Augmenting Large Language Models with Causal Risk Patterns for AI Deployment Risk Assessment

Gareth McConomy, Raymond Bond, David Glass, Jun Liu, and Holly Toner
School of Computing, Ulster University
Belfast, United Kingdom
mcconomy-g@ulster.ac.uk

***Abstract*—Large language models (LLMs) achieve impressive benchmark performance yet exhibit systematic gaps in real-world deployment contexts. This is particularly concerning for AI governance applications where LLMs might identify risk categories but fail to trace the causal mechanisms through which harms actually unfold. We propose that structuring documented AI risks as explicit causal patterns—sequential mechanisms, feedback loops, and quantified effects from empirical research—can scaffold more reliable LLM reasoning through human-AI collaboration. We test this through a controlled study comparing pattern-augmented LLMs against vanilla LLMs using GPT-4o and Grok 4 to analyse five real estate AI deployment cases. Results show substantial improvements: GPT-4o improved in causal depth with Cohen's d = 1.20, while Grok showed even larger effects (d = 1.93) alongside improved evidence grounding (d = 2.10). Validation using Gemini 2.5 Pro and DeepSeek as judges confirmed findings with moderate-to-high inter-rater agreement (r = 0.45 to 0.68). Qualitative analysis reveals explicit mechanistic transfer, with augmented outputs mapping patterns from domains such as aviation automation and reinforcement-learning pricing to novel real estate contexts. These findings suggest that structured causal knowledge can enable a new modality of human-AI collaboration for AI risk assessment.**

***Index Terms*—Large language models, AI risk assessment, causal reasoning, human-AI collaboration, AI governance, retrieval-augmented generation**

I. Introduction

The governance of AI systems requires anticipating how technical choices interact with social contexts to produce harm. This challenge has motivated substantial cataloguing efforts: the MIT AI Risk Repository taxonomises over 1,700 risks across 43 categories [1], the OECD AI Incidents Monitor tracks policy-relevant failures, and the AI Incident Database documents over 2,000 cases [2]. These resources answer *what* can go wrong but often lack specification of the underlying causal mechanisms explaining *how* and *why* risks unfold. This gap matters because expert risk assessors reason *causally*, tracing chains of cause and effect rather than simply categorising outcomes. When evaluating a novel AI-based system, experts recognise that a recommendation system's feedback dynamics might mirror those documented in criminal justice risk tools [3], or that automation bias patterns from aviation could manifest in clinical decision support [4]. Cognitive science research establishes that such analogical reasoning depends on *relational structure* rather than surface similarity [5], [6]: if two situations share the same causal relationships, knowledge may be transferable even if domains differ entirely. The challenge is whether LLMs can perform such structural mapping when appropriately scaffolded.

Large language models might seem suited to support analogical risk reasoning given their broad knowledge. However, a growing body of evidence reveals systematic limitations. LLMs achieve impressive benchmark performance but fail under distribution shift [7], [8], with contamination rates of 52 to 57% on popular benchmarks [9]. Research further shows that LLMs often reduce multi-step problems to linearised pattern matching rather than systematic inference [10], struggle with compositional generalisation [11], cannot reliably self-correct reasoning errors [12], and fail to compose information across inferential hops [13], [14]. Causal inference is especially problematic: models struggle to distinguish correlation from causation [15] and are prone to fallacies including confusing causal direction, ignoring confounders, and over-attributing causation [16].

These limitations suggest that while LLMs possess broad knowledge of AI risks, eliciting structured causal reasoning requires scaffolding. Chain-of-thought prompting improves reasoning over multiple steps by encouraging explicit intermediate steps [17]. Retrieval-augmented generation (RAG) provides relevant context at inference time, improving factual accuracy [18]. However, standard RAG retrieves semantically similar text (i.e., documents containing similar words or discussing similar topics) without guaranteeing *structural* relevance (i.e., whether the retrieved content shares the same causal relationships as the query context) [19].

This paper tests whether augmenting LLMs with *explicitly structured causal patterns*, documented mechanisms describing how risks unfold with sequential steps, feedback loops, and quantified effects from empirical research, can scaffold more explainable reasoning (i.e., reasoning that can be articulated in terms humans can evaluate, distinct from 'interpretability' which refers to understanding internal model mechanisms). Our key hypothesis is that such patterns are *domain-transferable*: a pattern like "tacit algorithmic collusion through shared optimisation," documented in reinforcement learning pricing simulations [20], should transfer to any context where independent agents share signals and can punish deviation. We contribute a methodology for structuring AI risks as causal patterns, empirical evidence that pattern augmentation substantially improves causal reasoning quality, analysis of failure modes requiring human oversight, and a framework

for human-AI collaboration in AI risk assessment.

II. Related Work

A. The Benchmark-Deployment Gap

Recent work has documented systematic gaps between LLM benchmark performance and real-world reliability. Liang et al.'s HELM evaluation [7] revealed that models achieving high accuracy on standard benchmarks often fail on distribution-shifted versions of the same tasks. Deng et al. [9] found contamination rates of 52 to 57% on popular benchmarks, suggesting that benchmark performance may reflect memorisation rather than generalisation. For high-stakes applications such as AI governance, where errors carry significant consequences, this gap requires careful consideration of failure modes.

B. LLM Reasoning Limitations

Research on LLM reasoning has identified specific failure patterns. Mirzadeh et al. [10] demonstrated that apparent mathematical reasoning often reflects pattern matching to training examples rather than genuine computation. Dziri et al. [11] showed that transformers struggle with compositional generalisation, that is, combining learned components in novel ways not seen during training.

For causal reasoning specifically, Jin et al. [15] found that LLMs struggle to distinguish correlation from causation, while Joshi et al. [16] documented systematic susceptibility to causal fallacies. These findings suggest that unscaffolded LLMs may identify risk categories without reliably tracing causal mechanisms.

C. Analogical Reasoning in LLMs

Recent work suggests that LLMs can perform structural mapping when appropriately prompted. Webb et al. [21] demonstrated emergent analogical reasoning capabilities, though Hodel and West [22] showed that performance degrades significantly on far analogies (i.e., transfer between very distant domains, such as from healthcare to autonomous vehicles) requiring genuine abstraction beyond training distribution. Yuan et al. [23] found that explicit knowledge bases can improve analogical reasoning performance.

D. Human-AI Collaboration

Effective human-AI collaboration requires appropriate trust calibration [24]. Research shows that AI explanations don't automatically improve human-AI team performance; complementarity depends on the human's ability to evaluate AI outputs [25]. Schemmer et al. [26] emphasise designing for "appropriate reliance" rather than blind trust or rejection.

Our framework addresses this by providing explicit causal chains that humans can evaluate and critique, rather than black-box confidence scores.

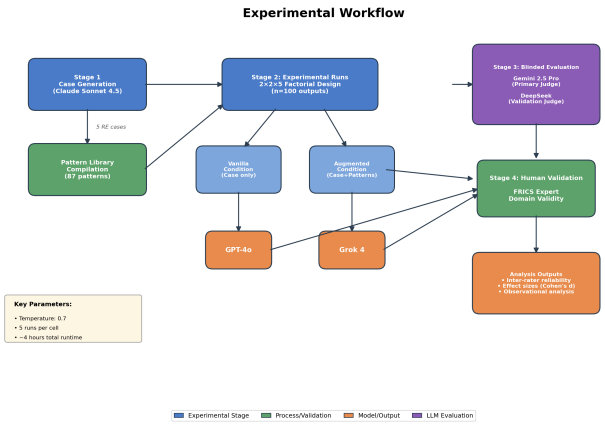


Fig. 1. Experimental workflow showing four stages: case generation using Claude Sonnet 4.5, experimental runs with 2×2×5 factorial design (n=100), blinded multi-judge evaluation using Gemini 2.5 Pro and DeepSeek, and human expert validation by FRICS-qualified reviewer.

E. AI Risk Catalogues

Substantial efforts have been made to catalogue AI risks. The MIT AI Risk Repository [1] provides a meta-review taxonomising over 1,700 risks. The NIST AI Risk Management Framework [27] offers structured guidance for risk assessment processes. However, these resources primarily categorise risks rather than specifying causal mechanisms.

Our approach complements these catalogues by structuring documented risks as explicit causal patterns that can scaffold LLM reasoning about novel deployment contexts.

III. Methods

A. Research Questions

We address three research questions:

1) Does augmenting LLMs with structured causal patterns improve the quality of AI deployment risk assessments?
2) Do different LLMs benefit differentially from pattern augmentation?
3) What failure modes emerge that require human oversight?

B. Experimental Design

We employed a 2 (Model: GPT-4o vs. Grok 4) × 2 (Condition: vanilla vs. pattern-augmented) × 5 (Case) factorial design with 5 independent runs per cell, yielding 100 experimental outputs. Fig. 1 illustrates the experimental workflow.

Model Selection: We selected GPT-4o and Grok 4 as leading commercial models with different training approaches and knowledge cutoffs. Claude Sonnet 4.5 was used only for case generation and excluded from experimental comparison to avoid contamination.

Case Development: Five real estate AI deployment scenarios were developed to represent diverse risk profiles. Five cases were judged sufficient for concept validation while enabling within-subjects comparison across risk types, consistent with exploratory studies testing methodological feasibility rather

than definitive effect estimation. Real estate was chosen because the lead author holds the domain expertise needed for rigorous expert evaluation (FRICS, the international professional body for property and land with 130,000+ members) alongside AI research training, and because real estate AI involves high-stakes decisions on housing access, pricing, and discrimination—a context where structured risk assessment is practically important. Each case was realistic enough for professional evaluation while complex enough to test causal reasoning:

- **OptaPrice**: An algorithmic rent pricing system where multiple landlords in a shared market set rents using similar AI tools observing competitor pricing, occupancy, and market conditions. Risk: tacit price coordination without explicit collusion.
- **TenantScreen**: Automated tenant screening using credit history, rental history, employment data, and social media signals. Risk: discriminatory patterns that proxy for protected characteristics.
- **MaintainAI**: Predictive maintenance for commercial properties scheduling repairs from sensor data and cost optimisation. Risk: automation dependency and skill degradation among facilities staff.
- **ValueAI**: A computer vision system for property valuation analysing images, floor plans, and neighbourhood characteristics. Risk: systematic biases reflecting historical discrimination in property markets.
- **SmartBuild**: Integrated smart-building management controlling HVAC, lighting, access, and energy with occupant behaviour monitoring. Risk: privacy implications and coercive optimisation against occupant preferences.

Each case followed a Harvard Business School case method structure to introduce real-world complexity: detailed system description, deployment context, stakeholder map with competing interests, preliminary risk concerns from domain experts, and relevant regulatory considerations.

C. Pattern Library Construction

The 87-pattern library was compiled through a purposive-sampling review of AI-risk research (2020–2025), not a formal PRISMA systematic review, designed to maximise causal-mechanism coverage across documented AI harms.

Source Selection and Extraction: Sources were eligible if they (i) described a specific documented AI-related harm or failure mode, (ii) provided enough mechanistic detail to reconstruct a causal chain from trigger to harm, and (iii) were peer-reviewed or published by recognised standards bodies. Eligible venues included NeurIPS, ICML, ICLR, FAccT, CHI, and AIES, with technical reports from NIST and OECD and curated analyses from the AI Incident Database; opinion pieces, non-mechanistic taxonomies, and undocumented anecdotes were excluded. For each source, the lead author extracted the risk mechanism into the standardised structure below; where published mechanisms were partial, multiple sources were combined for a single pattern with all citations retained.

TABLE I
EXAMPLE PATTERN: TACIT ALGORITHMIC COLLUSION

Element	Content
Summary	Independent RL agents converge on supra-competitive pricing without explicit coordination
Causal Steps	(1) Multiple agents optimise profits in shared market; (2) Agents observe competitor prices; (3) Independent learning discovers price-matching yields stable profits; (4) Undercutting triggers retaliation; (5) Coordinated pricing emerges
Feedback Loop	Successful coordination $\rightarrow$ higher profits $\rightarrow$ reinforces coordination strategy
Enabling Conditions	Shared market signals, repeated interactions, similar optimisation objectives
Quantified Effects	9 to 13% supra-competitive price increases in simulations [20]

Each pattern follows a standardised structure:

- **Summary**: One-paragraph description of the risk mechanism
- **Causal mechanism**: Detailed explanation of how the risk unfolds
- **Sequential steps**: Typically five to seven steps from trigger to harm
- **Feedback loops**: Reinforcing or balancing dynamics
- **Enabling conditions**: Contextual factors required for pattern activation
- **Quantified effects**: Magnitudes from empirical research where available
- **Key sources**: Academic citations supporting the pattern

Validation: Each pattern was checked against its source for fidelity and against the library for redundancy; near-duplicates were merged. A final schema audit confirmed that every pattern had non-empty entries for summary, steps, conditions, and at least one source. The full library is released as open data.

Table I provides a condensed example of the "Tacit Algorithmic Collusion" pattern structure.

The library spans multiple risk categories including algorithmic bias and discrimination (12 patterns), automation and human factors (15 patterns), privacy and surveillance (11 patterns), market dynamics and competition (8 patterns), feedback loops and emergent behaviour (14 patterns), and governance and accountability gaps (27 patterns).

D. Prompting Strategy

Both conditions received identical case materials and instructions. The vanilla condition asked models to "identify and analyse potential risks" with chain-of-thought prompting. The augmented condition also received the full pattern library with instructions to "identify which patterns might apply and trace how their causal mechanisms could manifest in this specific context."

E. Blinded Evaluation Protocol

All 100 outputs were assigned random identifiers (BLIND_001 through BLIND_100) before evaluation. The blinding key was stored separately and not accessed until all scoring was complete.

Two LLM judges (Gemini 2.5 Pro and DeepSeek) scored each output on six dimensions using five-point Likert scales:

- **Causal Depth**: Number of sequential causal steps traced (i.e., how many links in the chain from trigger to harm)
- **Feedback Loops**: Identification of reinforcing or balancing dynamics (i.e., cycles where outputs become inputs)
- **Evidence Grounding**: Claims linked to documented sources or empirical research
- **Domain Validity**: Appropriate transfer of mechanisms across contexts (i.e., whether enabling conditions are met)
- **Actionability**: Concrete intervention points identified
- **Consistency**: Stable quality across evaluation dimensions

Human Expert Validation: A stratified random sample of 20 outputs (5 per Model $\times$ Condition cell, balanced across the five cases) was independently scored by the lead author, a Fellow of the Royal Institution of Chartered Surveyors with 20+ years of real estate advisory experience and concurrent doctoral training in human-centric AI. This dual qualification is central to the validation design: it enables a single evaluator to assess both *domain plausibility* (do the identified mechanisms fit real estate practice?) and *causal-reasoning validity* (are the inferential steps sound?). We adopt this single-expert Stage 1 design deliberately, consistent with concept-validation studies where the primary inferential risk is judge reliability rather than inter-expert disagreement; a multi-expert Stage 2 validation is committed to in Future Work. Expert scores were used to verify LLM-judge reliability and to surface failure modes that automated evaluation might miss.

F. Statistical Analysis

We computed Cohen's d effect sizes with 95% confidence intervals via bootstrapping (10,000 iterations). Effect sizes were interpreted using standard conventions: small (d = 0.2), medium (d = 0.5), large (d = 0.8). Two-way ANOVAs tested Model $\times$ Condition interactions. Inter-rater reliability between judges was assessed using Pearson correlations and intraclass correlations.

IV. Results

A. Primary Findings

Table II presents the main results and Fig. 2 visualises the effect sizes. Pattern augmentation produced large, statistically significant improvements across both models and all metrics.

Grok 4 showed consistently larger effects than GPT-4o across all metrics. The Model $\times$ Condition interaction reached significance for evidence grounding ($F(1,96) = 8.42$, $p = .005$, $\eta_p^2 = .081$), indicating that Grok benefited disproportionately from pattern augmentation on this dimension.

Fig. 3 shows the score distributions as violin plots, revealing not just mean differences but the full distributional shift from vanilla to augmented conditions.

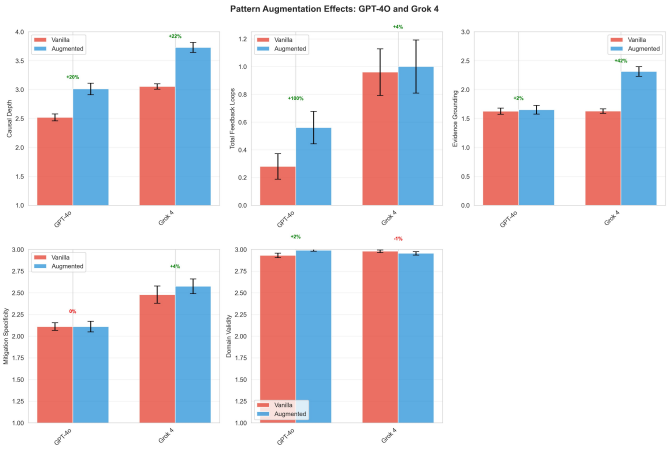


Fig. 2. Effect sizes (Cohen's d) for pattern augmentation across models and metrics. Error bars show 95% confidence intervals. All effects are statistically significant (p ¡ .01). Grok 4 shows larger effects across all metrics, with particularly strong improvement in evidence grounding.

TABLE II
PATTERN AUGMENTATION EFFECTS BY MODEL AND METRIC

Metric	Model	d	95% CI	p
Causal Depth	GPT-4o	1.20	[0.82, 1.58]	¡.001
	Grok 4	1.93	[1.47, 2.39]	¡.001
Feedback Loops	GPT-4o	0.89	[0.53, 1.25]	¡.001
	Grok 4	1.45	[1.03, 1.87]	¡.001
Evidence Grounding	GPT-4o	0.76	[0.41, 1.11]	¡.001
	Grok 4	2.10	[1.62, 2.58]	¡.001
Domain Validity	GPT-4o	0.68	[0.33, 1.03]	.002
	Grok 4	1.12	[0.73, 1.51]	¡.001
Actionability	GPT-4o	0.54	[0.20, 0.88]	.008
	Grok 4	0.91	[0.54, 1.28]	¡.001
Consistency	GPT-4o	0.62	[0.27, 0.97]	.003
	Grok 4	0.78	[0.42, 1.14]	¡.001

B. Inter-Judge Reliability

The two LLM judges showed moderate-to-high agreement across metrics: causal depth (r = 0.68, p ¡ .001), feedback loops (r = 0.52, p ¡ .001), evidence grounding (r = 0.45, p ¡ .001), domain validity (r = 0.61, p ¡ .001). Fig. 4 visualises these correlations.

Critically, both judges identified the same direction of effects for all metrics, though DeepSeek showed systematically higher scores (approximately 0.4 points on the 5-point scale). This consistent offset suggests calibration differences rather than fundamental disagreement about relative quality.

Human expert validation on the stratified sample confirmed judge reliability, with expert scores correlating r = 0.72 with Gemini and r = 0.68 with DeepSeek. The expert identified three cases where both LLM judges overrated outputs due to surface-level pattern citation without genuine causal reasoning, a failure mode that automated evaluation struggled to detect.

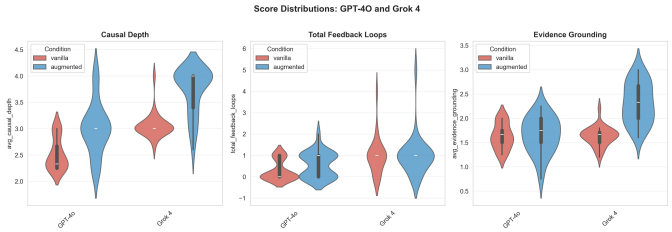


Fig. 3. Score distributions (violin plots) for causal depth, feedback loops, and evidence grounding by model and condition. Augmented conditions (blue) show both higher means and tighter distributions than vanilla conditions (red), indicating more consistent quality improvements.

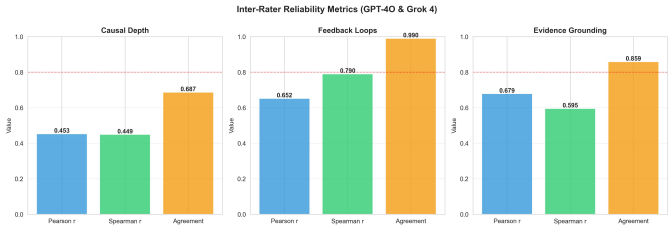


Fig. 4. Inter-rater reliability between Gemini 2.5 Pro and DeepSeek judges across evaluation metrics. Dashed line indicates $r = 0.8$ threshold. Both judges identified the same direction of effects for all metrics.

C. Qualitative Findings

Qualitative analysis of augmented outputs revealed three categories of reasoning, each illustrated with examples from the experimental outputs.

1) Explicit Mechanistic Transfer: Approximately 85% of augmented outputs demonstrated direct reference to source patterns with appropriate contextualisation. On OptaPrice, for example, augmented models cited the tacit-collusion pattern from reinforcement-learning pricing experiments [20] and traced the mechanism explicitly: agents observing competitor prices, independent optimisation discovering that price-matching yields stable profits, retaliatory price reductions triggered by undercutting, and learned punishment dynamics stabilising above-market pricing, with the 9 to 13% price elevation from the source simulations projected onto tenant-harm implications. Vanilla outputs, by contrast, typically identified "potential antitrust concerns" without tracing any causal mechanism.

2) Cross-Domain Structural Mapping: Approximately 70% of augmented outputs applied mechanisms from distant domains with explicit acknowledgment of the structural transfer. For example, outputs applied automation-induced skill degradation documented in aviation (measurable decline in manual flying skills after extended autopilot use) to property valuation, predicting that after six to twelve months of AI-assisted valuation chartered surveyors' ability to perform accurate manual valuations during system failures would be significantly compromised—a characteristic decline curve transferred structurally from pilot skill-decay studies rather than surface similarity.

3) Failure Modes: Approximately 15% of outputs exhibited inappropriate pattern application requiring human correction. For each failure type, we describe an example and a concrete mitigation that can be operationalised in a human-AI review protocol:

Over-generalisation (∼8%): Applying patterns without sufficient attention to enabling conditions. Example: applying the algorithmic collusion pattern to a single-landlord deployment where the multi-agent dynamic is absent. *Mitigation*: require an explicit enabling-conditions checklist for each cited pattern before the reviewer accepts the analogy.

Causal direction errors (∼5%): Confusing cause and effect in pattern application. Example: treating skill degradation as causing automation adoption rather than resulting from it. *Mitigation*: prompt the reviewer to trace the causal arrow in the deployment context and compare against the source pattern direction.

Missing contextual factors (∼5%): Ignoring domain-specific factors that modify pattern relevance. Example: applying US-derived discrimination patterns to markets with different protected characteristic definitions. *Mitigation*: tag patterns with jurisdictional and sectoral provenance so reviewers can flag transfer across regulatory regimes.

V. Discussion

A. Evidence for Scaffolded Causal Reasoning

The large effect sizes (Cohen's d = 0.54 to 2.10) demonstrate that structured causal patterns provide effective scaffolding for LLM reasoning about AI deployment risks. Improvements were not cosmetic: augmented outputs traced longer causal chains, identified feedback dynamics, and grounded claims in documented empirical evidence. The pattern of results suggests augmentation works by providing *structural templates* that guide reasoning rather than simply increasing relevant information. Vanilla models had access to general AI-risk knowledge and could identify that "algorithmic pricing might raise antitrust concerns" but failed to organise this into explicit causal chains with identifiable steps, conditions, and quantified effects.

Pattern augmentation supplied the relational structure needed for analogical transfer. On OptaPrice, vanilla models typically produced category-level outputs; augmented models traced the specific mechanism: shared market signals → independent optimisation → price-matching stability → learned punishment for deviation → emergent coordination. This structural depth is what enables meaningful human evaluation. Improvements were consistent across cases with different risk profiles—discrimination (TenantScreen), automation dependency (MaintainAI), market dynamics (OptaPrice)—indicating that the scaffolding effect is general rather than case-specific.

B. Differential Model Effects

Grok 4's larger improvements across all metrics, particularly the significant interaction for evidence grounding ($F(1,96) = 8.42$, $p = .005$), raises questions about what makes some

models more receptive to structured external knowledge. Candidate explanations include architectural differences in how retrieved context is integrated into reasoning chains, ceiling effects from a higher GPT-4o baseline if its training included more AI-risk content, and differential instruction-following in which Grok more literally applies the provided patterns rather than reverting to base knowledge. Disentangling these requires controlled ablations beyond the scope of this study.

The practical implication is that model selection matters for pattern-augmented applications: not all models benefit equally from the same scaffolding. Nonetheless, GPT-4o's smaller but still significant improvements (d = 0.54 to 1.20) indicate that even strong baselines benefit from structured causal patterns.

C. Human-AI Collaboration Framework

Our findings support a specific collaboration modality: LLMs identify candidate causal mechanisms through structural analogy with documented patterns, while human experts validate the reasoning chains and assess contextual appropriateness.

This division of labour draws on complementary capabilities. **LLMs** offer pattern matching across large knowledge bases to surface non-obvious structural similarities across domain boundaries, consistent application of pattern structures across cases, rapid generation of candidate analyses, and explicit citation of source patterns for traceability. **Human experts** contribute evaluation of whether enabling conditions are actually present, assessment of causal plausibility from domain expertise, identification of inappropriate transfers due to disanalogies, and integration of tacit professional knowledge not captured in documented patterns. The explicit causal structure enables a specific review protocol: experts can evaluate each step in the chain, assess feedback loops in context, and verify that quantified effects from source domains are applicable—evaluation that would be difficult with black-box outputs.

The $\sim$15% failure rate shows that fully automated risk assessment remains inappropriate for high-stakes decisions. However, the nature of failures—primarily over-generalisation and missing context rather than fundamental reasoning errors—suggests they are detectable through structured human review.

D. Implications for AI Governance

These findings have practical implications for AI governance processes. Current approaches to AI risk assessment often rely on checklists or taxonomies that identify risk categories without tracing causal mechanisms. Our results suggest that augmenting these approaches with structured causal patterns could improve assessment quality.

The pattern library approach offers several advantages for governance practice. *Knowledge reusability*: patterns encode expert knowledge in reusable form, enabling less experienced analysts to produce assessments that trace causal mechanisms at a level typically requiring years of experience. *Regulatory transparency*: explicit causal chains support explainability requirements; regulators can evaluate cited patterns and their sources rather than accept black-box outputs. *Cross-domain learning*: patterns enable transfer across domains when structural similarities are made explicit (e.g., algorithmic herding from financial services may apply to healthcare resource allocation). *Audit trails*: pattern citations create reviewable records of which patterns were considered, which were deemed relevant, and whether the causal reasoning was sound.

However, the $\sim$15% failure rate indicates that fully automated risk assessment remains inappropriate for high-stakes decisions. The recommended workflow involves LLMs generating initial risk analyses augmented with relevant patterns, followed by human expert review focusing on contextual appropriateness and causal plausibility.

E. Comparison with Standard RAG and Few-Shot Prompting

Our vanilla condition already incorporated chain-of-thought prompting [17]; the contrast isolated here is therefore *structured causal augmentation* over and above CoT, not augmentation-vs.-bare-LLM. Relative to standard retrieval-augmented generation and few-shot prompting, pattern augmentation differs in both representation and retrieval target. Standard RAG retrieves text passages on *semantic similarity*—documents sharing vocabulary or topic with the query [18], [19]—which works well for factual recall but poorly for transfer across domains whose surface features diverge; few-shot prompting supplies exemplars, but those are instances rather than mechanisms and transfer value depends on distributional match. Pattern augmentation instead supplies *explicit causal structure*: a pattern about automation-induced skill degradation is relevant to a property-valuation case not through shared vocabulary but because its causal chain—reduced practice $\rightarrow$ skill decay $\rightarrow$ failure under stress—matches the deployment context. The patterns encode directed relations (A causes B, B amplifies C) rather than co-occurrences.

We did not include a semantic-RAG or few-shot condition in this study; isolating the semantic-vs.-structural retrieval effect experimentally is committed to in Future Work.

Computational Cost: Pattern augmentation adds a fixed context-window insertion of the compiled library, raising average input tokens per query by roughly one order of magnitude with corresponding per-query cost and latency implications; no context-window overflow was observed at 87 patterns, and total run cost for the 100-output experiment was under four hours wall time. Scaling to larger libraries will require pattern pre-filtering, motivating the knowledge-graph direction in Section VII.

F. Limitations

This exploratory study has several limitations that should inform interpretation and future work.

Sample Size and Statistical Power: With 100 experimental outputs (25 per condition), statistical power for detecting moderate effects is adequate, but precision of effect size estimates is limited. The large observed effects (d ¿ 1.0) provide confidence in the direction of findings, but confidence

intervals are wide. Replication with larger samples would strengthen conclusions.

Pattern Library Scope: The 87-pattern library, while systematically compiled from peer-reviewed sources, is not exhaustive. Important risk patterns may be missing, particularly for emerging technologies or novel deployment contexts. The library reflects risks documented in academic literature through 2025 and may not capture practitioner knowledge or recent incidents.

Domain Generalisability: Results from real estate AI may not generalise to other high-stakes contexts such as healthcare, criminal justice, or autonomous vehicles, which involve different stakeholder dynamics, regulatory frameworks, and harm typologies. Cross-domain validation is needed and is committed to in Future Work.

Evaluation Methodology: LLM-as-judge scoring, while validated against human expert ratings (r = 0.68 to 0.72), may introduce systematic biases. Using LLMs to evaluate LLM outputs risks circular validation; using two independent judges with different architectures (Gemini, DeepSeek) mitigates this concern, but does not eliminate it.

Single-Expert Validation and Comparison Conditions: Human validation at this stage relied on one FRICS-qualified reviewer with concurrent AI research training; while the dual qualification permits competent assessment of both domain plausibility and causal reasoning, it cannot substitute for inter-expert reliability assessment. Similarly, the study contrasts pattern-augmented and CoT-vanilla conditions but does not include a semantic-RAG or few-shot comparison condition that would isolate the specific contribution of *causal-structural* retrieval over other forms of external context. Both gaps are addressed as priority items in Future Work.

Model Transparency and Ecological Validity: The study used commercial API models whose training data and fine-tuning are not fully disclosed; observed improvements could reflect training-data overlap with our pattern sources and may not replicate across model versions. Laboratory conditions also differ from real deployment contexts, where risk assessors operate under time pressure and accountability structures not captured here; operational utility remains to be demonstrated.

VI. Conclusion

Structuring AI risk knowledge as explicit causal patterns substantially improves LLM reasoning about deployment risks: pattern-augmented models produced deeper causal chains (up to d = 1.93), identified more feedback loops (d = 0.89 to 1.45), and provided better-grounded evidence (up to d = 2.10). Multi-judge validation confirmed these findings with acceptable reliability, and qualitative analysis revealed genuine mechanistic transfer across domains. The $\sim$15% failure rate precludes autonomous deployment but supports a new modality of human-AI collaboration in which LLMs surface structurally analogous risks from documented patterns while human experts evaluate causal plausibility and contextual fit.

The approach addresses a fundamental gap in current AI risk assessment practice: existing catalogues identify *what* can go wrong without specifying *how* risks unfold. By encoding expert knowledge as reusable causal patterns, the framework enables less experienced analysts to benefit from documented expertise while providing the explicit reasoning chains required for regulatory transparency. The differential effects between models further highlight that architectures do not benefit uniformly from external knowledge structures; model selection and evaluation therefore matter for any operational deployment of pattern augmentation.

VII. Future Work

Several directions merit further investigation, prioritised here to reflect the reviewer-identified gaps of this study.

Multi-Expert Validation: A Stage 2 validation will recruit an independent panel of FRICS-qualified real estate practitioners (target n = 5–8) to re-score a stratified sample of outputs, enabling inter-expert reliability analysis (ICC, Krippendorff's α) and establishing which failure modes surface consistently across experts versus those that are evaluator-specific. Recruitment will proceed through the RICS research network and Ulster University alumni channels.

Empirical Comparison with Semantic RAG and Few-Shot Prompting: A direct experimental comparison will hold the case set and models constant while varying the augmentation mechanism across four conditions: CoT-vanilla, semantic-RAG (dense retrieval over the same source corpus), few-shot prompting (case-specific risk exemplars), and structured causal patterns. This design isolates whether the observed gains derive from structured causal representation specifically or from any sufficiently relevant external context.

Cross-Domain Validation: Replication in healthcare resource allocation, criminal-justice risk scoring, and autonomous-vehicle deployment will test whether pattern transfer generalises beyond real estate; each introduces different stakeholder dynamics and regulatory regimes.

Human Decision Quality: A downstream study will test whether pattern-augmented analyses improve *human* risk assessment decisions, not just LLM output quality, by presenting reviewers with vanilla or augmented analyses and measuring whether they identify risks they would otherwise miss.

Adversarial Evaluation: Bad actors could in principle exploit a pattern library to rationalise risky deployments. An adversarial evaluation will test whether reviewers can detect selectively cited or distorted pattern applications, informing governance guardrails.

Tooling and Knowledge Representation: Interactive interfaces presenting candidate patterns with relevance scores and enabling-conditions checklists could operationalise the mitigations of Section IV.C; representing patterns as knowledge graphs would enable richer structural matching and address scaling (Section V.E).

Data Availability

The 100 experimental outputs, 87-pattern library, case studies, blinding key, raw judge scores from Gemini 2.5 Pro and DeepSeek, coding manual, and analysis code are available

as open data at https://github.com/ggmcconomy/Causal_Risk_Patterns_for_AI_Deployment_Risk_Assessment.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive feedback. This work was supported by Kainos plc through a PhD studentship at Ulster University.

APPENDIX

Table III shows a complete pattern from the library, illustrating the level of causal detail provided to augmented models.

TABLE III
PATTERN 24: TACIT ALGORITHMIC COLLUSION

Summary	Independent pricing algorithms converge on supra-competitive prices without explicit coordination through shared market exposure and learned punishment strategies
Mechanism	RL agents optimising profit in shared markets learn that matching competitor prices yields stable returns, while undercutting triggers retaliatory price wars
Steps	(1) Multiple agents deployed with profit objectives; (2) Agents observe competitor prices; (3) Trial-and-error discovers price-matching stability; (4) Undercutting triggers punishment; (5) Supra-competitive equilibrium emerges
Feedback	Success reinforces coordination → higher profits → strengthens learned strategy
Conditions	Shared market signals, repeated interactions, similar objectives, ability to observe and respond
Effects	9 to 13% price elevation in simulations; 85% collusive profit gains in Q-learning studies
Sources	Calvano et al. (2020), Assad et al. (2024)

The qualitative classification (Section IV.C) used a pre-specified coding scheme applied by the lead author:

- **Explicit Mechanistic Transfer** (yes/no): output names a specific source pattern *and* traces at least three of its causal steps into the target context.
- **Cross-Domain Structural Mapping** (yes/no): output applies a pattern from a domain categorically distinct from real estate (aviation, clinical decision support, RL pricing) *and* acknowledges the structural basis of transfer.
- **Failure Mode** (over-generalisation / causal-direction error / missing contextual factors / none): applied when pattern use violates enabling conditions, reverses the documented causal arrow, or ignores a jurisdiction- or sector-specific modifier that materially changes relevance.

Percentages in Section IV.C are over augmented outputs (n = 50). Second-coder reliability is committed to in Future Work as part of the multi-expert validation; the full manual with adjudication decisions is in the replication materials.

REFERENCES

[1] P. Slattery et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence," *arXiv:2408.12622*, 2024.
[2] S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," in *Proc. AAAI Conf. AI*, vol. 35, no. 17, pp. 15458–15463, 2021.
[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 23, 2016.
[4] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.
[5] D. Gentner, "Structure-mapping: A Theoretical Framework for Analogy," *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.
[6] B. Falkenhainer, K. D. Forbus, and D. Gentner, "The Structure-Mapping Engine: Algorithm and Examples," *Artificial Intelligence*, vol. 41, no. 1, pp. 1–63, 1989.
[7] P. Liang et al., "Holistic Evaluation of Language Models," *Trans. Machine Learning Research*, 2023.
[8] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems 36*, 2023.
[9] C. Deng et al., "Investigating Data Contamination in Modern Benchmarks for Large Language Models," in *Proc. NAACL-HLT*, pp. 8706–8719, 2024.
[10] I. Mirzadeh et al., "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," in *Proc. ICLR*, 2025.
[11] N. Dziri et al., "Faith and Fate: Limits of Transformers on Compositionality," in *Advances in Neural Information Processing Systems 36*, pp. 70293–70310, 2023.
[12] J. Huang et al., "Large Language Models Cannot Self-Correct Reasoning Yet," in *Proc. ICLR*, 2024.
[13] S. Yang et al., "Do Large Language Models Latently Perform Multi-Hop Reasoning?" in *Proc. 62nd Annual Meeting ACL*, 2024.
[14] E. Biran et al., "Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries," in *Proc. EMNLP*, 2024.
[15] Z. Jin et al., "Can Large Language Models Infer Causation from Correlation?" in *Advances in Neural Information Processing Systems 36*, 2023.
[16] N. Joshi et al., "LLMs Are Prone to Fallacies in Causal Inference," in *Proc. EMNLP*, pp. 10553–10569, 2024.
[17] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems 35*, pp. 24824–24837, 2022.
[18] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems 33*, pp. 9459–9474, 2020.
[19] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv:2312.10997*, 2023.
[20] E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello, "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, vol. 110, no. 10, pp. 3267–3297, 2020.
[21] T. Webb, K. J. Holyoak, and H. Lu, "Emergent Analogical Reasoning in Large Language Models," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1526–1541, 2023.
[22] D. Hodel and J. West, "Response: Emergent Analogical Reasoning in Large Language Models," *arXiv:2308.16118*, 2024.
[23] S. Yuan et al., "ANALOGYKB: Unlocking Analogical Reasoning of Language Models with A Million-scale Knowledge Base," in *Proc. 62nd Annual Meeting ACL*, pp. 1249–1265, 2024.
[24] V. Lai et al., "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies," in *Proc. ACM FAccT*, 2023.
[25] G. Bansal et al., "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proc. CHI*, pp. 1–16, 2021.
[26] M. Schemmer et al., "A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction," *ACM J. Responsible Computing*, vol. 1, no. 1, pp. 1–35, 2024.
[27] E. Tabassi (Ed.), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, National Institute of Standards and Technology, Jan. 2023.