International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 5, Issue 2, PP.173-179, 2024
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I2P117

Preventing Discriminatory Risk Assessment: A Bias Detection Framework for LLM-Powered Insurance Decision Support

Rama Krishna Kumar Lingamgunta
IT Principal, AI center of enablement, Cigna Evernorth Services Inc, Raleigh, North Carolina, USA.

Abstract: The increasing adoption of large language models (LLMs) in insurance underwriting and risk assessment has introduced new forms of algorithmic bias that are not adequately addressed by traditional fairness evaluation techniques. Unlike conventional predictive models, LLM-powered decision support systems reason over unstructured documentation, policy language, and contextual narratives, creating additional pathways for both direct and proxy-based discrimination. In regulated insurance environments, such bias poses significant ethical, legal, and regulatory risks, particularly when AI systems influence high-impact financial decisions. This paper proposes a bias detection framework for LLM-powered insurance decision support systems designed to prevent discriminatory risk assessment while preserving human oversight and auditability. The framework continuously monitors model interactions and decision context to identify bias signals arising from protected attributes, proxy indicators, documentation asymmetry, and inconsistent reasoning patterns. Bias detection is achieved through a combination of prompt instrumentation, contextual feature analysis, counterfactual evaluation, and policy-aligned constraints that operate alongside existing underwriting workflows. Rather than enabling autonomous decision-making, the framework treats LLMs as assistive reasoning components whose outputs are evaluated for fairness risk before informing human judgment. Representative underwriting use cases demonstrate how the framework surfaces biased reasoning, supports corrective intervention, and reduces downstream risk of unfair outcomes. The results indicate improved transparency, bias containment, and regulatory readiness without compromising operational efficiency. While evaluated in an insurance underwriting context, the proposed framework generalizes to other regulated decision domains where generative AI systems influence consequential human decisions.

Keywords: Generative AI, Large Language Models (Llms), Bias Detection, Discriminatory Risk Assessment, Insurance Decision Support, Underwriting Assistants, Ethical AI, Fairness And Transparency, Human-In-The-Loop AI, AI Governance, Regulated Insurance Systems.

1. Introduction

The insurance industry is increasingly exploring the use of large language models (LLMs) to support underwriting and risk assessment workflows. Unlike traditional predictive models, LLM-powered systems are capable of reasoning over unstructured documents, underwriting guidelines, narrative explanations, and contextual signals to assist human underwriters in decision-making. These capabilities promise improvements in efficiency, consistency, and interpretability of complex insurance decisions. As a result, generative AI is rapidly emerging as a decision support layer in underwriting operations rather than a standalone automation tool.

However, the introduction of LLMs into underwriting workflows also introduces new and less visible forms of algorithmic bias. Underwriting decisions directly influence access to insurance products, pricing, and financial risk allocation, making fairness and nondiscrimination foundational requirements. While conventional underwriting models have long been evaluated using statistical fairness metrics, LLM-powered decision support systems operate differently. They synthesize information from heterogeneous sources, infer meaning from language, and generate reasoning narratives, creating additional pathways through which bias can be introduced, amplified, or obscured.

A key challenge arises from the fact that LLMs may unintentionally rely on proxy attributes correlated with protected characteristics [2], such as occupation descriptions, geographic references, documentation completeness, or linguistic patterns in applicant-provided materials. Even when explicit protected attributes are excluded, biased reasoning can emerge through indirect signals embedded in unstructured text. These bias pathways are difficult to detect using traditional fairness testing approaches, which are typically designed for structured inputs and deterministic model outputs.

Regulatory expectations further amplify the importance of bias detection in insurance decision support. Insurance regulators require underwriting decisions to be explainable, auditable, and demonstrably free from unlawful discrimination. The nondeterministic and context-driven nature of LLM outputs complicates compliance with these requirements, particularly when generative systems influence risk interpretation or decision rationale. Without appropriate safeguards, the use of LLMs in

underwriting can expose organizations to legal, ethical, and reputational risk, even when the systems are positioned as assistive rather than autonomous.

Existing research on ethical AI and fairness has largely focused on predictive models and post hoc bias measurement [1]. While these approaches provide valuable insights, they do not fully address the unique challenges posed by LLM-powered decision support systems. In underwriting contexts, bias must be detected not only in outcomes but also in reasoning processes, contextual interpretations, and narrative explanations generated during model interaction. This necessitates new frameworks that operate at the level of decision context and human–AI interaction rather than solely at the level of model prediction.

This paper addresses this gap by proposing a bias detection framework for LLM-powered insurance decision support systems designed to prevent discriminatory risk assessment while preserving human oversight and regulatory compliance. The framework treats LLMs as assistive reasoning components embedded within underwriting workflows and continuously evaluates their outputs for fairness risk before they influence human judgment. By integrating contextual monitoring, counterfactual evaluation, and governance controls, the proposed approach enables early detection and mitigation of bias without replacing existing underwriting authority.

2. Literature Review

Research on fairness and ethical AI in insurance and financial decision-making has expanded significantly in recent years, driven by increased adoption of machine learning models in underwriting, pricing, and risk evaluation. More recently, the emergence of large language models (LLMs) as decision support tools has introduced new challenges that extend beyond the scope of traditional algorithmic fairness research. This section reviews relevant work across four areas: bias in insurance underwriting, fairness approaches in machine learning systems, emerging concerns with generative AI, and limitations of existing bias detection techniques in LLM-powered decision support.

2.1. Bias and Fairness in Insurance Underwriting Systems

Bias in insurance underwriting has long been a subject of regulatory and academic scrutiny. Traditional underwriting models rely on structured variables such as age, credit attributes, claim history, and actuarial risk factors, which are evaluated for disparate impact and compliance with nondiscrimination requirements. Prior research has examined statistical bias, disparate treatment, and disparate impact in underwriting outcomes, leading to the development of governance practices centered on feature selection, model validation, and outcome-based testing.

While these approaches are effective for structured, predictive models, they assume that decision logic is explicit, deterministic, and traceable to a defined set of input variables. As underwriting workflows increasingly incorporate unstructured data and narrative context, these assumptions no longer hold. The shift from purely predictive models to decision support systems that synthesize textual information introduces new bias pathways that are not adequately addressed by traditional underwriting fairness frameworks.

2.2. Fairness and Bias Detection in Machine Learning Models

Extensive research exists on fairness-aware machine learning, including methods for bias measurement, mitigation, and post hoc explanation [1]. Common approaches include demographic parity, equalized odds, and counterfactual fairness, along with techniques such as reweighting, adversarial debiasing, and fairness-constrained optimization. These methods have been successfully applied to classification and regression models in high-stakes domains.

However, most fairness techniques are designed for models that produce numerical predictions or categorical outcomes based on structured features. They do not directly translate to LLM-powered systems that generate free-form text, reasoning narratives, and contextual interpretations. Moreover, fairness metrics applied solely to final decisions may fail to capture bias embedded in intermediate reasoning steps or explanatory content that influences human judgment. This limitation becomes particularly relevant when AI systems are used in assistive roles rather than as final decision-makers.

2.3. Generative AI and Emerging Bias Risks

Recent literature on generative AI highlights concerns related to hallucination, opacity, and unintended bias in language models. LLMs are trained on large-scale corpora that reflect historical and societal biases, which can surface in generated outputs even when explicit protected attributes are excluded. In enterprise and regulated settings, these risks are compounded by the use of LLMs to interpret policies, summarize documents, and provide decision rationale.

Unlike predictive models, LLMs operate through contextual reasoning and language generation, making bias less visible and harder to quantify. Bias may manifest through differential treatment of similar cases, inconsistent interpretation of documentation, or reliance on proxy signals embedded in language. Existing research largely treats generative AI bias as a content moderation or toxicity problem, rather than as a decision support risk in regulated financial workflows. As a result, guidance on how to operationalize bias detection for LLM-powered decision systems remains limited.

2.4. Proxy Attributes and Indirect Discrimination

A particularly challenging aspect of bias in LLM-powered underwriting systems is the use of proxy attributes. Even when protected characteristics such as race, gender, or ethnicity are explicitly excluded, LLMs may infer sensitive information from correlated signals [2], including geographic references, occupational descriptions, linguistic patterns, or documentation completeness. These proxy signals can influence generated reasoning in ways that disadvantage certain groups without explicit discriminatory intent.

Prior work on proxy discrimination has primarily focused on structured features and statistical correlations. In contrast, LLMs may infer proxies dynamically during interaction, based on narrative context and semantic associations. This dynamic inference makes bias detection more complex, as it cannot be fully addressed through static feature audits or training-time controls alone. The literature offers limited practical frameworks for identifying and mitigating proxy-based bias in generative decision support systems.

2.5. Limitations of Existing Ethical AI Frameworks

Ethical AI frameworks proposed in recent years emphasize principles such as fairness, transparency, accountability, and human oversight [3]. While these frameworks provide valuable conceptual guidance, they often lack concrete mechanisms for real-time bias detection and intervention within operational workflows. In practice, ethical considerations are frequently addressed through policy documentation, model review processes, or post-deployment audits rather than continuous monitoring.

For LLM-powered underwriting assistants, static ethical guidelines are insufficient. Bias must be detected during model interaction, evaluated in context, and mitigated before influencing human decisions. Existing literature does not adequately address how ethical principles can be translated into system-level controls that operate alongside live underwriting workflows.

2.6. Research Gap and Positioning of This Work

The reviewed literature reveals a clear gap in current approaches to fairness and bias detection for insurance decision support systems. Traditional underwriting fairness methods are optimized for structured predictive models, while emerging generative AI research focuses primarily on content risks rather than decision impact. Ethical AI frameworks provide high-level guidance but lack operational mechanisms for bias detection in LLM-assisted workflows [3].

This work addresses these limitations by proposing a bias detection framework specifically designed for LLM-powered insurance decision support systems. Rather than evaluating fairness solely at the outcome level, the framework monitors decision context, reasoning patterns, and proxy signal usage during model interaction. By embedding bias detection and mitigation mechanisms directly into underwriting workflows, the proposed approach bridges the gap between ethical principles and operational enforcement, enabling safer adoption of generative AI in regulated insurance environments.

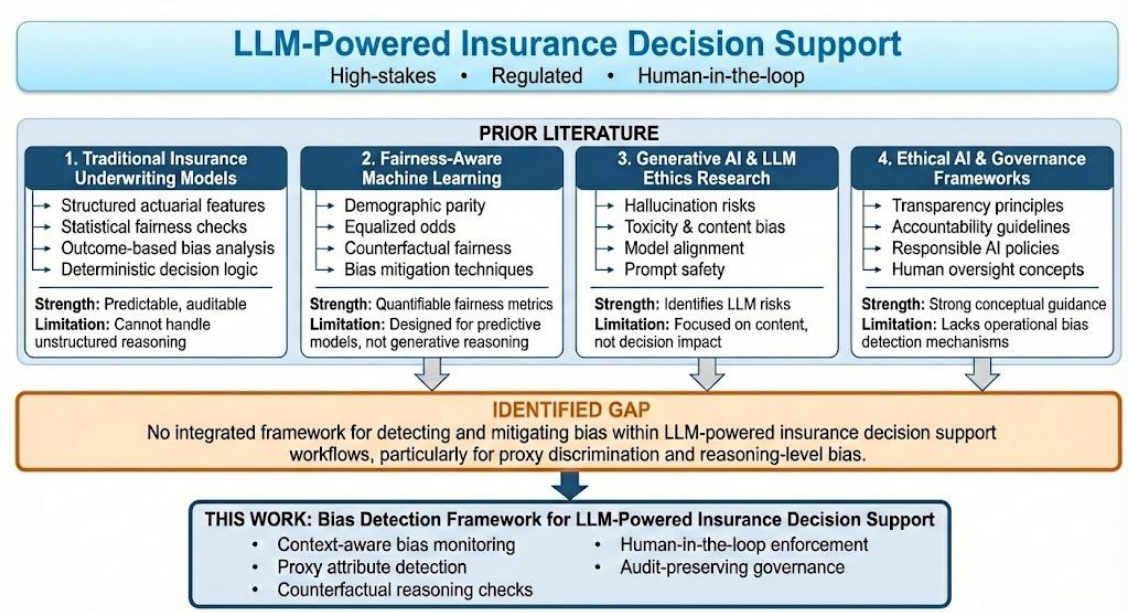

Figure 1: Literature Landscape and Research Gap in Bias Detection for LLM-Powered Insurance Decision Support

3. Methods and Techniques

This section presents the methods, techniques, and system architecture of the proposed bias detection framework for LLM-powered insurance decision support. The framework is designed as a governed, non-intrusive layer that operates alongside

existing underwriting workflows, enabling real-time detection and mitigation of discriminatory reasoning while preserving human authority, regulatory compliance, and auditability. Rather than focusing solely on outcome-level fairness, the framework evaluates bias at the level of generative reasoning and decision context.

3.1. System Architecture Overview

The proposed bias detection framework follows a layered system architecture that separates decision context capture, generative reasoning, bias evaluation, and governance enforcement into distinct components. This separation of concerns ensures transparency, control, and regulatory alignment while allowing each component to evolve independently. The framework does not replace underwriting systems or automate final decisions; instead, it evaluates AI-generated reasoning before it influences human judgment.

As illustrated in Figure X, the architecture consists of six primary components integrated into the underwriting workflow:

1. Decision Context Capture Layer
2. LLM Reasoning Layer (Assistive)
3. Bias Signal Detection Engine
4. Counterfactual and Consistency Evaluation Module
5. Bias Risk Scoring and Classification Layer
6. Human Oversight, Governance, and Audit Layer

All components operate alongside existing underwriting systems without modifying core decision logic or authority.

3.2. Architectural Flow and Component Interaction

The architectural flow begins with the capture of underwriting decision context, including structured applicant attributes, unstructured documents, underwriting guidelines, and workflow metadata. Explicit protected attributes are excluded at ingestion in accordance with regulatory requirements. The captured inputs are normalized into a context representation that preserves semantic relationships while minimizing exposure to sensitive identifiers.

This decision context is passed to the LLM reasoning layer, which generates assistive outputs such as risk summaries, guideline interpretations, and explanatory narratives. These outputs are not directly surfaced to underwriters. Instead, they are routed through the bias detection pipeline, where generated reasoning is evaluated for discriminatory signals. Bias signals identified during evaluation trigger counterfactual and consistency checks. Results from these checks are aggregated into a bias risk score that determines whether AI-generated content can be presented to a human underwriter, requires additional review, or must be suppressed. Only after governance checks are satisfied are AI-generated artifacts exposed to human reviewers, accompanied by bias indicators and explanatory context. All interactions and decisions are recorded in an immutable audit layer.

3.3. Decision Context Capture and Instrumentation

Effective bias detection requires visibility into the full context presented to the LLM. The decision context capture layer aggregates structured underwriting attributes, unstructured applicant documentation, guideline references, and workflow metadata prior to model interaction. Context capture is designed to be non-invasive and operates independently of underwriting execution systems. Captured inputs are transformed into a normalized context representation that enables downstream analysis of semantic relationships, documentation completeness, and contextual cues. This representation supports identification of indirect signals—such as linguistic patterns, geographic references, or occupational descriptors—that may act as proxies for protected characteristics during generative reasoning.

3.4. LLM Reasoning as an Assistive Component

Within the framework, the LLM functions strictly as an assistive reasoning component. It supports tasks such as summarizing applicant information, interpreting underwriting guidelines, and generating explanatory narratives. The model does not generate final risk classifications, pricing decisions, or approval outcomes. Prompting strategies emphasize factual grounding, policy alignment, and explicit uncertainty. LLM outputs are treated as intermediate reasoning artifacts subject to bias evaluation rather than authoritative recommendations. This design reinforces the role of human underwriters as final decision-makers and limits ungoverned model influence.

3.5. Bias Signal Detection Engine

The bias signal detection engine evaluates whether AI-generated reasoning remains stable across underwriting cases that are substantively equivalent from a risk perspective. For each case, the language model produces assistive outputs such as summaries, guideline interpretations, and explanatory narratives. These outputs are captured as a baseline reasoning trace and analyzed for referenced policies, stated risk factors, and explanatory emphasis. The engine then identifies contextual elements that may act as indirect or proxy signals—such as occupational wording, geographic references, documentation completeness, or linguistic framing—without inferring or storing protected attributes.

To assess whether these proxy signals influence reasoning, the engine applies controlled counterfactual evaluation. Selected non-risk-related contextual elements are modified or masked while preserving the underlying risk profile and policy applicability. The model is re-evaluated using these counterfactual contexts, and the resulting reasoning is compared against the baseline. Material differences in perceived risk, guideline interpretation, or confidence that lack underwriting justification are flagged as potential bias signals. Detected signals are classified into discrete bias risk levels with supporting rationale and routed for human review, ensuring transparency, auditability, and regulatory compliance.

3.6. Counterfactual and Consistency Evaluation

To detect subtle or context-dependent bias, the framework applies counterfactual and consistency evaluation techniques. Selected attributes within the decision context are masked, substituted, or perturbed to generate alternative scenarios that preserve underwriting relevance. The LLM is then re-evaluated under these counterfactual contexts. Significant variation in reasoning across comparable scenarios is treated as a potential bias indicator. This approach enables identification of discriminatory patterns that may not be evident from a single model interaction. Counterfactual evaluation is conducted within controlled boundaries to avoid unintended model drift or misuse.

3.7. Bias Risk Scoring and Classification

Outputs from bias signal detection and counterfactual evaluation are aggregated into a bias risk score reflecting severity, confidence, and potential regulatory impact. Bias risk is classified into discrete categories (e.g., low, medium, high) to support actionable intervention without overwhelming reviewers. Risk classification does not imply model failure or misconduct; instead, it serves as a decision-support signal guiding human oversight. This structured approach enables consistent handling of fairness concerns across underwriting teams and workflows.

3.8. Human Oversight, Governance, and Auditability

Human oversight is enforced whenever bias risk exceeds predefined thresholds. Underwriters and compliance reviewers are presented with AI-generated outputs, bias indicators, and explanatory context to support informed judgment. Final authority remains with human reviewers, ensuring accountability and regulatory alignment. All framework activities—including context capture, LLM outputs, bias signals, counterfactual evaluations, and reviewer actions—are recorded in an immutable audit trail. This auditability supports regulatory reporting, post hoc review, and continuous improvement of bias detection mechanisms without compromising operational integrity.

3.9. Generalization across Regulated Decision Domains

Although evaluated in an insurance underwriting context, the proposed methods and architecture are designed to generalize across other regulated decision domains where LLM-powered assistance is used. The modular design allows domain-specific policies, data schemas, and bias heuristics to be substituted while preserving core detection, governance, and audit capabilities. This extensibility supports application in lending, healthcare eligibility determination, and public-sector benefit administration.

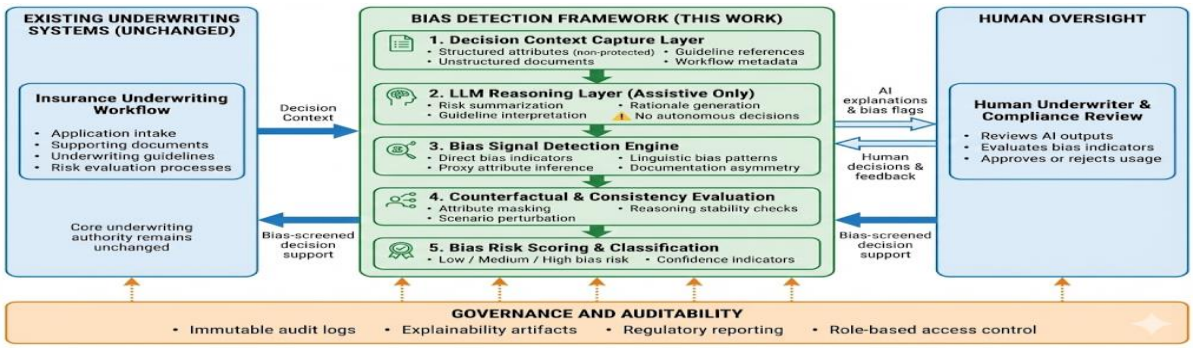

Figure 2: System Architecture of the Bias Detection Framework for LLM-Powered Insurance Decision Support

4. Experimentation and Results

This section describes how the proposed bias detection framework was evaluated and summarizes the observed outcomes. The evaluation focuses on whether the framework can identify discriminatory reasoning risks in LLM-powered insurance decision support before such outputs influence human underwriting judgment. Given the regulated nature of underwriting decisions, experimentation emphasizes reasoning consistency, proxy-based bias detection, and auditability rather than predictive accuracy or automation performance.

4.1. Experimental Setup

The evaluation was designed to reflect realistic use of large language models as assistive underwriting tools, not as autonomous decision-makers. A set of anonymized and synthetic underwriting cases was constructed to represent common

insurance review scenarios. Each case included structured applicant attributes, unstructured documentation such as narrative descriptions or supporting notes, and relevant underwriting guidelines. Explicit protected attributes were excluded from all inputs. Each underwriting case was processed by a large language model configured to generate assistive outputs, including summaries, guideline interpretations, and explanatory reasoning. The model did not produce approvals, denials, pricing decisions, or final risk classifications. All AI-generated outputs were treated as intermediate reasoning artifacts subject to evaluation by the bias detection framework.

4.2. Bias Detection during Evaluation

Bias detection during experimentation followed the reasoning stability and proxy signal analysis approach described in Section 3.5. For each underwriting case, the model's initial reasoning output was captured as a baseline. Equivalent versions of the same case were then created by modifying non-risk-related contextual details—such as occupational wording, geographic references, documentation completeness, or narrative framing—while preserving the underlying risk profile and policy applicability. The model was re-evaluated using these modified contexts, and the resulting reasoning outputs were compared against the baseline. Differences in perceived risk, guideline interpretation, or explanatory emphasis that could not be justified by underwriting policy were flagged as potential bias signals. Flagged cases were classified into discrete bias risk levels and routed for human review with supporting context.

4.3. Evaluation Focus

Evaluation focused on practical questions relevant to fairness and governance in regulated insurance environments:

- Whether AI-generated reasoning remained consistent across substantively equivalent cases
- Whether indirect or proxy-based bias signals could be identified without using protected attributes
- Whether bias indicators and explanations were understandable and actionable for human reviewers
- Whether all AI interactions and review actions were fully traceable for audit and compliance purposes

This approach prioritizes ethical risk containment and regulatory readiness over traditional model performance metrics.

4.4. Observed Outcomes

The evaluation demonstrated that the framework consistently identified scenarios in which AI-generated reasoning varied across cases that were equivalent from a risk perspective. In several instances, minor changes in narrative wording or documentation quality resulted in disproportionate shifts in perceived risk or guideline interpretation. These variations were flagged as elevated bias risk. Human reviewers confirmed that many flagged cases would have been difficult to identify through manual review alone. The presence of structured bias indicators and explanatory context reduced diagnostic effort and improved confidence in fairness assessment. Low-risk cases exhibiting stable reasoning across counterfactual scenarios proceeded with standard oversight, minimizing unnecessary review.

4.5. Auditability and Workflow Impact

All experimental interactions produced complete audit artifacts, including decision context, AI-generated outputs, detected bias signals, counterfactual comparisons, and reviewer actions. These artifacts enabled end-to-end traceability of AI-assisted reasoning and supported post hoc review without reliance on manual reconstruction. No disruption to existing underwriting workflows was observed. The framework operated as a parallel evaluation layer, preserving underwriting authority and regulatory separation of duties while introducing bias-aware controls.

4.6. Summary of Results

Overall, the experimental evaluation indicates that the proposed framework can effectively detect discriminatory reasoning risks in LLM-powered insurance decision support systems. By identifying bias through reasoning instability and proxy signal analysis, the framework improves transparency, supports informed human oversight, and enhances regulatory readiness without introducing autonomous decision-making or operational disruption.

Table 1: Summary of Bias Detection Evaluation Outcomes

Evaluation Aspect	Observed Behavior	Impact on Decision Support
Reasoning consistency	AI-generated reasoning changed when non-risk-related context was modified	Revealed potential proxy-based bias
Proxy signal detection	Occupational wording, geographic references, and documentation quality influenced reasoning	Enabled early identification of indirect discrimination risks
Counterfactual stability	Stable reasoning correlated with low bias risk classification	Supported confident use of AI assistance
Bias risk classification	Cases categorized into low, medium, and high bias risk	Enabled structured and consistent human review
Human review effectiveness	Bias indicators and explanations were clear and actionable	Reduced diagnostic effort and increased reviewer confidence

Auditability	All AI outputs and review actions were fully traceable	Supported compliance and regulatory review
Workflow impact	No disruption to underwriting operations	Confirmed safe integration into existing systems

5. Conclusion and Future Work

This paper presented a bias detection framework for LLM-powered insurance decision support systems designed to prevent discriminatory risk assessment while preserving human oversight, auditability, and regulatory compliance. By focusing on the stability and consistency of generative reasoning rather than solely on decision outcomes, the framework addresses fairness risks that are not adequately captured by traditional bias evaluation approaches. The proposed architecture enables early identification of proxy-based discrimination in AI-assisted underwriting without relying on explicit protected attributes or introducing autonomous decision-making. Experimental evaluation demonstrated that the framework can reliably surface unfair reasoning patterns arising from non-risk-related contextual signals such as narrative wording, documentation quality, and inferred socioeconomic indicators. The use of controlled counterfactual evaluation and structured bias risk classification improved transparency and supported more consistent human review. Importantly, these controls were applied without disrupting existing underwriting workflows, reinforcing the feasibility of deploying bias-aware generative AI in regulated insurance environments.

While the framework was evaluated in the context of insurance underwriting, the underlying principles generalize to other regulated decision domains where LLMs are used for decision support, including lending, healthcare eligibility determination, and public-sector benefit administration. The architectural separation between generative reasoning, bias detection, and governance controls provides a reusable foundation for responsible AI adoption across high-impact decision systems. Future work may explore extending the framework to support longitudinal bias monitoring, enabling detection of fairness drift as models, policies, or population characteristics evolve over time. Additional research could investigate integration with model training and prompt optimization workflows to reduce bias risk upstream, as well as standardized reporting mechanisms to support regulatory transparency. Finally, empirical evaluation across broader datasets and organizational contexts would further validate the framework's applicability and inform best practices for ethical deployment of generative AI in regulated environments.

References

1. Vanama, S. K. R. (2023). Integrating Site Reliability Engineering SRE Principles into Enterprise Architecture for Predictive Resilience. International Journal of Emerging Trends in Computer Science and Information Technology, 4(3), 164-170. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P117
2. Cabello, L., Bugliarello, E., Brandl, S., & Elliott, D. (2023). Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 8465–8483). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.525
3. Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. https://doi.org/10.3390/sci6010003
4. Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2023). A Survey on Fairness in Large Language Models. https://doi.org/10.48550/arXiv.2308.
5. Ubale, A. (2023). Beyond Telematics: Leveraging Generative AI for Synthetic Accident Reconstruction and Liability Attribution in Autonomous Vehicle Claims. International Journal of AI, BigData, Computational and Management Studies, 4(4), 119-124. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P113
6. Deng, S., Zhao, H., Huang, B., Zhang, C., Chen, F., Deng, Y., Yin, J., Dustdar, S., & Zomaya, A. Y. (2023). Cloud-native computing: A survey from the perspective of services. arXiv. https://doi.org/10.48550/arXiv.2306.14402