single correction.

4.4. The integration effect

The headline finding is that C4 (CBR plus OWA) significantly outperformed C2 (Vanilla plus OWA) with a MAE difference of 0.034 (95% CI [0.020, 0.048], p<0.001). This integration effect is driven by CBR improving the raw input scores rather than by OWA's aggregation mechanics. When the inputs to OWA are better calibrated through CBR scaffolding, the aggregated output is correspondingly more accurate.

However, the best single condition was GPT-4o with CBR alone (C3, MAE 0.171), outperforming every aggregated condition. The aggregated C4 (MAE 0.218) is worse than GPT-4o alone because OWA pools all three models and the weaker Gemini and Grok 4 inputs dilute GPT-4o's strong performance. This creates a tension between peak

accuracy from a single model and robustness from multi-model aggregation that is less sensitive to any one model's failures or version changes.

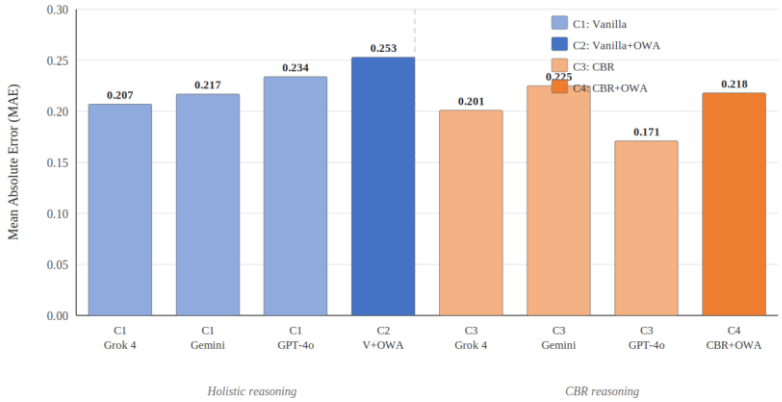


Figure 4. Mean absolute error by condition. Blue bars represent holistic (vanilla) reasoning; orange bars represent CBR structured reasoning. Darker shades indicate OWA aggregated conditions. Lower values indicate better alignment with expert benchmark.

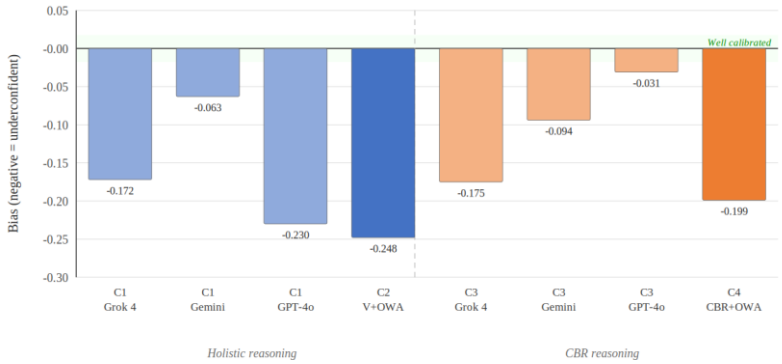


Figure 5. Calibration bias by condition. All conditions show negative bias (model scores below expert benchmark). The green band marks the well-aligned zone near zero. CBR scaffolding reduces the gap, particularly for GPT-4o. OWA at orness 0.6 amplifies the gap by weighting lower scores more heavily.

5. Discussion

5.1. Design principles for hybrid human-AI risk assessment

Structure the reasoning layer. CBR scaffolding forces models to systematically assess enabling conditions, trace causal mechanisms, and consider contextual adjustments rather than producing a snap judgment. This is only possible because the knowledge graph patterns provide an explicit causal structure to scaffold against, with each pattern's components forming the skeleton for the CBR phases. The result is a reasoning chain that is auditable by design: a human reviewer can inspect not just the final score but the intermediate assessments that produced it (Figure 3), which is essential for governance contexts where decision rationale must be documented and defensible.

Calibrate to the expert, not to an absolute standard. OWA's orness parameter encodes an organisational risk posture, with the appropriate setting determined by comparison with expert judgment, positioning the domain expert as an irreplaceable element of the pipeline. Different organisations with different risk appetites set different values, a feature of the framework rather than a limitation.

Validate per model and per task. CBR benefits are strongly model- and task-dependent. GPT-4o benefited substantially from CBR for applicability scoring while Grok 4 and Gemini did not. Yet prior work found Grok 4 benefited substantially from the same scaffolding on a pattern ranking task [7], suggesting the benefit depends on assessment type as well as model. Supplementary analysis of Grok 3 vs Grok 4 on this scoring task showed Grok 3 benefited (MAE δ 0.034, $p<0.001$) while Grok 4 did not (δ 0.011, $p=0.07$), adding a version dimension. CBR scaffolding should be empirically validated for each model, task, and version rather than assumed to transfer.

Retain human authority over governance parameters. The expert's role is threefold: providing the calibration anchor, setting the governance posture through the orness parameter, and reviewing CBR reasoning chains for face validity. None of these can be automated without undermining the framework. Together, the four principles map onto three NIST AI RMF [2] trustworthiness characteristics: *valid and reliable* through CBR's structured decomposition; *accountable and transparent* through auditable reasoning chains that let a human inspect each enabling-condition determination rather than face an opaque score; and *manageable* through OWA's tuneable orness, which surfaces the organisation's risk posture as an explicit, defensible parameter. The framework does not claim the LLM itself becomes trustworthy; it claims that the *pipeline*, LLM reasoning, structured scaffolding, expert-calibrated aggregation, and human review of intermediate steps, produces assessments an organisation can stand behind. The locus of trust shifts from the model to the workflow, with the expert as the calibration anchor that makes that trust defensible.

5.2. Reframing OWA for practice

The finding that OWA at orness 0.6 worsened rather than improved accuracy here carries a practical lesson. It would be tempting to treat a specific orness as a default; our results show why this would be a mistake. OWA is a calibration tool and like any calibration tool must be set for the specific measurement context. In practice, an organisation would first run a calibration exercise: have the expert score a sample of cases, compare with LLM outputs, determine the gap, and set orness accordingly. A precautionary regulator might set a lower orness; a higher-tolerance technology firm a higher one. The framework accommodates both without modification.

A related point on robustness: GPT-4o with CBR alone outperformed every aggregated condition. Multi-model aggregation looks costly when one model dominates, but model capabilities shift with updates that the user does not control. Aggregation insures against this fragility, C4 is worse than GPT-4o alone but substantially better than the worst single model. The portfolio approach trades peak accuracy for resilience over time.

5.3. Limitations and future work

Single expert benchmark. The benchmark reflects one expert's judgment. This is not a sample-size limitation: as set out in §3.3, professional risk assessment has no ground

truth, only calibration targets, and pooling across experts to chase an "average" benchmark would obscure rather than reveal the methodological behaviour the study characterises. Different experts would produce different scores and likely different optimal orness values, but the structural design principles, which layer to intervene at, whether to use CBR or fuzzy decomposition, the need to validate per model, are independent of any specific expert. The contribution is the method, not the numeric values it produces against this target.

Single domain. All cases involve real estate AI deployment. The procedure is domain-general: the pattern structure, the CBR phase mapping (RETRIEVE→REUSE→REVISE), and OWA calibration are not specific to real estate, and the 87-pattern knowledge graph from which the 16 patterns were drawn was synthesised across healthcare, criminal justice, financial services, and real estate [5]. Applying the framework in another domain requires substituting the deployment cases and re-running the expert calibration step; it does not require redesigning the pipeline. Testing across healthcare, financial services, and public sector deployments would strengthen generalisability.

Static evaluation and model versions. This study evaluates a single round of assessment with three models as available in February 2026. Risk assessment in practice is iterative, and as the Grok 3 vs Grok 4 comparison in §5.1 shows, model updates can change whether an intervention is beneficial, reinforcing the practical value of a framework that can be recalibrated rather than one assuming fixed model characteristics. Testing iterative feedback loops is an important next step.

6. Conclusion

The findings across 2,880 trials support layer-appropriate augmentation: CBR improves reasoning quality and score calibration at the input layer for models that benefit from structured decomposition, OWA provides a tuneable governance dial at the output layer, and their combination outperforms calibrated holistic assessment, though the optimal configuration requires empirical calibration to expert judgment and validation per model.

These design principles position the domain expert as the essential calibration anchor for trustworthy AI risk assessment. As governance frameworks mature and the demand for structured assessment grows, hybrid systems combining scalable LLM reasoning with human expert calibration offer a practical path forward. The expert is not replaced. The expert is amplified.

References

[1] European Parliament. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Off J Eur Union. 2024.

[2] National Institute of Standards and Technology. AI Risk Management Framework (AI RMF 1.0). Gaithersburg (MD): NIST; 2023. Report No.: NIST AI 100-1.

[3] Geng J, Cai F, Wang Y, Koeppl H, Nakov P, Gurevych I. A survey of confidence estimation and calibration in large language models. In: Proceedings of NAACL; 2024. p. 6577–6595.

[4] Tian K, Mitchell E, Zhou A, et al. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In: Proceedings of EMNLP; 2023. p. 5433–5442.

[5] McConomy G, Bond R, Glass D, Liu J, Toner H. Augmenting large language models with causal risk patterns for AI deployment risk assessment. In: Proc. IEEE International Conference on AI and Data Analytics (ICAD 2026). In press.
[6] McConomy G, Liu J, Glass D, Bond R. Bridging causal reasoning and fuzzy calibration for trustworthy AI risk assessment. In: Proc. FLINS-ISKE 2026. In press.
[7] McConomy G, Bond R, Glass D, Liu J, Toner H. Structured reasoning for AI risk assessment: empirical evidence of model-dependent benefits from case-based decomposition. Under review at KI 2026.
[8] Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of ACM FAT*; 2020. p. 33–44.
[9] Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun. 1994;7(1):39–59.
[10] Yager RR. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans Syst Man Cybern. 1988;18(1):183–90.
[11] Yager RR. Quantifier guided aggregation using OWA operators. Int J Intell Syst. 1996;11(1):49–73.
[12] Floridi L, Cowls J, Beltrametti M, et al. AI4People: an ethical framework for a good AI society. Minds Mach. 2018;28(4):689–707.
[13] Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. In: Proceedings of ACM FAT*; 2019. p. 59–68.
[14] Methnani L, Aler Tubella A, Dignum V, Theodorou A. Let me take over: variable autonomy for meaningful human control. Front Artif Intell. 2021;4:737072.
[15] Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. 2016.
[16] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. Int Data Priv Law. 2017;7(2):76–99.
[17] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
[18] Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems - an introductory survey. WIREs Data Min Knowl Discov. 2020;10(3):e1356.
[19] Binns R. Fairness in machine learning: lessons from political philosophy. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 81); 2018. p. 149–59.
[20] Shneiderman B. Human-centered AI. Issues Sci Technol. 2021;37(2):56–61.