

Generative AI in Policy-Informed Insurance Adjudication: A Comparative Analysis

Emmanuel Osamau¹ and Chun-Hua Tsai²

College of Information Science & Technology, University of Nebraska Omaha ,
Omaha, 68182, NE, USA

Abstract. This paper addresses how Generative AI-mediated policy framing influences insurance outcomes and the broader adjudication ecosystem. We explore a hybrid decision-making framework that combines an XGBoost model for initial claim likelihood prediction and OpenAI’s GPT-4 for policy application. This approach is tested using two distinct policies: a contextual risk-benefit assessment (CRBA) and an equal needs and fairness evaluation (ENFE). The study specifically analyzes how GenAI’s interpretation and application under these different policy framings impact claim approval rates and associated financial liabilities. Furthermore, we examine the broader implications of integrating this AI-driven framework for multiple stakeholders involved in the adjudication process, including ethical considerations.

Keywords: Machine Learning, Large Language Models, Claims, Policy

1 Introduction

In the traditional decentralized claim structure, insurance claims are typically processed through a hierarchical insurance claims modeling system, where each claim moves through multiple layers of evaluation before a final decision is made [5]. To improve accuracy, efficiency, and reduce costs, the integration of AI enables insurers to process complex and diverse data sources more efficiently, thereby enhancing prediction accuracy and claims handling across the value chain [4]. For instance, machine learning (ML) methods could be used to streamline the insurance claims process, particularly through advancements in natural language processing (NLP) [9, 13]. Over the next twenty years, the insurance sector is anticipated to harness the power of emerging AI technologies, including data from smart connected devices, robotics, and expansive data ecosystems, to substantially improve claims processing efficiency and accuracy [1].

Beyond claims processing, AI also plays a role in shaping insurance policies, as outlined in this survey [10]. One outcome of AI-driven analytics is the identification of potentially fraudulent or unnecessary treatments, which may lead insurers to deny certain claims. While this can reduce fraudulent payouts and potentially increase insurer revenue [8], it can result in billions of dollars in administrative costs to the providers, as well as a significant burden for the patients [12]. AI claims processing extends beyond mere cost reductions [11], with

the advancement of AI and its extension to different components of the claim processing system, it has been argued that the use of AI needs to follow a responsible framework and ethical guidelines to avoid potential risks in identifying claims that should be denied [7].

The ongoing pursuit of operational efficiency driven by technological advancements in automation and AI is fundamentally reshaping industries like insurance. In this highly regulated sector, the ability to make swift, accurate, and compliant decisions for tasks like claims authorization is crucial [9]. These decisions must balance cost-effectiveness with fairness, transparency, and align with complex policy objectives and ethical standards, which eventually impact key stakeholders, including claimants, insurers, and industry regulators. A critical challenge lies in embedding diverse and sometimes conflicting policy goals into automated systems in a way that is both effective and responsible.

Traditional ML models have been instrumental in improving the efficiency of claims management, especially for routine, high-volume claims that follow predictable patterns. Models such as decision trees, logistic regression, and gradient boosting have proven effective in extracting insights from structured claims data to predict outcomes such as fraud likelihood or approval probability. However, systems built on such ML models are designed to fulfill specific tasks, which often fall short when confronted with complex, borderline, or ambiguous cases that require contextual reasoning, interpretation of policies, or ethical considerations that extend beyond the standard numerical and categorical features in a given dataset. This makes the traditional ML method may not be adaptive and aligned with the stakeholders' policy goals.

To overcome the aforementioned limitations, this study introduces and assesses a novel hybrid decision-support framework that leverages the strengths of both traditional ML and Generative AI (Gen-AI). The proposed framework employs XGBoost, a well-established ML algorithm, to identify claims that are likely candidates for approval based on historical trends and patterns. These preliminary predictions are then subjected to a more in-depth, policy-driven analysis by a Large Language Model (LLM), specifically GPT-4, which has been informed by relevant policy framing. By integrating the predictive power of ML with the contextual understanding and interpretive capabilities of General AI, this hybrid approach seeks to provide a more comprehensive and nuanced decision-support system for claims management and policy goals.

Specifically, this paper aims to answer the research question: *How can Gen-AI-mediated policy framing influence insurance claim approval rates and financial liability?* To answer this question, we first apply an XGBoost model to generate initial claim predictions, and then use AI as a final approval mechanism guided by policy. We specifically assess the impact of AI policy review on ML approvals when directed by two distinct policy frameworks: a Contextual Risk-Benefit Assessment (CRBA), prioritizing systemic efficiency and risk minimization, and an Equitable Needs-and-Fairness Evaluation (ENFE), emphasizing fairness and claimant-specific needs. These frameworks were selected because they reflect real-world tensions in insurance adjudication, that is, the

trade-off between actuarial efficiency, which emphasizes minimizing risk, and individualized justice, which centers on the fair and context-sensitive treatment of claimants. CRBA and ENFE thus serve as archetypal representations of two fundamentally different approaches to decision making.

By tasking the LLM to adjudicate insurance claim cases alternately on these two guiding principles, we analyze not only how the chosen paradigm shapes approval decisions and potential financial liabilities but also how it creates implications for the involved stakeholders. Ultimately, this work aims to assess the impact of the distinct AI-mediated policy paradigms and gain critical insights into the implications for multiple stakeholders in the adjudication process.

2 Related Works

2.1 Machine Learning and LLMs in Claims Processing

The insurance claims process is often characterized by large volumes of both structured and unstructured data, and this presents opportunities for AI-driven automation and analysis. Advanced AI applications now allow insurers to predict key characteristics of claims, such as the likelihood of fraud, total loss, or litigation, even at the First Notice of Loss (FNOL). AI systems can segment claims based on their complexity using both known and predicted attributes enabling real-time triage and routing to appropriate processing pathways. These systems can also help determine the most suitable handling procedure for each claim [9]. [13] postulates that in cases where insurance policies and claims share clear textual correlations, AI systems demonstrate strong performance, with high accuracy and logical consistency showing that language models fine-tuned for the insurance domain can effectively learn and apply coverage rules. However, key challenges remain, especially due to the complex language used in certain industries like health insurance.

Studies have demonstrated the utility of leveraging unstructured claims notes, which contain rich information often absent in structured data fields. [3] developed Insurance-Based Language Models (IBLMs) by pre-training models like ULMFiT and BERT on extensive claims notes. These IBLMs showed significant improvement over traditional methods (TFIDF, BiLSTM) in certain classification tasks such as identifying attorney retention, predicting bodily injury, and detecting potential fraud early in the claims lifecycle. Their work highlights the value of domain-specific pre-training and the challenges of handling insurance-specific terminology with varied note quality. [2] provides a detailed examination of the claims process (reporting, adjusting, investigating, etc.) and identifies numerous points where LLMs can add value. Applications include extracting structured information (e.g., incident details, involved parties) from free-text descriptions or call transcripts, performing sentiment analysis on claimant interactions, categorizing claims by peril or severity, identifying inconsistencies for potential fraud investigation, and summarizing complex documents like police or medical reports. This automation aims to support human handlers by providing structured insights and reducing manual workload.

2.2 LLMs for Policy and Regulatory Interpretation

Beyond processing claims data, LLMs are being explored for their ability to interpret and reason based on complex documents like insurance policies and legal regulations. This is particularly relevant for tasks requiring decisions aligned with specific policy constraints. For instance, [6] developed an LLM agent using GPT-4 Turbo and Retrieval-Augmented Generation (RAG) to improve the accessibility of Finnish traffic insurance law. The RAG system retrieves relevant information from selected legal documents to answer user questions, aiming to provide factually grounded responses and mitigate the hallucination issues common with LLMs. While expert evaluation indicated the system could generate satisfactory responses, challenges remained, particularly concerning the optimality and relevance of the retrieved documents. This finding correlates with other research suggesting that improving retrieval mechanisms, possibly through metadata filtering or multi-database categorization, is crucial for RAG effectiveness in specialized domains like law. Related work cited by [6] includes systems like ChatLaw, a multi-agent legal LLM using RAG that demonstrated enhanced reliability, and another Finnish legal assistant for housing disputes where GPT-3.5 with RAG struggled with reliability, suggesting model capability (e.g., GPT-4 vs. GPT-3.5) is also a key factor. [2] also discusses using LLMs in compliance by ingesting regulatory requirements as context and highlights the use of vector databases to overcome context length limitations when dealing with extensive documents like IFRS 17 or Solvency Assessment and Management (SAM) standards. A case study demonstrates querying a regulatory knowledge base built using OpenAI embeddings and a Pinecone vector database to provide domain-specific answers regarding SAM requirements. This approach of augmenting LLMs with external, verified knowledge sources aligns with the goal of making context-sensitive, policy-aligned decisions.

The hybrid approach proposed in this paper is situated within a growing interest in modular AI systems, where different components (e.g., classifiers, language models, search engines) collaborate to achieve more refined outputs. Prior studies, including [3], have experimented with multi-modal and ensemble architectures that combine structured classifiers with unstructured language models. Multi-agent systems and decision-support frameworks where LLMs serve as reasoning modules alongside deterministic predictors have been explored in legal tech and compliance. However, few implementations have explicitly modeled the interaction between a prediction model and a reasoning model operating over policy documents, as is proposed in this work. This sequential architecture, ML prediction followed by LLM-based policy alignment is still underexplored.

3 Method

The adjudication framework consists of two components: (1) a screening stage using an XGBoost classifier trained on structured claim data, and (2) a policy review stage mediated by OpenAI’s GPT-4 model. The goal of this architecture is

to simulate how language embedded within policy documents, when interpreted by Gen-AI models, can influence claim approval rates.

3.1 Data Processing and Feature Engineering

This research utilized an insurance claims dataset sourced from Kaggle. The dataset used in this study consists of 10,293 anonymized insurance policy records, each representing a unique customer and their associated policy details. Each record includes 19 fields capturing a combination of demographic information (e.g., Customer ID, Age, Gender, Marital Status, Occupation, Income Level, Education Level), policy metadata (e.g., Policy Start Date, Policy Renewal Date, Policy Type, Coverage Amount, Premium Amount, Deductible), behavioral and financial indicators (e.g., Claim History, Previous Claims History, Credit Score, Driving Record, Risk Profile), and the target label (Approved) indicating the outcome of a prior claim or policy decision. Prior to modeling, a final hold-out set of 250 records was reserved for downstream evaluation, simulating real-world deployment. The remaining 10,043 records were randomly split into 80% training (8,034 samples) and 20% testing (2,009 samples) sets for validation.

The Policy Duration feature was engineered by calculating the difference in days between the Policy Start Date and Policy Renewal Date. Numerical features such as Age, Income Level, Coverage Amount, and Credit Score were median-imputed and standardized using `StandardScaler`, while categorical variables such as Gender, Marital Status, and Policy Type were imputed with the most frequent value and encoded via one-hot encoding (`OneHotEncoder`). Data imbalance was addressed using Synthetic Minority Over-sampling Technique (SMOTE), incorporated via an `imblearn` pipeline to avoid data leakage during cross-validation.

3.2 Machine Learning–Based Quantitative Screening

The initial screening model is a supervised binary classifier implemented using the `XGBoost` library. Hyperparameters were tuned to balance model complexity and training efficiency, including a depth of 25, a learning rate of 0.01, and early stopping with 50 rounds of patience. The model was trained on 80% of the dataset, stratified by the target variable, and evaluated on the held-out 20% using metrics such as accuracy, AUC-ROC, and a confusion matrix. The complete pipeline, including preprocessing, resampling, and model training, was subsequently saved for reuse. This stage was designed to perform an initial risk-based filtering of claims. Claims predicted as "Denied" by the ML model are excluded from subsequent LLM review, simulating the division between quantitative and interpretative decision thresholds in real-world adjudication systems.

3.3 Large Language Model–Mediated Policy Review

The second adjudication layer uses the GPT-4 model, accessed via the OpenAI Python API, for each ML-approved claim. All API calls were made between April

16 and May 01 2025. The ML-approved claim, together with the full policy document, was passed to the LLM using a zero-shot prompt. GPT-4’s interpretability was tested in an initial evaluation phase by analyzing its reasoning and explanations for each processed claim to ensure it correctly applied the policy language. This setup ensured that the model’s decisions were grounded solely in the policy content, allowing for a focused evaluation of its ability to interpret and apply the guiding principles.

The system prompt defined the LLM as an “Insurance Policy Adjudicator AI” and instructed it to base decisions strictly on the policy language, ignoring any extraneous information. The user prompt embedded both the claim details and the full text of the policy. The LLM was required to respond with a binary decision either “Approved” or “Denied”. A deterministic response was enforced by setting a low temperature (temperature=0.2) and truncating the output (max tokens=100). This ensured consistency and repeatability of decisions. Two separate policy documents were used to evaluate the influence of policy language framing. The ML-LLM pipeline reads the policy text from a local folder to process claims independently against both policies, allowing for a comparison of LLM-mediated approval rates and coverage liabilities. The following prompt was used:

```
messages = [
    {"role": "system", "content": "You are an Insurance Policy Adjudicator
    AI. Your task is to review an insurance claim against the provided policy
    text and make a final decision based *only* on the policy rules and
    the claim details. The claim has already passed an initial quantitative
    check by an ML model."},
    {"role": "user", "content": f""" Insurance Policy: {policy_text (A
    or B)}
    Claim Details: {claim_details_text}
    **Instructions:Review the Claim Details strictly based on the Insurance
    Policy provided above. Your response MUST be one of the following two
    options - Approved or Denied
    Only state the final decision ('Approved' or 'Denied'), do not include
    any additional comments or explanations.""" }
]
```

3.4 Policy Framing

To evaluate the impact of policy framing on LLM-mediated decisions, we implemented two distinct policy paradigms designed to simulate divergent adjudication philosophies: Policy A – Contextual Risk–Benefit Assessment (CRBA): This framing emphasizes a narrative-based evaluation of each claim, encouraging the LLM to weigh interdependent factors (e.g., income, risk profile, driving record) holistically. Claims are approved when the claimant’s overall profile suggests a favorable balance of responsibility and risk, even in the presence of borderline indicators. Policy B – Equitable Needs-and-Fairness Evaluation (ENFE): This approach prioritizes fairness through peer-group benchmarking and proportional

reasoning. The LLM is instructed to assess whether the claimant aligns with approval norms for their demographic cohort and to adjust for aggravating or mitigating factors using a compensatory, sliding-scale logic. The policy text used by the LLM is as follows:

Policy A: Under this framework each claim must be evaluated through a holistic narrative lens rather than a fixed checklist. The adjudicator should interpret the claimant’s profile—age, gender, marital status, occupation, income, education, coverage and premium levels, deductible, policy type, risk profile, credit score, driving record, and policy duration as an interconnected story. Weigh each factor according to its real-world impact: for example, a high-income engineer with a spotless driving history may offset a slightly elevated risk profile, while a lower deductible might signal financial prudence that reduces moral-hazard concerns. Rather than hard cutoffs, apply soft thresholds and allow borderline cases if the overall balance of stability, responsibility, and need suggests the claim can be safely approved. Document the reasoning path, highlighting which elements tipped the scale so that the decision remains transparent and explainable. The adjudication must require weighing each factor according to its demonstrated real-world impact and predictive validity, recognizing that factors often interact in complex ways that simple additive models might miss. Similarly, a claimant’s conscious selection of a lower deductible might signal financial prudence and careful planning that substantially reduces moral-hazard concerns that might otherwise be present.

Policy B: This policy centers on fairness and proportionality ensuring that similar claimants receive comparable outcomes while also recognizing individual circumstances. The adjudicator should first cluster claimants into peer groups (e.g., by occupation bracket, income range, or policy type) and assess whether the incoming profile aligns with the typical approval patterns for that group. Next, adjust for mitigating or aggravating factors such as a credit score well above the group average or a long policy duration without prior claims modify the decision. Rather than strict binary rules, employ a sliding scale approach: exceptional strengths (like a bachelor’s-degree engineer with a 725 credit score and no accidents) can compensate for minor concerns (such as a slightly higher coverage amount), and vice versa. The ultimate goal is to achieve equitable treatment approving claims when the claimant’s overall circumstances justify it, and flagging only those that meaningfully deviate from established fairness norms. This compensatory approach must extend in both directions allowing favorable factors to offset concerns while also permitting unfavorable indicators to counterbalance otherwise positive profiles when the severity warrants such consideration.

These two framings were chosen to reflect contrasting principles commonly found in real-world insurance adjudication: actuarial risk optimization versus equity and procedural fairness.

4 Findings

4.1 Quantitative Classification Performance

The initial machine learning screening stage demonstrated moderate discriminative ability with a test set accuracy of 71.98% and an AUC of 0.7309. This suggests the XGBoost classifier achieved reasonable but not exceptional performance in identifying potentially approvable claims. The model showed better precision for approved claims (0.80) compared to denied claims (0.78), indicating a tendency to be more confident when predicting approvals. Feature importance analysis revealed that *policy type*, *education level*, *marital status*, *occupation*, and *driving record* were among the most influential factors in the ML screening decisions.

Table 1. ML Model Performance Metrics

Metric	Value
Test Set Accuracy	0.72
Test Set AUC	0.73
F1-Score (Denied)	0.74
F1-Score (Approved)	0.78

4.2 Policy Framing Impact on Approval Rates

A significant divergence in approval rates emerged between the two Gen-AI-mediated policy paradigms. Table 2 summarizes the approval outcomes across the adjudication pipeline.

Table 2. Approval Rates by Processing Stage

Stage	Claims	Rate (%)	Coverage (\$)	Mean (\$)	Variance
Total Claims	250	100.0	122,808,288	491,233	276,824
ML Approved	177	70.8	91,327,432	515,974	271,545
Policy A Approved	165	66.0	86,591,068	531,810	272,500
Policy B Approved	142	56.8	75,517,088	524,794	274,887

This represents a 9.2 percentage point difference in approval rates between the two policy frameworks when applied to same claims cases. This finding demonstrates that the linguistic framing of strategic policy guidelines when interpreted by Gen-AI systems can significantly influence adjudication outcomes even when the underlying claim data remains the same.

The differing approval rates translated into substantial financial liability variations. Under Policy A, the total approved coverage was \$86,591,068 compared to \$75,517,088 under Policy B—a difference of \$11,073,980.00 (16.1%). On average this equates to about \$44,296 per differential claim.

4.3 Policy Interpretation by Gen-AI

Policy A’s emphasis on narrative context and holistic evaluation encouraged the model to approve claims where interdependent factors could be interpreted favorably, even when certain individual risk indicators were present. In contrast, Policy B’s focus on peer-group benchmarking and proportional reasoning applied a more stringent standard that resulted in fewer approvals. This differential interpretation reveals that Gen-AI systems could do more than just executing rule-based decision frameworks by engaging with the underlying strategic value systems of insurers which are embedded in policy language. The 12.9% reduction in approvals when moving from Policy A to Policy B indicates that Gen-AI systems may amplify subtle differences in policy framing into significant outcome disparities. This could have diverse implications on stakeholders involved in the claims process especially policyholders.

4.4 ML-LLM Pipeline Dynamics

An important relationship between the quantitative ML screening and qualitative LLM review stages is highlighted as illustrated in Table 3.

Table 3. ML-LLM Pipeline Preservation Rates

Policy	ML LLM Preservation (%)		
Policy A	177	165	93.2
Policy B	177	142	80.2
Difference	0	23	13.0

This confirms that the Gen-AI policy review can have noticeable impact on preservation of ML approvals with Policy B exercising a more critical evaluation of claims that passed the initial ML screening. This two-stage architecture reveals how different components of an AI-mediated adjudication system interact with the LLM review stage serving as either a minor refinement mechanism (under Policy A) or a significant secondary filter (under Policy B) for claims that passed the initial ML screening.

5 Discussion

Our study demonstrates that Gen-AI-mediated policy framing has substantial influence on both insurance claim approval rates and the resulting financial

liability. By embedding two contrasting adjudication philosophies into policy texts—Policy A’s contextual risk–benefit narrative versus Policy B’s equitable needs-and-fairness benchmarking—and passing identical claims through an XG-Boost pre-screening followed by GPT-4–mediated policy review, we observed a 9.2 percentage-point divergence in final approval rates. These findings confirm that LLMs can be developed do not merely apply static rule sets but can be developed to actively interpret and amplify the strategic goals within an insurer’s policy. Under Policy A’s holistic framing, GPT-4 was more inclined to weigh interdependent claimant factors favorably, whereas Policy B’s fairness-oriented framing induced stricter peer-group comparisons and proportional reasoning, resulting in a more conservative approval posture. The two-stage ML–LLM architecture demonstrates how strategic policy framings can fundamentally alter the balance between algorithmic consistency and contextual judgment in hybrid AI decision systems. Our findings reveal profound implications for multiple stakeholders in the insurance ecosystem:

Insurers: The \$11.1M variance in financial liability between policy framings demonstrates that policy language choices translate directly to bottom-line impact. Beyond the immediate financial implications, this illuminates a strategic decision point for insurers: Policy A’s narrative-based approach may enhance customer satisfaction and market differentiation through higher approval rates but carries elevated risk exposure, while Policy B’s fairness-benchmarking orientation offers stronger fiscal controls but potentially at the cost of customer goodwill. Particularly striking is our finding that policy framing disproportionately affected high-value claims (\$44.3K per differential claim), suggesting that policy framing may have its most significant financial impact where traditional underwriting would apply the greatest scrutiny, because of this insurers must recognize that the adoption of Gen-AI adjudication systems does not absolve them of responsibility for outcomes but it only shifts the control from direct claim-by-claim decisions to the upstream design of policy frameworks that guide AI interpretation. This necessitates new governance approaches where policy framing becomes as important as actuarial modeling in determining expected loss ratios.

Policyholders: While the strategic policy framings are beyond the control of the policyholder, the approval rate differential indicates that claimants subject to different policy paradigms may experience fundamentally different outcomes despite identical underlying circumstances. This raises important questions about transparency and predictability, as policyholders may struggle to anticipate adjudication outcomes. For vulnerable policyholders with limited resources to challenge denials, the choice between Policy A and Policy B may represent the difference between receiving needed coverage and facing significant financial hardship. This introduces potential equity concerns if the strategic policy framings of insurers systematically give advantage to certain demographic groups that are better

positioned to navigate the reasoning approach the policy language encourages the AI to adopt.

Healthcare Providers Policy framing in claim approvals has direct financial and operational consequences for healthcare providers. Hospitals may face inconsistent reimbursement rates and administrative burdens depending on whether patients are insured under an insurer with strategic policy guidelines, despite identical clinical scenarios. This unpredictability can affect cash flow and increase appeal volumes for providers. For providers serving vulnerable populations, these effects are amplified and can potentially restrict access to critical care. As such, providers become peripheral stakeholders in the design and governance of AI-mediated insurance adjudication despite having limited influence over the policy decisions shaping these systems as insurers do.

Regulators: To mitigate the gap of transparency from the perspective of policyholders, regulators must be proactive in the use of AI in claims adjudication. Our findings present significant regulatory challenges regarding consistency, fairness, and transparency in AI-mediated insurance decisions. The observed outcome variability based solely on policy framing without changing the underlying claim data, ML algorithm, or LLM suggests that existing regulatory approaches focused primarily on algorithmic transparency may be insufficient to protect policyholders from insurers who may take advantage of this opportunity to pay out fewer claims. Instead, regulators need frameworks that account for how strategic policy framing of insurers interact with AI systems to produce systematically different outcomes. This research also highlights potential risks of regulatory arbitrage, where insurers might adjust policy language to achieve desired approval thresholds while maintaining technical compliance with regulatory guidelines. Standard measurement approaches may be needed to assess the stringency of different policy frameworks and their effects on various claimant groups, particularly those historically marginalized in insurance markets.

6 Conclusions

Building on these findings, future work should pursue hybrid policy frameworks that blend narrative flexibility with fairness-benchmarking to strike an optimal balance between claim accessibility and risk control. A longitudinal impact assessment is also warranted, tracking how policy framing influences claimant satisfaction, renewal behavior, moral hazard, and overall profitability over time. Extending this inquiry across other insurance lines such as health, property, and liability and also into adjacent domains like lending and benefits administration will test the generalization of our results. Taken together, these avenues underscore the necessity of a principled, multi-stakeholder approach to crafting policy language as Gen-AI continues to inform high-stakes decisions, the strategic policy framing of insurers will be as essential as model architecture in aligning AI-driven outcomes with legal, ethical, and business imperatives.

References

- [1] Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. “Insurance 2030—The impact of AI on the future of insurance”. In: *McKinsey & Company* (2018), pp. 1–10.
- [2] Caesar Balona. “ActuaryGPT: applications of large language models to insurance and actuarial work”. In: *British Actuarial Journal* 29 (2024), pp. 1–42. DOI: 10.1017/S1357321724000102.
- [3] Anuj Dimri et al. “Enhancing Claims Handling Processes with Insurance Based Language Models”. In: *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. American Family Insurance. Madison, WI USA, 2019.
- [4] Martin Eling, Davide Nuessle, and Julian Staubli. “The impact of artificial intelligence along the insurance value chain and on the insurability of risks”. In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 47.2 (2022), pp. 205–241.
- [5] Edward W Frees and Emiliano A Valdez. “Hierarchical insurance claims modeling”. In: *Journal of the American Statistical Association* 103.484 (2008), pp. 1457–1469.
- [6] Amir Ingher. “Large Language Model Agent as Insurance Law Assistant”. Supervisor: Senior University Lecturer Arto Hellas, Advisor: Ms Elina Eriksson (MBA). Master’s thesis. Computer, Communication and Information Sciences, Sept. 2024.
- [7] Marina Johnson, Abdullah Albizri, and Antoine Harfouche. “Responsible artificial intelligence in healthcare: Predicting and preventing insurance claim denials for economic and social wellbeing”. In: *Information Systems Frontiers* 25.6 (2023), pp. 2179–2195.
- [8] Neel Mandar Khurjekar. “An integrated three stage predictive framework for health insurance claim denials”. PhD thesis. State University of New York at Binghamton, 2017.
- [9] Ramesh Chandra Aditya Komperla. “Ai-Enhanced Claims Processing: Streamlining Insurance Operations”. In: *Journal of Research Administration* 3.2 (2021), pp. 95–106.
- [10] Olivier Koster, Ruud Kosman, and Joost Visser. “A checklist for explainable AI in the insurance domain”. In: *International Conference on the Quality of Information and Communications Technology*. Springer. 2021, pp. 446–456.
- [11] Ashok Kumar Reddy Sadhu et al. “The Financial Transformation of Healthcare Through AI-Driven Claims Management”. In: *European Economic Letters (EEL)* 13.5 (2023), pp. 2054–64.
- [12] BJ Sanborn. “Change Healthcare analysis shows \$262 billion in medical claims initially denied, meaning billions in administrative costs”. In: *Health Care Finance News* 27 (2017).
- [13] Wen Zhang et al. “AI-powered decision-making in facilitating insurance claim dispute resolution”. In: *Annals of Operations Research* (2023), pp. 1–30.