Bridging Causal Reasoning and Fuzzy Calibration for Trustworthy AI Risk Assessment

Gareth McConomy[1], Jun Liu[1], David Glass[1], Raymond Bond[1], and Holly Toner[2]

[1] Ulster University, Northern Ireland, UK
[2] Kainos plc, Belfast, UK

Abstract. When organisations deploy AI systems in high-stakes domains, they need risk assessments they can trust. Large language models offer a promising path to scalable assessment, but two problems undermine their reliability: shallow reasoning that flags risks without tracing causal mechanisms, and untuned aggregate output that sits systematically above cautious expert postures. We address both problems with a two-layer framework. **Causal risk patterns** structure the reasoning input: each pattern documents a known harm mechanism with its trigger conditions, causal steps, and empirical evidence. **Ordered Weighted Averaging (OWA)** calibrates the output by weighting multiple LLM assessments according to rank rather than source, systematically aligning aggregate output with a chosen expert calibration target.
In experiments with three frontier LLMs assessing 125 case-pattern pairs across five real estate AI deployments, all models produced ratings sitting +0.07 to +0.16 above our expert target on a 0–1 scale. OWA aggregation calibrated to this target reduced mean absolute error by 7.7% at orness=0.6 versus the best individual model (p=0.023); we frame orness as a tunable governance-posture parameter that must be calibrated to context, not a universal constant. Our most surprising finding concerns a *layer asymmetry*: fuzzy aggregation at the output layer improves calibration, but fuzzy scaffolding at the reasoning layer *degrades* performance (34.8% error increase). This asymmetry, where LLMs reason better holistically than when forced into structured decomposition, has practical implications for anyone designing human-AI workflows.

Keywords: OWA operators · LLM calibration · AI risk assessment · fuzzy aggregation · causal reasoning · NIST AI RMF · human-AI teaming

1 Introduction

1.1 The Problem: AI Risk Assessment Doesn't Scale

Consider a property management company evaluating whether to deploy an AI system that automatically sets rental prices. A thorough risk assessment would trace how such a system could lead to harm: if multiple landlords in the same market use similar pricing algorithms, those algorithms might learn to

coordinate on higher prices without any explicit collusion, a pattern documented in reinforcement learning research where pricing agents learnt supra-competitive strategies purely through market feedback [1].

This kind of causal reasoning is exactly what expert risk assessors do. They don't just check boxes; they trace mechanisms from trigger to harm. The problem is that expert assessment doesn't scale. A large enterprise might deploy hundreds of AI systems, each requiring this kind of careful analysis.

Large language models seem like a natural solution. They have broad knowledge about AI risks and can process information quickly. But can their assessments be trusted?

1.2 Two Gaps in LLM Risk Assessment

Our research identifies two distinct problems:

Gap 1: Shallow reasoning. When asked to assess AI deployment risks, LLMs typically identify relevant risk categories but fail to trace the causal mechanisms through which harms actually unfold [2,3]. They might note "potential for price manipulation" without explaining *how* algorithmic pricing could lead to coordination, *what conditions* would enable this, or *what evidence* supports the concern. This shallow reasoning makes it hard for human reviewers to judge whether the assessment is sound.

Gap 2: Untuned aggregate output. Even when LLMs identify the right risks, their applicability ratings sit systematically above the postures of cautious risk professionals [4,5,6]. A risk mechanism that an experienced expert would rate at 0.5 applicability may receive 0.65 from raw LLM aggregation. Without calibration to a chosen risk posture, this gap means LLM output cannot be acted upon directly: organisations must either escalate every warning (alert fatigue) or apply ad-hoc filters that lack governance traceability.

1.3 Our Solution: Causal Patterns + Fuzzy Calibration

We propose a two-layer solution (Fig. 1):

Layer 1: Causal risk patterns address the reasoning gap. A causal risk pattern is a structured documentation of a known AI harm mechanism: the causal chain from trigger to harm, the conditions that must be present, and empirical evidence from research or real incidents.

Layer 2: Ordered Weighted Averaging (OWA) addresses the calibration gap [7]. Rather than taking a single LLM's assessment at face value, we collect multiple assessments and aggregate them using OWA, which weights assessments by their *rank position* rather than their source. If LLMs share an offset from a chosen target, we can configure OWA to give more weight to lower assessments, systematically aligning the aggregate with the target.

Our main contribution is the critical insight that these interventions must operate at *different layers*. Causal patterns structure the **input** to reasoning; OWA calibrates the **output**. When we tried applying structured decomposition

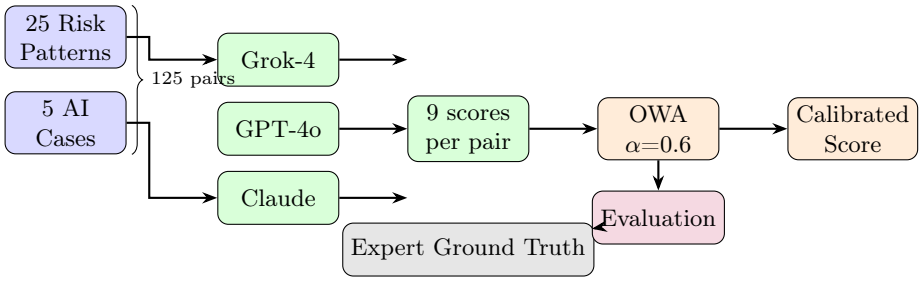


Fig. 1. The Causal-Fuzzy Framework. Causal patterns and deployment cases are assessed by three LLMs (3 runs each = 9 scores per pair). OWA aggregation produces calibrated scores evaluated against expert ground truth.

to the reasoning process itself, performance got worse, not better. We call this the *layer asymmetry*.

1.4 Contributions

This paper makes five contributions:

1. **Documents the untuned LLM aggregate offset**: Three frontier models produce ratings +0.07 to +0.16 above the expert calibration target on a 0–1 scale.
2. **Demonstrates OWA calibration correction**: Aggregation against our expert target reduces MAE by 7.7% at orness=0.6 (p=0.023).
3. **Reveals the layer asymmetry**: Fuzzy methods at aggregation reduce error by 7.7%, but at reasoning they *increase* error by 34.8%.
4. **Frames OWA as a tunable calibration recipe**: The optimal orness is expert-, domain-, and context-conditional, requiring monitoring rather than universal specification.
5. **Aligns with NIST AI RMF**: The framework implements "Measure" function requirements [8].

2 Background

Before diving into our method, we need to cover two concepts that may be unfamiliar: AI risk patterns and OWA aggregation.

2.1 AI Risk Patterns: Documenting How Things Go Wrong

An AI risk pattern is essentially a "recipe for harm", a documented mechanism by which AI systems have caused or could cause problems. The idea borrows from design patterns in software engineering [9], but instead of documenting reusable solutions, we document reusable problems. Similar approaches appear in AI auditing frameworks [10] and risk repositories [11].

Each pattern includes four elements:

- **Causal mechanism**: The pathway from system properties to harm
- **Enabling conditions**: What must be present for the mechanism to operate
- **Empirical evidence**: Documented cases or research findings
- **Quantified effects**: How bad it gets, where data exists

Example: Tacit Algorithmic Collusion. Independent pricing algorithms can converge on above-market prices through shared market exposure and learned punishment strategies. Calvano et al. [1] demonstrated this in simulation: Q-learning pricing agents achieved 9–13% supra-competitive price increases without any explicit collusion programming. The enabling conditions are: (1) multiple agents with similar optimisation objectives, (2) ability to observe competitor prices, (3) repeated interactions over time, and (4) similar algorithmic approaches.

Why structure risks this way? Because it enables *analogical reasoning* [12,13]. The tacit collusion pattern was documented in abstract pricing simulations, but the mechanism could apply anywhere the enabling conditions are met: rental pricing, airline tickets, hotel rooms. By making the causal structure explicit, we help both LLMs and human reviewers spot when a pattern from one domain might transfer to another.

2.2 OWA: A Flexible Way to Combine Multiple Opinions

Ordered Weighted Averaging combines multiple values into a single aggregate [7,14]. What makes OWA special is that it weights values by their *rank position* rather than their source.

The basic idea: Suppose three LLMs assess a risk and produce scores of 0.8, 0.7, and 0.6. Simple averaging gives 0.7. With OWA, we first sort the values high-to-low (0.8, 0.7, 0.6), then apply weights to each position. Using weights [0.2, 0.3, 0.5] that favour lower values: $0.2 \times 0.8 + 0.3 \times 0.7 + 0.5 \times 0.6 = 0.67$.

Why this matters for LLM calibration: If all three LLMs share a similar offset from a chosen calibration target, and they tend to because they were all trained with similar RLHF techniques that reward confident-sounding answers, then simple averaging preserves the offset rather than correcting it. OWA lets us systematically pull the aggregate toward the target by re-weighting ranked positions, without knowing which specific model contributes most to the shared offset.

The orness parameter: Rather than specifying individual weights, we control OWA's behaviour with a single number called "orness" ($\alpha \in [0, 1]$) [15,16]:

- $\alpha = 0.0$: Take the maximum (pure optimism, trust the highest estimate)
- $\alpha = 0.5$: Equal weights (same as simple averaging)
- $\alpha = 1.0$: Take the minimum (pure pessimism, trust the lowest estimate)
- $\alpha = 0.6$: Mild pessimism, slightly discount high values

We interpret orness as a *governance posture*: how cautious should the organisation be? Nuclear safety might warrant orness of 0.7 (strong precaution), while consumer product recommendations might tolerate 0.3 (mild optimism).

Algorithm 1 Exponential OWA Weight Derivation

Require: Orness $\alpha \in (0, 1)$, number of inputs n
Ensure: Weight vector $\mathbf{w} = (w_1, \ldots, w_n)$
if $\alpha = 0.5$ **then**
  $w_j \leftarrow 1/n$ for all j {Uniform weights}
else if $\alpha > 0.5$ **then**
  $\beta \leftarrow \alpha/(1-\alpha)$
  $\tilde{w}_j \leftarrow \beta^{j-1}$ {Favour higher ranks}
else
  $\beta \leftarrow (1-\alpha)/\alpha$
  $\tilde{w}_j \leftarrow \beta^{n-j}$ {Favour lower ranks}
end if
Normalise: $w_j \leftarrow \tilde{w}_j / \sum_k \tilde{w}_k$
return $\mathbf{w}$

This connects OWA to broader work on fuzzy decision-making under uncertainty [17,18].

Algorithm 1 computes weights from a target orness using an exponential approach.

Why exponential weights? The fuzzy literature offers several methods for deriving OWA weights from a target orness: maximum entropy [19], minimax disparity [20], least-squares optimisation, and RIM quantifier functions [21]. We chose the exponential approach because it yields weights in closed form, produces a smooth monotonic distribution (avoiding the discontinuities of minimax methods), and gives intuitive gradation. With $\alpha = 0.6$ and $n = 9$, the highest-ranked score receives weight $\approx$0.16 and the lowest $\approx$0.08: a gentle ramp suited to mild correction of shared bias rather than aggressive outlier handling. Empirical ablation against alternative methods is direct future work (Section 7).

2.3 Why Untuned LLM Aggregation Drifts Toward Confident Output

Understanding *why* LLMs tend toward confident-sounding output helps explain why calibration matters.

Modern LLMs are trained using Reinforcement Learning from Human Feedback (RLHF) [22]. Human raters evaluate model outputs, and the model learns to produce responses that score well. The problem is that humans rate confident-sounding responses more favourably than hedged ones [23]. "The risk is significant" sounds more authoritative than "The risk might be moderate, depending on context." Over many training iterations, models learn that confidence gets rewarded.

Research has documented this across multiple settings. Studies report that 84% of LLM scenarios exhibit confident outputs on biomedical tasks [24], and calibration drift increases on complex judgement tasks where ground truth is uncertain [25]. Recent work proposes calibration methods [26,27], but these focus

on factual question-answering with objective targets, not on calibration to a chosen risk posture.

For AI risk assessment, this creates a systematic problem: untuned LLM aggregation will tend to sit above the postures of cautious risk professionals, making raw output unsuitable for direct governance use without calibration.

3 Method

3.1 Research Questions

We designed an experiment to answer four questions:

RQ1: Does untuned LLM aggregate output sit systematically away from a chosen expert calibration target?
RQ2: Can OWA aggregation correct this bias?
RQ3: What orness value produces the best calibration?
RQ4: Does it matter *where* we apply fuzzy methods, reasoning or aggregation?

3.2 Materials

Risk patterns: The 25 AI risk patterns used here derive from a structured literature review of documented AI harms, conducted in companion work [39]. Patterns were extracted where the literature provides a documented causal mechanism, identifiable enabling conditions, and at least one cited piece of empirical evidence (research finding or incident). The 25 patterns cover algorithmic bias and discrimination (5 patterns), automation and human factors (5), privacy and surveillance (4), market dynamics (4), and feedback loops and emergent behaviour (7), drawing on AI risk repositories [11,28]. We make no claim of exhaustive coverage: the catalogue is a documented working sample, scoped for tractable evaluation; expansion of the catalogue is ongoing. Each pattern follows the structure from Section 2.1.

Deployment cases: We created five realistic AI deployment scenarios in real estate:

1. **OptaPrice**: Algorithmic rent pricing used by 15 property managers in the same metropolitan market, observing competitor pricing, occupancy rates, and market conditions.
2. **TenantScreen**: ML tenant screening based on credit history, rental history, employment verification, and social media signals.
3. **MaintainAI**: Predictive maintenance for commercial properties with 128 facilities staff affected, scheduling repairs based on sensor data and cost optimisation.
4. **ValueAI**: Computer vision property valuation using images, floor plans, and neighbourhood characteristics.
5. **SmartBuild**: Integrated building management with occupancy monitoring, controlling HVAC, lighting, and access.

Why real estate? Two reasons. First, the lead author is a Fellow of the Royal Institution of Chartered Surveyors with 20+ years in the industry, which means we can rigorously evaluate whether LLM outputs reflect genuine domain understanding or just superficial pattern matching. Second, real estate AI deployments involve high-stakes decisions affecting housing access, pricing, and potential discrimination [29].

Ground truth: The FRICS-qualified author assessed all 125 case-pattern pairs, providing applicability scores from 0 to 1. The assessment followed the same holistic approach we asked LLMs to use: read the case and pattern, then judge how plausibly the mechanism could operate in that context. Distribution: 47 low (<0.4), 59 medium (0.4–0.7), 19 high (>0.7).

3.3 LLM Assessors and Approaches

We used three frontier LLMs: Grok-4 (xAI), GPT-4o (OpenAI), and Claude Sonnet 4 (Anthropic). Using models from different providers helps ensure any shared bias isn't specific to one company's training approach [30].

Each model assessed each case-pattern pair two ways:

Vanilla (holistic) assessment: The model received the case description and pattern documentation, then provided a single applicability score with reasoning. This mirrors how a human expert would approach the task.

Fuzzy-scaffolded assessment: The model first scored each enabling condition separately (e.g., "Are multiple agents present? Score 0–1"), then provided an overall assessment. We thought this structured decomposition might improve accuracy, based on research showing explicit reasoning steps can help LLMs [31,32].

Each model-pair-approach combination ran 3 times to capture variability: 3 models × 125 pairs × 3 runs × 2 approaches = 2,250 total assessments.

All assessments used temperature 0.3 to balance consistency with some exploration.

3.4 Aggregation Methods

We compared three approaches:

1. **Best individual model**: The single model with lowest MAE
2. **Simple mean**: Average of all 9 assessments (3 models × 3 runs)
3. **OWA**: Ordered weighted averaging with various orness values

For OWA, we tested orness from 0.1 to 0.9 in increments of 0.1.

3.5 Evaluation Metrics

We measured three things:

- **Mean Absolute Error (MAE)**: Average of |prediction − target|. Lower is better.
- **Correlation**: Pearson correlation with the target. Higher is better.

Table 1. Individual LLM performance against the expert calibration target. Bias = mean (model − target).

Model	MAE ↓	Correlation ↑	Bias
Grok-4	0.145	0.696	+0.071
GPT-4o	0.138	0.687	+0.074
Claude Sonnet 4	0.202	0.603	+0.157

Table 2. Aggregation comparison: OWA outperforms alternatives. For MAE, lower is better; ΔMAE shows change from baseline (negative = improvement).

Method	MAE ↓	Correlation ↑	Bias	ΔMAE
GPT-4o (best individual)	0.138	0.687	+0.074	—
Simple mean	0.140	0.722	+0.100	+0.002 (worse)
OWA (orness=0.6)	**0.127**	0.714	**+0.036**	**−0.011 (7.7% better)**

– **Bias**: Average of (prediction − target). Positive indicates output sits above the target.

Statistical significance came from bootstrap resampling (10,000 iterations) to compute 95% confidence intervals [33].

4 Results

4.1 RQ1: All LLMs Sit Above the Expert Posture

Table 1 confirms what we suspected: all three models produce applicability ratings systematically above our expert calibration target.

All three show positive bias against this target. Claude sits furthest above (+0.157); Grok and GPT-4o sit similarly above (+0.07). This isn't random noise: it's a systematic offset that calibration must correct.

What this means in practice: Raw LLM aggregation cannot be trusted to produce output already aligned with a chosen risk posture; the offset is large enough that direct use without calibration would systematically over-flag deployments relative to what an experienced expert would consider warranted.

4.2 RQ2: OWA Aligns Aggregation with the Target

Table 2 compares aggregation methods. Two findings stand out:

Simple averaging fails to correct the offset. The mean has *larger* bias (+0.100) than the best individual model (+0.074). Why? All three models share a similar offset from this target, plausibly because they all carry similar RLHF-induced patterns. Averaging values that share the same offset preserves the offset rather than removing it.

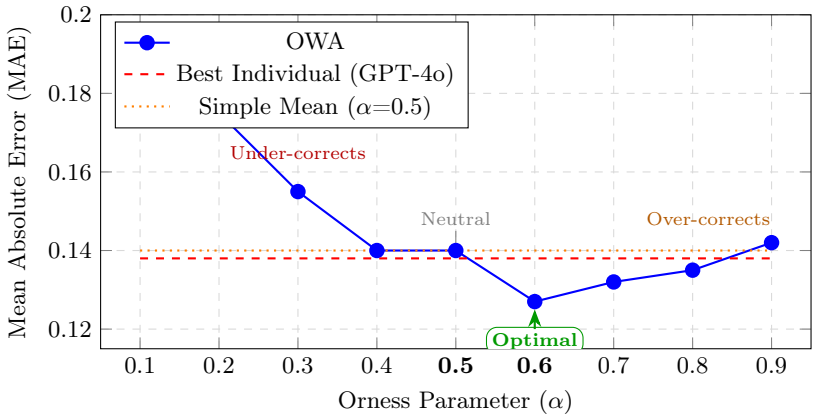


Fig. 2. The "Goldilocks Zone" for orness in this calibration. MAE is minimised at α=0.6 against our expert target. The location of the minimum is target- and context-conditional.

OWA significantly aligns aggregation. With orness=0.6, MAE drops to 0.127, a 7.7% reduction from the best individual model. Bias drops by 51% (from +0.074 to +0.036). This improvement is statistically significant: bootstrap 95% CI for ΔMAE is $[-0.021, -0.0002]$, p=0.023.

4.3 RQ3: The Goldilocks Zone for Orness

Testing across orness values reveals a clear sweet spot (Fig. 2):

- **Orness 0.1–0.3** (optimism): Under-corrects; aggregate sits too close to the highest LLM estimate, leaving the upward offset uncorrected.
- **Orness 0.5** (neutral): Equivalent to simple averaging, which preserves the offset.
- **Orness 0.6**: The empirical optimum for our calibration target.
- **Orness 0.7–0.9** (strong pessimism): Over-corrects; pulls aggregate too far below the target.

The optimal orness of 0.6 in this experiment represents a "mildly precautionary" stance calibrated to one FRICS-qualified expert assessing real estate AI deployments. We expect this value to vary across experts, domains, and over time as models and contexts evolve. The contribution is the calibration recipe and its tunability, not a universal value of α; organisations adopting the framework should expect to calibrate and periodically recalibrate against their own expert benchmarks.

4.4 RQ4: The Layer Asymmetry

This is our most surprising result (Table 3):

Table 3. The layer asymmetry: fuzzy methods help at aggregation but hurt at reasoning. ΔMAE shows change from vanilla baseline; negative = improvement.

Configuration	Fuzzy Layer	MAE	ΔMAE	Result
Vanilla LLM	None	0.138	—	Baseline
Fuzzy-scaffolded LLM	Reasoning	0.186	+0.048 (+34.8%)	Worse
OWA aggregation	**Aggregation**	**0.127**	**−0.011 (−7.7%)**	**Better**

Fuzzy scaffolding at the reasoning layer increases error by 34.8%. When we forced LLMs to score each enabling condition separately before making an overall assessment, accuracy got substantially worse.

Fuzzy aggregation at the output layer reduces error by 7.7%. When we let LLMs reason holistically but then applied OWA to aggregate their outputs, accuracy improved.

Why does this happen? We hypothesise three mechanisms, which together constitute what we call the "reductive fallacy" in AI-assisted judgement: the mistaken assumption that decomposing a complex assessment into parts will improve accuracy:

1. **Fragmentation**: Scoring conditions separately loses the holistic understanding needed for risk mechanisms that require conditions A *and* B *and* C operating together in specific configurations [37].
2. **Anchoring**: Once the LLM generates individual condition scores, those numbers anchor the overall assessment in ways that don't reflect expert judgement [38].
3. **Modality mismatch**: Our expert (calibration target) assessed holistically; forcing a different reasoning structure creates a mismatch between how we evaluate and how the expert actually thinks.

This finding contrasts with results on chain-of-thought prompting [31], suggesting that *eliciting* natural reasoning ("think step by step") differs fundamentally from *imposing* external structure ("score each condition separately"). The former asks the model to show its work; the latter constrains how that work must be done.

4.5 Illustrating the Asymmetry: A Concrete Example

To make the layer asymmetry concrete, consider how GPT-4o assessed the "Tacit Algorithmic Collusion" pattern against the OptaPrice case (algorithmic rent pricing with 15 competing property managers):

Vanilla (holistic) assessment:

> *"OptaPrice operates in a shared metropolitan market where 15 property managers use similar algorithmic pricing systems. The combination of price observability through market data feeds, repeated daily pricing interactions, and aligned profit-maximisation objectives creates fertile ground*

for emergent coordination. While there's no explicit communication, the market structure mirrors the conditions under which Calvano et al. observed learned collusive behaviour. The mechanism is plausible though not certain—smaller markets with fewer players show stronger effects in the literature." **Score: 0.72**

Fuzzy-scaffolded assessment:

Condition 1 (Multiple agents with similar objectives): "15 property managers, all optimising for rental yield." **0.90**
Condition 2 (Price observability): "Market data feeds provide competitor pricing." **0.85**
Condition 3 (Repeated interactions): "Daily pricing decisions over months/years." **0.95**
Condition 4 (Similar algorithms): "Likely similar ML approaches given vendor market." **0.80**
Overall assessment: "All enabling conditions are substantially present." **Score: 0.88**

The expert calibration target for this case-pattern pair was **0.70**.

The holistic assessment reasoned about how conditions *interact*, tempering confidence with "smaller markets with fewer players show stronger effects." The scaffolded version scored each condition in isolation, found all high, and anchored on those high sub-scores. Across our dataset, scaffolded assessments scored +0.09 above holistic ones on average, enlarging rather than reducing the offset from the target.

4.6 Performance by Case Difficulty

Breaking down performance by ground truth difficulty reveals where OWA helps most (Fig. 3):

OWA helps most on edge cases: situations where the risk mechanism has low applicability against the calibration target but raw LLM aggregation rates it as moderate. These are exactly the cases where the offset matters most for governance: false positives that distract from genuine risks and erode analyst attention.

5 Discussion

5.1 Practical Implications

For practitioners building AI risk assessment systems, our findings point to a specific workflow:

1. **Use multiple LLMs from different providers and aggregate with OWA, not simple averaging.** The offset from any cautious target appears across providers; simple averaging cannot correct a shared offset.

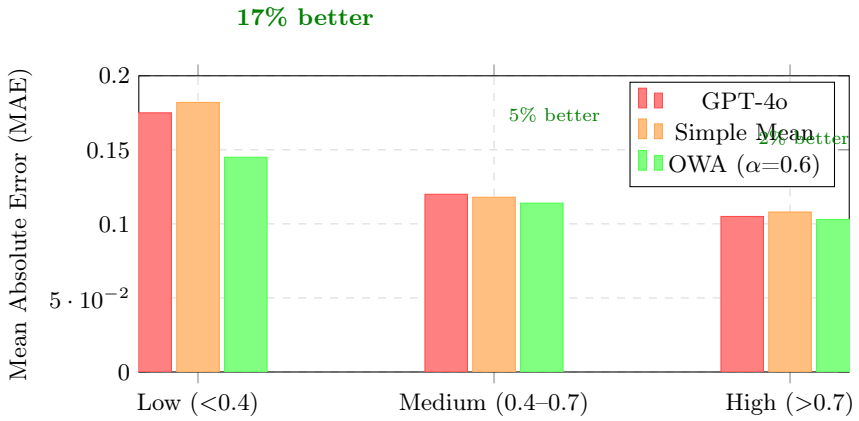


Fig. 3. OWA improvement by case difficulty. OWA provides the greatest benefit on low-applicability cases (17% error reduction), exactly where the raw-LLM offset causes the most governance problems.

2. **Calibrate to your expert target.** Orness ≈ 0.6 worked as a starting point against our FRICS-qualified expert; treat this as a calibration starting value, not a default. Dial up (0.7–0.8) for safety-critical contexts; dial down (0.4–0.5) for lower-stakes applications. Monitor and recalibrate as models and contexts shift.
3. **Prioritise low-applicability cases for human review.** This is where the raw-LLM offset causes the biggest governance problems, and where OWA helps most.
4. **Let LLMs reason holistically; calibrate their outputs.** Avoid imposing structured reasoning templates; calibrate after natural assessment.

5.2 Alignment with NIST AI Risk Management Framework

Our framework maps directly to requirements from the NIST AI RMF "Measure" function [8]:

- **MEASURE 1.1**: "Approaches for measuring AI risks are selected." Causal pattern applicability provides a principled measurement approach grounded in documented mechanisms.
- **MEASURE 2.1**: "Evaluation results are documented." Explicit scoring and aggregation creates an audit trail showing how assessments were derived.
- **MEASURE 2.6**: "Regular evaluation of AI systems for safety." The scalable pipeline enables ongoing monitoring across many deployments.
- **MEASURE 4.1**: "Measurement approaches for trustworthy AI are employed." Bias correction directly addresses trustworthiness by quantifying and correcting systematic errors.

The orness parameter offers something particularly valuable: a transparent, tunable, and *monitorable* way to encode governance posture. Rather than hiding risk tolerance in opaque model weights, organisations can state explicitly: "We use orness=0.5 for routine assessments and orness=0.4 for high-stakes deployments," and revise these values as expert benchmarks, models, and contexts evolve.

5.3 Theoretical Contribution: Layer-Appropriate Augmentation

The layer asymmetry points to a broader principle for human-AI system design:

AI augmentation should be layer-appropriate. Methods that structure inputs (like causal patterns providing context) or calibrate outputs (like OWA correcting bias) can improve performance. But methods that constrain the reasoning process itself may interfere with capabilities the AI already has.

Before imposing structure on an AI workflow, ask: does this structure *enable* the AI to do something it couldn't otherwise do, or does it *constrain* something it was already doing well? This connects to emerging work on human-AI complementarity [34,35] and appropriate reliance [36].

The layer asymmetry also echoes classical systems thinking: decomposing wholes into parts can destroy emergent properties that arise from interactions [37]. LLMs, trained on vast corpora of human reasoning, may have internalised holistic assessment patterns that structured decomposition disrupts.

5.4 Validation Approach and Effect Size in Context

We should clarify the role of expert assessment. Rather than claiming objective "ground truth," our expert (FRICS-qualified, 20 years experience) provides a **professional calibration benchmark**, a realistic target representing how an experienced practitioner may assess risk applicability. The contribution is the calibration recipe: OWA systematically aligns aggregate LLM output with this expert target, with orness as a tunable parameter that should be re-calibrated as expert profiles, domains, and models change.

Effect size, small in the lab, meaningful at scale: The 7.7% MAE reduction corresponds to a modest effect size (Cohen's $d \approx 0.11$). In a laboratory setting with 125 case-pattern pairs, this might seem incremental. But at enterprise scale (a property firm running 500 automated risk assessments per quarter), uncalibrated LLM aggregation would systematically over-flag relative to the chosen target. A 7.7% MAE reduction, concentrated in low-applicability cases (17% improvement there), translates to roughly 35–40 fewer unnecessary escalations per quarter, plus the harder-to-quantify benefit of reduced alert fatigue. OWA is not a replacement for expert judgement; it is a filter that lets experts focus where it matters most.

6 Conclusion

This paper presents a two-layer framework for trustworthy AI risk assessment:

- **Causal risk patterns** provide qualitative structure: the "how" of risk mechanisms, enabling LLMs to trace chains from trigger to harm.
- **OWA calibration** provides quantitative correction: the "how much" of risk magnitude, aligning aggregate output with the chosen calibration target.

Our experiments with three frontier LLMs assessing 125 case-pattern pairs yield four main findings:

1. Frontier LLM aggregate output sits systematically above our expert calibration target (+0.07 to +0.16) when assessing AI risk applicability.
2. OWA aggregation with orness=0.6 reduces mean absolute error by 7.7% (p=0.023).
3. The layer asymmetry: fuzzy methods at aggregation reduce error by 7.7%, but at reasoning they increase error by 34.8%.
4. In this calibration, the optimal orness of 0.6 represents a "mildly precautionary" governance posture for our expert target; the value is expected to vary across experts, domains, and over time, requiring monitoring and recalibration rather than fixed specification.

The framework enables scalable, calibrated AI risk assessment aligned with NIST AI RMF requirements, supporting trustworthy AI governance in practice. More broadly, the layer asymmetry offers a design principle for human-AI workflows: augment capabilities at the right layer, rather than assuming that more structure is always better.

For the fuzzy systems community, our work opens a new application domain (LLM calibration) where classical aggregation operators like OWA provide principled solutions to emerging AI governance challenges. The orness parameter, reinterpreted as a tunable governance posture calibrated to a specific expert target, bridges technical calibration methods and organisational risk management policy in a way that is transparent, tunable, and monitorable.

7 Future Work

Our findings suggest three directions for fuzzy systems research in AI governance.

First, **dynamic orness via fuzzy control**: treating calibration as a feedback loop where α adjusts based on recent MAE trends against validation cases would enable long-term deployment despite model drift. A PI (Proportional-Integral) controller, with $\alpha_{t+1} = \alpha_t + k_p \cdot \text{error}_t + k_i \cdot \int \text{error}\, dt$, could smoothly increase α when bias creeps upward, adapting governance posture automatically without manual retuning.

Second, **type-2 fuzzy calibration**: the orness parameter reflects organisational risk tolerance, which is itself vague and context-dependent. Representing $\tilde{\alpha}$ as an interval $[0.55, 0.65]$ would produce calibrated score ranges rather than point estimates, directly encoding governance ambiguity into outputs.

Third, **fuzzy measures for correlated offsets**: OWA assumes independence, yet LLM offsets from a calibration target tend to correlate across providers

due to shared RLHF training. Choquet integrals could learn interaction indices from calibration data, formalising how much to discount consensus drift versus genuine model disagreement.

Fourth, **broader empirical comparisons**: three direct extensions are outstanding. Systematic comparison of OWA against established LLM aggregation techniques (Self-Consistency [31], temperature-based ensembling, learned calibrators) would situate our results in the broader calibration literature. Ablation of OWA weight-derivation methods (exponential, maximum entropy, minimax disparity) would substantiate the qualitative choice argued in Section 2.2. Pairwise correlation analysis among model predictions would empirically substantiate the shared-offset mechanism we hypothesise for the simple-mean failure.

A journal extension of this work, currently in preparation, develops design criteria for target-relative LLM calibration methods (faithfulness, tunability, stability, auditability, and monitorability) and evaluates OWA-family methods against alternative calibration approaches under these criteria.

These extensions preserve the layer asymmetry while building robust, adaptive fuzzy augmentation for trustworthy AI assessment.

8 Limitations

Our work has three main constraints. First, the framework was calibrated against one FRICS-qualified expert. This is by design: calibration is to a chosen target, not toward objective truth. The contribution is bounded: results characterise OWA's behaviour against this specific calibration target. How calibration recipes transport across context (model versions, deployment contexts, temporal drift) under a single chosen target is the natural direction for follow-on work.

Second, the 7.7% MAE reduction is modest in effect size ($d \approx 0.11$) but meaningful at enterprise scale, translating to dozens of avoided false escalations quarterly. Whether this improvement changes actual governance decisions requires field trials, not just benchmark evaluation.

Third, our real estate domain provided a rigorous testbed but bounds the calibration target. Risk pattern structures and optimal orness values are expected to vary across sectors (healthcare, criminal justice, finance) and over time as models evolve, a variation predicted by the calibration-recipe framing rather than a defect in it. Cross-domain calibration mapping is direct future work.

These limitations bound current evidence and align with the fuzzy extensions proposed above.

Acknowledgments. The authors acknowledge the use of Claude (Anthropic) to assist with literature discovery, drafting, and prose refinement during manuscript preparation. All substantive intellectual contributions, experimental design, data collection, analysis, and interpretation remain the authors' own work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S.: Artificial intelligence, algorithmic pricing, and collusion. American Economic Review **110**(10), 3267–3297 (2020)
2. Jin, Z., Liu, Y., Belinkov, Y., et al.: Can large language models infer causation from correlation? In: Advances in Neural Information Processing Systems, vol. 36 (2023)
3. Joshi, N., Ruan, Q., Pierson, E., Kang, M.: LLMs are prone to fallacies in causal inference. In: Proc. EMNLP, pp. 10553–10569 (2024)
4. Kadavath, S., Conerly, T., Askell, A., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)
5. Xiong, M., Hu, Z., Lu, X., et al.: Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation. In: Proc. ICLR (2024)
6. Tian, K., Mitchell, M., Zhou, Y.: Just ask for calibration: Strategies for eliciting calibrated confidence scores. In: Proc. EMNLP (2023)
7. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on Systems, Man, and Cybernetics **18**(1), 183–190 (1988)
8. Tabassi, E. (ed.): Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology (January 2023)
9. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley, Reading (1994)
10. Raji, I.D., Smart, A., White, R.N., et al.: Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proc. FAT*, pp. 33–44 (2020)
11. Slattery, P., Saeri, A., Grundy, E., et al.: The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. arXiv preprint arXiv:2408.12622 (2024)
12. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive Science **7**(2), 155–170 (1983)
13. Holyoak, K.J., Thagard, P.: Mental Leaps: Analogy in Creative Thought. MIT Press, Cambridge (1995)
14. Yager, R.R.: Centered OWA operators. Soft Computing **11**(7), 631–639 (2007)
15. Filev, D., Yager, R.R.: On the issue of obtaining OWA operator weights. Fuzzy Sets and Systems **94**(2), 157–169 (1998)
16. Xu, Z.: An overview of methods for determining OWA weights. International Journal of Intelligent Systems **20**(8), 843–865 (2005)
17. Zadeh, L.A.: Fuzzy sets. Information and Control **8**(3), 338–353 (1965)
18. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York (1980)
19. O'Hagan, M.: Aggregating template or rule antecedents in real-time expert systems with fuzzy set logic. In: Proc. 22nd Asilomar Conference on Signals, Systems and Computers, pp. 681–689 (1988)
20. Wang, Y.-M., Parkan, C.: A minimax disparity approach for obtaining OWA operator weights. Information Sciences **175**(1–2), 20–29 (2005)
21. Yager, R.R.: Quantifier guided aggregation using OWA operators. International Journal of Intelligent Systems **11**(1), 49–73 (1996)

22. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, vol. 35 (2022)
23. Zhou, C., Liu, P., Xu, P., Iyer, S., et al.: LIMA: Less is more for alignment. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
24. Cagnoli, M., Tasnim, S., Ge, J., Magge, A.: Calibration as trustworthiness in biomedical NLP. JAMIA Open (2025)
25. Geng, J., Cai, F., Wang, Y., et al.: A survey of confidence estimation and calibration in large language models. In: Proc. NAACL, pp. 6577–6595 (2024)
26. Shen, M., Yin, Y., Li, J., et al.: Thermometer: Towards universal calibration for large language models. arXiv preprint arXiv:2403.08819 (2024)
27. Zhuge, M., Liu, H., Li, Y., et al.: Overconfidence in LLM-as-a-Judge: Diagnosis and confidence-driven solution. arXiv preprint arXiv:2508.06225 (2025)
28. OECD: OECD AI Incidents Monitor. https://oecd.ai/en/incidents (2024)
29. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica (May 23, 2016)
30. Liang, P., Bommasani, R., Lee, T., et al.: Holistic evaluation of language models. Transactions on Machine Learning Research (2023)
31. Wei, J., Wang, X., Schuurmans, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837 (2022)
32. Yao, S., Yu, D., Zhao, J., et al.: Tree of thoughts: Deliberate problem solving with large language models. In: Advances in Neural Information Processing Systems, vol. 36 (2023)
33. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007)
34. Bansal, G., Wu, T., Zhou, J., et al.: Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proc. CHI, pp. 1–16 (2021)
35. Lai, V., Carton, S., Bhatnagar, R., et al.: Towards a science of human-AI decision making: A survey of empirical studies. In: Proc. ACM FAccT (2023)
36. Schemmer, M., Kühl, N., Benz, C., Satzger, G.: A systematic review on fostering appropriate trust in human-AI interaction. ACM Journal on Responsible Computing **1**(1), 1–35 (2024)
37. Koestler, A.: The Ghost in the Machine. Hutchinson, London (1967)
38. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157), 1124–1131 (1974)
39. McConomy, G., Liu, J., Glass, D., Bond, R.: 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 (2026)