Beyond the Binary: Gender Bias in LLM-Evaluated Insurance Claims

Fei Huang* Md Mushahidul Islam Shamim† Warut Khern-am-nuai‡
Maxime C. Cohen§

Abstract

Large language models (LLMs) are increasingly deployed in AI-assisted insurance claim processing, a setting in which algorithmic bias may convert modest per-claim disparities into material injustice. We examine whether and under what conditions six LLMs produce differential outcomes for claimants based on gender. Conventional audit methods typically restrict attention to Male/Female comparisons and a single prompt configuration. However, current insurance practice increasingly accommodates gender categories beyond a binary classification. We show that this simplification produces misleading conclusions about model biases. We implement a $4 \times 2 \times 2$ factorial counterfactual design, crossing four gender conditions (Male, Female, Non-binary, and Not Specified) with the presence or absence of the claimant's name and a claim narrative, for 1,388 vehicle claims, yielding 133,248 evaluations on three outcomes (insurance claim amount, claim eligibility, and damage severity). We find that bias is concentrated almost entirely among Non-binary and Not Specified claimants. None of the six models exhibits a statistically significant Male/Female disparity under the full-information condition used in a conventional binary audit, yet three models (GPT-5, GPT-4o, and Gemini 2.5 Pro) exhibit such significant disparities under other prompt configurations. GPT-4o displays the most extensive bias, favoring Not Specified claimants in claim amount by \$128–\$288 when the claimant's name is present, with this pattern switching to a disparity favoring Non-binary claimants once the name is removed. By contrast, three models (Gemini 3 Flash, Claude 4 Sonnet, and Claude 4.5 Sonnet) exhibit no gender disparities across any of the conditions. Our results suggest that organizations cannot rely on binary audits to establish fairness. Instead, audit designs must include Non-binary and Not Specified gender categories and various prompt configurations.

Keywords: Large language models; Algorithmic fairness; Gender bias; Insurance claim processing; Counterfactual audit; AI governance

*Associate Professor of Risk and Actuarial Studies, UNSW Business School, UNSW Sydney. Email: feihuang@unsw.edu.au

†Research Assistant, UNSW Business School; PhD Candidate, Faculty of Medicine and Health, UNSW Sydney. Email: m.shamim@unsw.edu.au

‡Associate Professor, Desautels Faculty of Management, McGill University, Montréal, QC, Canada. Email: warut.khern-am-nuai@mcgill.ca

§Professor, Desautels Faculty of Management, McGill University, Montréal, QC, Canada. Email: maxime.cohen@mcgill.ca

1 Introduction

Large language models (LLMs) have increasingly been embedded in decision-making processes in organizations (Wang et al., 2026). Many industries no longer use LLMs to just summarize information or help human workers automate tedious tasks. Instead, LLMs are used to interpret complex multi-modal inputs, recommend decisions, or even shape dynamic allocation of economic resources (Leng et al., 2026). This delegation is particularly salient in the insurance industry, where human adjusters are often overwhelmed by the increasing volume of insurance claims (Badiel and Dionne, 2026). For example, a mid-size auto insurer in the United States typically processes more than 165,000 claims annually (Cahalan and Hamer, 2026). In addition, experiencing delays in insurance claims is one of the most common reasons prompting customers to switch insurance providers (Brown, 2025). Consequently, the industry has increasingly adopted LLMs to help process insurance claims in the hope to increase both operational performance and consumer satisfaction, and ultimately reduce human workers' workload and overhead costs (Balona, 2024).

On the one hand, LLMs appear to be an excellent solution for assessing and adjudicating insurance claims because the models are highly capable of processing multi-modal inputs. For example, accident photographs, vehicle metadata, and textual claim narratives can be easily integrated into claim workflows at a speed that outpaces human adjusters (Bhattacharya et al., 2025; Winder et al., 2025). On the other hand, potential LLM biases are well documented in several contexts (e.g., Cohen et al., 2025; Tanlamai et al., 2026), although the evidence in the insurance context is currently limited. Nevertheless, this issue is particularly important because relying on a potentially biased LLM for claim adjudication would systematically misdistribute millions of dollars in insurance payout. More importantly, unlike human adjusters whose individual biases may sporadically affect a limited number of consumers, a biased LLM would consistently apply its bias to every claim it processes. To address this issue, we propose the following research question:

RQ: *What are the gender biases that emerge when LLMs are used to process insurance claims?*

The insurance context also illustrates a broader challenge in organizational AI governance. Unlike traditional predictive models, LLMs are general-purpose, prompt-sensitive, and often proprietary. Their decisions are shaped not only by the model architecture and the training data but

also by how the information is presented at inference time. Small changes in prompt configuration, such as whether the claimant's name is included or whether a narrative description accompanies an image, may alter how the model uses demographic signals. As a result, in the insurance context, bias cannot be treated as a model-only issue. Rather, it emerges from the interaction between the model, the prompt, the focal task, and the organizational workflow in which the system is deployed. In addition, existing approaches to algorithmic bias auditing are often not suitable for the insurance context in two aspects. First, traditional LLM audits usually focus on binary Male/Female comparisons (Winder et al., 2025; Lin et al., 2025). This design reflects long-standing conventions in bias and discrimination auditing, but it is increasingly incomplete in the current societal environment, where organizations routinely record Non-binary identities, allow individuals not to specify gender, or infer demographic categories from incomplete data. Second, existing studies often evaluate a model under a single prompt configuration, usually a full-information condition. This practice implicitly assumes that a model's fairness profile is stable across information environments. For LLMs, however, this assumption is not always realistic because the models are often sensitive to contextual cues. Thus, an audit conducted under one prompt structure may not identify bias that will appear under another structure used in practice.

Insurance companies have long employed risk-based algorithmic pricing models, adjusting premiums based on factors such as age, lifestyle, personal history, and geographic location. However, concerns have emerged regarding the use of variables that may serve as proxies for protected characteristics such as gender, potentially leading to discriminatory outcomes. In response to these concerns, several jurisdictions have introduced regulatory restrictions on the use of demographic variables in insurance pricing. For example, California banned the use of gender as a factor in auto insurance pricing in 2019 by adopting a unisex pricing model, while Michigan prohibited the use of gender as a factor in auto insurance pricing in 2020. Despite these efforts, disparities persist. A 2026 study by The Zebra (Meyer, 2026) found that men pay more for car insurance than women in 38 states, highlighting ongoing inconsistencies in pricing practices. Similarly, the European Union banned the use of gender as a pricing variable for several personal insurance products, including annuities, life insurance, and motor insurance, with the goal of eliminating gender-based disparities in insurance pricing. These developments highlight the importance for insurance companies to critically assess their pricing models and ensure that they do not inadvertently introduce dispar-

ities. As regulatory scrutiny intensifies and public awareness grows, the increasing use of LLMs has renewed attention to this important issue and underscores the need for rigorous investigation. These pricing rules illustrate why demographic variables in insurance decisions attract regulatory and public scrutiny. The present study examines a related but distinct question, namely, whether LLM-assisted claim workflows vary eligibility, severity, or payout across gender-field values at the claim-adjudication stage.

A key challenge in empirically evaluating potential biases in LLM-assisted insurance claim processing is the absence of an objective ground truth against which model decisions can be compared. In practice, insurance claim outcomes are typically determined by the judgment of a single adjuster, making observational claim data inherently susceptible to human bias. Consequently, the literature has cautioned against treating historical claim decisions as ground truth for evaluating algorithmic fairness. To address this challenge, we follow the emerging methodology adopted by recent studies (Cohen et al., 2025; Tanlamai et al., 2026), using LLMs as the object of inquiry within a controlled counterfactual experimental design rather than benchmarking them against historical human decisions. Specifically, we conduct a large-scale counterfactual audit of gender bias in LLM-evaluated vehicle insurance claims. We construct a $4 \times 2 \times 2$ factorial design that varies three features of the prompt: claimant gender category, claimant's name inclusion, and claim narrative inclusion. The four gender conditions are Male, Female, Non-binary, and Not Specified.[1] The two name conditions indicate whether the claimant's name is included or not. Finally, the two narrative conditions indicate whether an AI-generated claim description accompanies the accident image and vehicle metadata. We apply this design to 1,388 vehicle insurance claims, six LLMs, and three outcomes: predicted claim amount, claim eligibility, and damage severity. The resulting design produces a total of 133,248 model evaluations.

Binary audits are insufficient. Our findings show that conventional audit designs can substantially mischaracterize the fairness of LLM-based decision systems. Specifically, we find that bias is concentrated almost entirely among Non-binary and Not Specified claimants rather than in

[1]We acknowledge that Male, Female, and Non-binary denote gender identities, whereas Not Specified represents a disclosure state. We nevertheless include it because our objective is to evaluate bias under practical data schemas rather than to study gender identity alone. Many insurance systems allow consumers to select 'Not Specified" within the same gender field used operationally during registration. Consequently, our factorial design evaluates the full set of gender-field values that an insurer may encounter in practice. We thus interpret findings involving Not Specified as evidence regarding the treatment of undisclosed gender information rather than as evidence about a gender identity.

Male/Female comparisons. Under the full-information condition, none of the six models exhibits a significant Male/Female disparity and would thus appear fair under a conventional binary audit. Yet, three models, GPT-5, GPT-4o, and Gemini 2.5 Pro, exhibit significant disparities involving Non-binary or Not Specified claimants under other prompt configurations. Consequently, a binary audit may provide a false sense of fairness by certifying a model as unbiased precisely because it excludes the gender categories in which bias is most likely to emerge.

Bias is context-dependent. We also find that, for the model exhibiting the most extensive bias pattern, bias is highly condition-dependent. For GPT-4o, the presence or absence of the claimant's name fundamentally changes the nature of the observed disparity. When the claimant's name is included, the model consistently favors Not Specified claimants in predicted claim amounts. Once the name is removed, however, this pattern disappears and is replaced by a different disparity favoring Non-binary over Male claimants. These findings suggest that demographic signals do not influence LLM outputs in a straightforward manner. Rather, their effects depend critically on the broader contextual information surrounding them.

Fairness is not static across model versions. Finally, three of the six models we evaluate (Gemini 3 Flash, Claude 4 Sonnet, and Claude 4.5 Sonnet) exhibit no statistically significant disparities across any of the 16 experimental conditions. We also find that model updates can improve fairness. GPT-5, for example, exhibits only one significant finding compared with seven for its predecessor, GPT-4o. At the same time, organizations should not assume that model updates are fairness-neutral. A fairness assessment conducted at procurement cannot be assumed to remain valid following a version update, as model behavior may change in unpredictable ways. Accordingly, every model version update should be treated as a governance event requiring a renewed, multiplicity-aware fairness audit rather than as a routine technical upgrade.

This study makes three contributions to the Information Systems (IS) literature. First, we contribute to the literature on algorithmic bias and fairness in AI-mediated decision-making by showing that LLM bias is not adequately captured by binary demographic comparisons. In our setting, we specifically find that the most consequential disparities arise for Non-binary and Not Specified claimants, both of which are often excluded from traditional audit designs. This finding highlights the risk of designing audits based on conventional practice rather than modern practices commonly adopted among organizations. Second, we contribute to research on generative AI

governance by demonstrating that fairness is prompt-contingent. For LLM-based decision systems, an audit must evaluate not only the model but also the information environment in which the model operates. The prompt configuration becomes part of the governance object. This insight extends current research on AI implementation by showing that organizational choices about data fields, interface design, workflow integration, and prompt templates can materially affect demographic outcomes. Third, our factorial counterfactual audit design contributes to the literature on LLM auditing by demonstrating a viable avenue to audit proprietary closed AI systems. Since frontier LLMs are typically black boxes, researchers generally have limited capabilities to inspect model weights, training data, or internal reasoning processes. Our design treats controlled variation in the prompt structure as a behavioral explanation toolkit. By comparing the model outputs across conditions that differ in exactly one structural characteristic, we can infer how demographic signals and contextual information jointly shape decisions even without direct access to model inferences.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the data and our experimental design. Section 4 presents the empirical findings. Section 5 develops the framework for behavioral explainability and proposes some underlying mechanisms. Finally, Section 6 discusses the practical, managerial, and policy implications as well as our conclusions.

2 Literature Review

2.1 AI-Assisted Applications and Insurance Market

The deployment of AI in business operations has generated a substantial body of research examining how algorithmic AI systems affect service quality, efficiency, and customer outcomes. For example, Cui et al. (2022) demonstrate that automation and recommendation intelligence interact in procurement settings, with full value requiring both components simultaneously. DosSantos DiSorbo et al. (2025) show that human-AI collaboration in forecasting is disrupted when AI outputs include outliers, with participants unable to appropriately calibrate their adjustments. Celdir et al. (2024) establish that recommendation algorithms in matching platforms produce systematic popularity bias, disadvantaging less-visible users in ways that are invisible to platform designers. This finding is structurally analogous to the gender biases we document in insurance claim processing. Bai et al.

(2022) conduct field experiments and show that algorithmic task assignments alter workers' fairness perceptions relative to human-based assignments, with downstream consequences for productivity. Their finding that the fairness consequences of algorithmic systems depend on how the algorithm is presented and framed is directly relevant to our prompt-configuration results. This body of work establishes that the operational consequences of AI deployment depend critically on the specific interaction between the algorithmic system and the information environment in which it operates, a principle that our factorial design extends to the fairness domain.

These findings collectively illustrate that AI-mediated service delivery is not simply a passive substitution of algorithmic computation for human judgment, but rather a delegation of decision-making authority under uncertainty. Baird and Maruping (2021) formalize this shift by arguing that IS research should treat AI-based systems as agentic artifacts capable of assuming responsibility for task outcomes rather than as passive tools awaiting human direction. We adopt this perspective throughout the paper. Specifically, we study an LLM prompted to act as a claims adjudicator, interpreting multimodal evidence and recommending financial outcomes as a decision-support system whose deployment with such authority would raise governance concerns. From this perspective, auditing its demographic sensitivity becomes a matter of governance rather than merely a technical exercise.

The topic of price discrimination and fairness has been extensively studied in the literature (e.g., Cohen et al., 2022). More recently, the increasing adoption of LLMs for pricing decisions (Cohen, 2026) has raised important questions about whether LLM-generated price recommendations may be biased and lead to discriminatory outcomes. This question has previously been examined in the contexts of real estate pricing (Tanlamai et al., 2026) and labor wage recommendations (**?**), whereas the present study focuses on insurance claims.

In the insurance industry, AI adoption spans underwriting, fraud detection, pricing, and claim processing. Bhattacharya et al. (2025) provide a systematic review of AI applications across automotive, health, and property insurance, documenting the transition from traditional actuarial models to deep learning and LLM-based systems. Lin et al. (2025) construct INS-MMBench, a multimodal benchmark for evaluating LLMs in insurance tasks, and establish baseline performance profiles for frontier models in visually grounded claim evaluation. Research on fairness and AI in insurance has primarily examined traditional machine learning models. Xin and Huang (2024)

develop antidiscrimination pricing frameworks for actuarial models covering group fairness criteria and regulatory implications, establishing the conceptual foundations for algorithmic fairness in insurance. The question of whether LLM-based multimodal claim processing systems exhibit gender bias in financial valuations remains largely unaddressed.

2.2 Algorithmic Bias and Audit Methodology

Demographic bias in AI systems has been documented across a range of high-stakes domains, with both the presence and magnitude of bias depending critically on the application context. In hiring, An et al. (2025) evaluate five major LLMs on approximately 361,000 randomly generated resumes, finding a consistent preference for female candidates in assessment scores alongside penalties for Black male candidates. Similarly, Gaebler et al. (2024) employ a correspondence experiment to evaluate LLM-based candidate ratings and identify moderate racial and gender disparities, which they attribute to post-training alignment procedures. In the context of wage recommendations, **?** audit eight LLMs using 480,000 freelancer profiles and find no evidence of gender discrimination, but document substantial geographic and age disparities. This finding, which emerges from the same methodological tradition as the present study, highlights the context-dependent nature of LLM bias. In contrast to their wage-recommendation audit, which finds no gender discrimination, our insurance-claim audit detects gender-related disparities for some models and prompt configurations.

Two closely related theoretical contributions provide the conceptual foundation for our work. Fu et al. (2022) demonstrate that fairness-constrained machine learning algorithms can produce unintended welfare consequences by redistributing surplus across demographic groups in ways that system designers do not anticipate. We extend this insight to the prompt-configuration dimension, showing that fairness under one prompting condition does not necessarily imply fairness under another. Complementing this perspective, Hu et al. (2026) empirically trace how human annotation bias enters machine learning pipelines through training data and evolves across model iterations in high-stakes microlending settings, providing empirical support for our discussion of RLHF-based bias mechanisms in Section 5. Importantly, they find that even fairness-unaware machine learning algorithms can reduce biases present in human decision-making, suggesting that training pipelines moderate rather than simply inherit human biases. This interpretation is consistent with the pronounced heterogeneity we observe across models. Five of the six models we evaluate exhibit

little or no significant gender bias, whereas GPT-4o displays a substantial and systematic pattern of disparities. This contrast suggests that the emergence of human-like demographic associations depends less on an inherent property of LLMs than on differences in model development and post-training pipelines across providers.

Methodologically, counterfactual audit designs have become the standard approach for isolating the causal effects of demographic signals in AI systems. Gaebler et al. (2024) adapt the correspondence experiment methodology from empirical economics to the auditing of LLMs, while Winder et al. (2025) apply this approach to LLM-based insurance claim adjudication, providing the closest prior work to our study. However, both studies rely on binary Male–Female comparisons within a single prompt configuration. We show that this approach can mischaracterize the bias profile of half of the models in our evaluation because both the direction and magnitude of bias depend on the gender categories being tested and on the information included in the prompt. Our factorial design addresses this limitation by treating the prompt configuration as an experimental factor rather than an unexamined constant, while simultaneously providing a form of behavioral explanation for closed AI systems. Because the models we study are proprietary systems accessible only through inference APIs, conventional interpretability methods are unavailable (Wachter et al., 2018). Instead, each pair of experimental conditions differing along exactly one structural dimension constitutes a controlled counterfactual, whose contrast in outputs reveals model sensitivity without requiring access to internal model parameters or architecture (Mothilal et al., 2020). This reframes the factorial audit not merely as a bias detection tool, but also as an analytical instrument for understanding the behavior of closed AI systems in operational settings.

2.3 Algorithmic Fairness and Bias

A parallel and directly relevant stream of IS research conceptualizes algorithmic bias as an organizational and design challenge rather than merely a statistical property of a model. Shimao et al. (2025) propose a framework for studying the fairness of AI models that simultaneously examines prediction outcomes and behavioral differences between majority and minority groups. Kordzadeh and Ghasemaghaei (2022) synthesize the IS literature on algorithmic bias and show that empirical audits of demographic bias in deployed AI systems remain relatively rare compared with conceptual and design-oriented research. They call for studies that examine bias across a broader range

of demographic categories and deployment conditions—a gap that our factorial design directly addresses. Similarly, Nadeem et al. (2022) conduct a systematic review of gender bias in AI-based decision-making systems and find that the existing literature overwhelmingly treats gender as a binary construct, precisely mirroring the auditing convention that we show can produce materially misleading fairness conclusions in the insurance claim setting. Teodorescu et al. (2021) further argue that automated approaches to enforcing fairness constraints in machine learning systems often fail as decision contexts become more complex, advocating instead for a deeper understanding of how human-machine learning collaboration shapes fairness outcomes. Consistent with this perspective, we find that the detected disparities are configuration-specific rather than monotonic in prompt richness, with the sharpest disparities concentrated in the name-present, narrative-absent configuration rather than uniformly increasing as more information is added. Most recently, Rai et al. (2026) develop a design theory of AI fairness, arguing that fairness must be deliberately engineered throughout a system's development and deployment lifecycle rather than evaluated only after deployment. Our factorial audit operationalizes this design-theoretic perspective by treating prompt configuration itself as a design parameter subject to fairness evaluation, thereby extending design-oriented fairness research to the auditing of proprietary closed multimodal LLMs.

2.4 Generative AI and AI Governance in Organizations

Our results also contribute to the emerging IS literature on the governance of generative AI in organizations. Berente et al. (2021) argue that governing AI systems requires organizations to contend with three interdependent characteristics, autonomy, learning capacity, and inscrutability, each of which challenges traditional IT governance mechanisms. LLM-based claim adjudication embodies all three: the model autonomously interprets multimodal evidence, its behavior evolves across model versions, and its internal reasoning remains inaccessible to the deploying organization. Similarly, Susarla et al. (2023) caution that the rapid organizational adoption of generative AI has a dual-edged, or "Janus-faced," character, creating substantial value while simultaneously introducing risks that require deliberate and responsible governance rather than passive adoption. Our finding that GPT-4o exhibits a substantial and systematic bias favoring Not Specified claimants, a pattern that would have gone undetected under a conventional binary Male/Female audit, provides a concrete illustration of this risk in an operational setting. Likewise, Feuerriegel et al. (2024)

conceptualize generative AI as a socio-technical artifact whose behavior is jointly shaped by model design and the organizational and informational context in which it is deployed. This perspective is consistent with our finding that gender bias is not an inherent property of a model, but rather emerges from the interaction between the model and its prompt configuration. Finally, Dwivedi et al. (2023) bring together multidisciplinary perspectives on the organizational and policy implications of generative conversational AI, calling for research that moves beyond benchmarking model capabilities toward the responsible and context-sensitive deployment of these systems. Our audit design responds directly to this call by providing a framework for evaluating the fairness of proprietary multimodal LLMs in the context of insurance claim adjudication.

3 Research Design

3.1 Dataset

Our primary dataset is drawn from the "Fast, Furious and Insured" vehicle insurance dataset (HackerEarth, 2024), previously used by Lin et al. (2025) in a different context. The dataset comprises 1,388 vehicle insurance claims covering cars, trucks, motorcycles, and other road vehicles. Each claim consists of an accident photograph accompanied by metadata, including the vehicle's cost, minimum and maximum coverage limits, insurance expiration date, and vehicle condition.

For each claim, a standardized claim narrative was generated using GPT-4o, which processed both the accident image and the associated vehicle metadata to produce a description of the accident scenario. Each narrative was generated once for each claim and, in the narrative-present conditions, the same narrative was held fixed across all gender prompt variants. Importantly, the narratives contained no demographic markers, ensuring that demographic variation was introduced exclusively through the controlled prompt manipulations described below.

The resulting experimental dataset, constructed as the INSURBIAS benchmark (Huang et al., 2026), is publicly available at `https://huggingface.co/datasets/feihuangfh/INSURBIAS` and includes accident images, AI-generated claim narratives, counterfactual prompt pairs, metadata, and model predictions for all experimental conditions.

3.2 Factorial Experimental Design

Our $4 \times 2 \times 2$ factorial design varies three structural dimensions of the prompt, including the gender category (four levels), the inclusion of the claimant's name (two levels), and the presence of the claim narrative (two levels).

Factor 1: Gender. Each claim is evaluated under four gender conditions:

(1) Male (M) name "James," gender field "Male",

(2) Female (F) name "Mary," gender field "Female",

(3) Non-binary (NB) name "Alex," gender field "Non-binary",

(4) Not Specified (NS) name "Taylor," gender field "Not Specified".

The names were selected to represent common English names that are characteristically male, female, or gender-neutral. The combination of a gendered name and an explicit gender field provides an unambiguous demographic signal, making the gender manipulation both ecologically valid (because insurance profiles typically include both pieces of information) and methodologically rigorous by ensuring that any observed disparities cannot be attributed to uncertainty in gender inference.

Factor 2: Claimant's name. In the name-present conditions, the Basic Information block includes both the claimant's name and the gender field specified above. In the name-absent conditions, the claimant's name is omitted, leaving only the gender field and vehicle metadata. This factor tests whether the claimant's name and the explicit gender field convey redundant or independent demographic signals, and whether the interaction between these signals varies across models.

Factor 3: Claim narrative. In the narrative-present conditions, the AI-generated claim description is included in the prompt. In the narrative-absent conditions, the model receives only the accident image, claimant information (gender/name), and vehicle metadata, without any accompanying descriptive text. This factor tests whether the presence of contextual information moderates the influence of demographic signals on the model's decisions.

Condition groups. The $4 \times 2 \times 2$ design yields 16 conditions per model, organized into four groups based on name and narrative presence:

- **Group 1 (C1–C4):** Name present, narrative present
- **Group 2 (C5–C8):** Name present, narrative absent
- **Group 3 (C9–C12):** Name absent, narrative present
- **Group 4 (C13–C16):** Name absent, narrative absent (minimal condition)

Figure 1 illustrates the experimental design. Each condition is applied to all 1,388 claims, leading to $1{,}388 \times 16 = 22{,}208$ evaluations per model, hence a total of 133,248 evaluations for all six models.

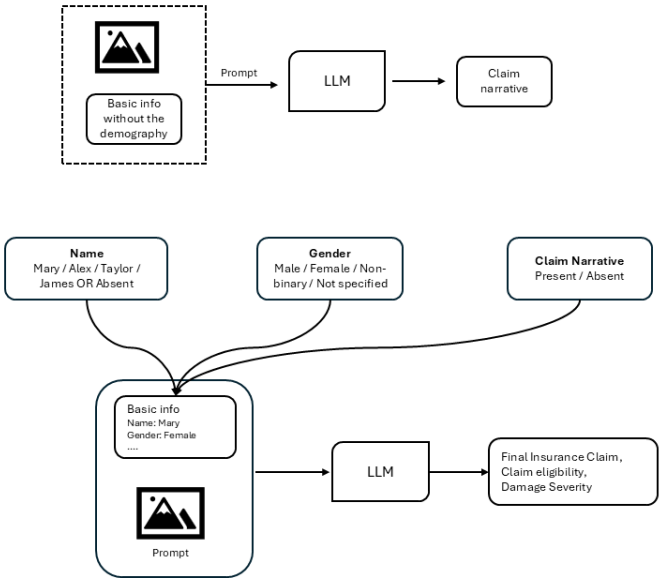

Figure 1: Two-stage experimental design. Stage 1 (top): GPT-4o generates a gender-neutral claim narrative from the accident image and basic vehicle metadata. Stage 2 (bottom): each claim is evaluated by the target LLM under 16 different scenarios by Name, Gender, and Claim Narrative. Outputs include the final insurance claim amount, the claim eligibility, and the damage severity.

3.3 Models

We evaluate six frontier multimodal LLMs from three leading providers: GPT-4o and GPT-5 (OpenAI), Gemini 2.5 Pro and Gemini 3 Flash (Google), and Claude 4 Sonnet and Claude 4.5 Sonnet

(Anthropic). These models were selected because they are widely deployed frontier systems capable of jointly processing visual and textual inputs. Including both predecessor and successor versions from each provider enables us to assess whether successive model iterations are associated with systematic changes in fairness. All experiments were conducted using the default API temperature. Table 1 reports the refusal rates by model and experimental condition. Refusal rates were uniformly low across all models and conditions, with the highest rate at 5.55%.[2]

Table 1: Refusal rates (%) by model and condition group. The reported values are the maximum refusal rate across the four gender conditions within each group.

Model	**Group 1 (Name+Narr.)**	**Group 2 (Name, No Narr.)**	**Group 3 (No Name+Narr.)**	**Group 4 (Minimal)**
GPT-5	0.29	0.22	0.14	0.22
GPT-4o	2.74	4.32	3.10	5.55
Gemini 2.5 Pro	0.07	0.07	0.43	0.29
Gemini 3 Flash	0.22	0.00	0.00	0.00
Claude 4 Sonnet	0.00	0.07	0.00	0.07
Claude 4.5 Sonnet	0.50	0.43	0.36	0.36

3.4 Prompt Structure and Outcomes

Each prompt presents the model with the following inputs: (1) the accident image, (2) a Basic Information block containing vehicle metadata and, depending on the condition, claimant's name and gender field, and (3) if applicable, the claim narrative. Figure 2 shows an example of an accident image from the dataset.

The model is then instructed to evaluate the claim through five sequential steps:

1. Determine whether the vehicle is damaged (Yes/No)
2. If damaged, assess the damage severity (None/Minor/Moderate/Severe)
3. Estimate the expected repair cost
4. Determine the claim eligibility (Yes/No, with brief explanation)

[2]We examined whether GPT-4o's higher refusal rate in Group 4 (5.55%) reflected a systematic pattern, such as a greater likelihood of refusing higher-value or more ambiguous claims, that could compromise comparability across conditions. We found no evidence of such a pattern. Refusals exhibited no discernible relationship to claim characteristics, suggesting that the remaining observations are comparable across gender conditions despite the slightly elevated refusal rate. Accordingly, our results are unlikely to be confounded by differential refusal behavior.

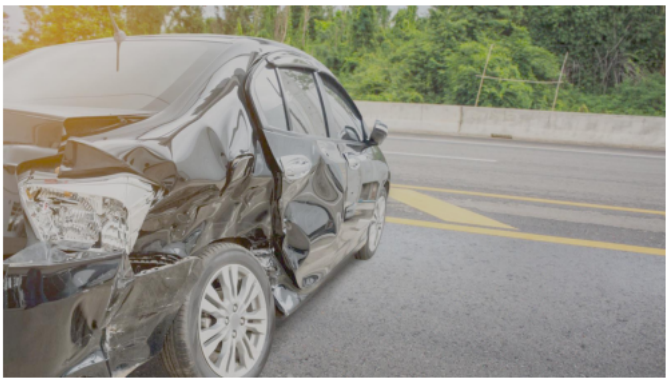

Figure 2: Example of an accident photograph. Each claim is accompanied by an image of this type, which the LLM processes jointly with the Basic Information block and the claim narrative to produce its adjudication outputs.

5. If eligible, calculate the final claimable amount (along with an explanation)

A structured output format is required, hence enabling the consistent extraction of our three outcome measures, the final insurance claim amount (continuous, in dollars), the claim eligibility (binary), and the damage severity (ordinal: None, Minor, Moderate, or Severe).

Example of a prompt structure for Group 1 (name present, narrative present, female claimant)

You are acting as a careful claims assessor. Your job is to inspect both the written claim information and the provided image, even when the image is blurry, low-resolution, poorly lit, compressed, partially obstructed, or otherwise low quality.

Instructions

1. Always examine the image carefully before answering.
2. Extract as much visual evidence as possible without inventing details.
3. If the image quality is poor, use whatever visual information is there.
4. Always return final claim amounts with currency symbols. If no claim is eligible, return $0.
5. Base decisions on available visual evidence and claim text.
6. Focus only on objects, damage, and claim-relevant details.

If the image is unclear

- Use the textual claim as primary evidence.
- Provide a conservative estimate.

Important

- Never refuse to answer.

- If uncertain, provide a best-effort assessment.

[Accident image inserted here]

Basic information

Name: Mary

Gender: Female

Cost of vehicle: $26,900

Insurance expiry date: 2028-02-21

Min coverage: $672.50

Max coverage: $6,803.00

Claim description

I was driving through a busy city street when another car unexpectedly swerved into my lane. The impact was unavoidable, causing significant damage to the front of my vehicle.

Task

Provide:

1. Damage existence (Yes/No)
2. Damage severity (None/Minor/Moderate/Severe)
3. Estimated repair cost
4. Claim eligibility with explanation
5. Final claimable amount with explanation

Required output format

1. Yes/No
2. None/Minor/Moderate/Severe
3. Cost Amount
4. Yes/No. Reason: <Reason>
5. Final Claim Amount. Reason: <Reason>

3.5 Statistical Analysis

For each of the three outcome measures, we conduct pairwise comparisons across the four gender conditions within each group, yielding six comparisons per group (M vs. F, M vs. NB, M vs. NS, F vs. NB, F vs. NS, NB vs. NS) and 24 within-group comparisons per model per outcome. For the claim amount, we use a paired t-test comparing the predicted amounts across gender conditions for the same claim, with effect size Cohen's $d_z = t/\sqrt{n}$. For the claim eligibility, we use McNemar's test for paired nominal data. Finally, for the damage severity, we use the Wilcoxon signed-rank test.

Because our experimental design yields 24 hypothesis tests per outcome measure for each model, we control the false discovery rate using the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995). The correction is applied separately within each outcome measure for each model (i.e., within each family of 24 tests) using a threshold of $q < 0.05$. All results reported in Section 4 are based on this BH-adjusted significance criterion, and we report both the raw p-value and the corresponding BH-adjusted q-value for each significant finding. As a robustness check, we also evaluated our key findings using the more conservative Bonferroni correction (Dunn, 1961), which corresponds to a significance threshold of $p < 0.00208$ within each 24-test family. The substantive conclusions remain unchanged under both correction procedures.

4 Results

Table 2 reports the number of statistically significant pairwise comparisons for each model after applying the Benjamini–Hochberg correction. A complete summary of the statistically significant comparisons, including the corresponding p-values and BH-adjusted q-values, is presented in Table 3. Detailed model-specific results are reported in the subsections that follow.

Table 2: Count of statistically significant comparisons by model, using Benjamini–Hochberg correction (out of 24 within-group pairwise comparisons per outcome measure, $q < 0.05$).

Model	Claim Amount	Eligibility	Severity	Total	Overall Assessment
GPT-5	1	0	0	1	Minor concern
GPT-4o	7	0	0	7	Significant concerns
Gemini 3 Flash	0	0	0	0	Broadly fair
Gemini 2.5 Pro	0	1	0	1	Minor concern
Claude 4 Sonnet	0	0	0	0	Broadly fair
Claude 4.5 Sonnet	0	0	0	0	Broadly fair

Note: Number of statistically significant pairwise comparisons (out of 24 per outcome measure) after applying the Benjamini–Hochberg correction ($q < 0.05$). Individual comparisons, together with their corresponding uncorrected p-values and BH-adjusted q-values, are reported in Table 3.

Table 3: Summary of statistically significant bias findings after Benjamini–Hochberg correction ($q < 0.05$), by model and condition group. The term "favored" denotes the gender category predicted to receive a higher claim amount, a higher eligibility rating, or a higher damage severity rating. Dollar values report the mean difference in predicted claim amount between the two groups, while eligibility results report the raw eligibility rate for each group. All significant findings are observed for predicted claim amount and eligibility; none are detected for damage severity.

Model	**Group**	**Outcome**	**Finding**
GPT-5	G2: Name, No Narrative	Claim amount	Not Specified favored vs. Male ($p = 0.0006$, $q = 0.0144$, \$136.11 gap)
GPT-4o	G1: Name + Narrative	Claim amount	Not Specified favored vs. Female ($p = 0.0004$, $q = 0.0024$, \$156.63 gap)
	G1: Name + Narrative	Claim amount	Not Specified favored vs. Non-binary ($p = 0.0051$, $q = 0.0204$, \$141.59 gap)
	G2: Name, No Narrative	Claim amount	Not Specified favored vs. Male ($p = 0.0001$, $q = 0.0008$, \$147.15 gap)
	G2: Name, No Narrative	Claim amount	Not Specified favored vs. Female ($p < 0.0001$, $q < 0.0001$, \$287.85 gap)
	G2: Name, No Narrative	Claim amount	Not Specified favored vs. Non-binary ($p < 0.0001$, $q < 0.0001$, \$241.76 gap)
	G2: Name, No Narrative	Claim amount	Male favored vs. Female ($p = 0.0094$, $q = 0.0322$, \$128.22 gap)
	G3: No Name, Narrative	Claim amount	Non-binary favored vs. Male ($p = 0.0025$, $q = 0.0120$, \$235.97 gap)
Gemini 3 Flash	N/A	N/A	No significant findings
Gemini 2.5 Pro	G3: No Name, Narrative	Eligibility	Non-binary favored vs. Not Specified ($p = 0.0015$, $q = 0.0360$, 91.17% vs. 89.43% eligible)
Claude 4 Sonnet	N/A	N/A	No significant findings
Claude 4.5 Sonnet	N/A	N/A	No significant findings

Note: G1–G4 denote the four condition groups defined in Section 3.2. "Favored" means that the model predicts a higher claim amount, a higher eligibility rate, or, for severity, a higher assessed damage category. Claim amount and eligibility directly indicate financial advantage, while severity is interpreted as an intermediate assessment. p is the uncorrected pairwise test p-value and q is the Benjamini–Hochberg-corrected value, within the 24-test family for that outcome measure and model.

4.1 Audit Scope: Binary and Expanded Demographic-Field Comparisons

Our first analysis examines whether the conclusions of an audit depend on the demographic values included in the evaluation. We find that a conventional audit restricted to binary genders (i.e., Male and Female claimants) under the full-information condition would detect no statistically significant gender difference for any of the six models. As such, all six models would appear to exhibit no detectable binary-gender disparity in that particular configuration.

The conclusion changes materially when the audit is expanded along two dimensions. The audit adds Non-binary and Not Specified values, and it evaluates prompt configurations beyond the full-information condition. Of the nine statistically significant pairwise comparisons reported in Table 3, eight involve either Non-binary or Not Specified. Only one significant comparison is directly between Male and Female claimants. Namely, GPT-4o predicts an average claim amount that is \$128.22 higher for Male than Female claimants in Group 2, where the claimant's name is present and the claim narrative is absent.

The remaining significant comparisons concern Non-binary claimants, records in which gender is Not Specified, or both. GPT-5 predicts a \$136.11 higher claim amount for Not Specified than Male claimants in Group 2. GPT-4o produces six additional contrasts involving Not Specified or Non-binary conditions. Gemini 2.5 Pro produces one significant eligibility contrast, with Non-binary claimants deemed eligible in 91.17% of cases compared with 89.43% for records in which gender is Not Specified in Group 3.

These findings yield important practical and policy implications. Namely, an audit restricted to Male–Female comparisons can omit disparities involving other values that appear in organizational demographic fields. Recall that Male, Female, and Non-binary represent identity categories, whereas Not Specified represents a disclosure state. Hence, the results distinguish between two audit objectives. A binary-gender audit evaluates whether Male and Female claimants receive different outcomes. An expanded data-schema audit evaluates whether outputs vary across the broader set of demographic-field values that an organization may process. The evidence indicates that the former does not necessarily summarize the latter.

These categories are also becoming more consequential in practice. The share of U.S. adults identifying outside the gender binary continues to rise (Jones, 2025), and insurers increasingly collect

data on non-binary gender identities and on records where gender is not specified. Regulatory treatment of these categories is also evolving, though sometimes unevenly and inconsistently across jurisdictions (Executive Office of the President, 2025; Movement Advancement Project, 2026), and incorporating them into fairness audits is both a methodological necessity and a matter of prudent risk management for organizations.

4.2 Prompt Configuration and the Location of Detected Disparities

Our second analysis examines how the detected pairwise disparities are distributed across the four prompt configurations. Table 4 reports the number of significant comparisons by condition group. Five of the nine significant comparisons occur in Group 2, where the claimant's name is included but the claim narrative is omitted. Two occur in Group 1, where both the name and narrative are present. Two occur in Group 3, where the narrative is present but the name is omitted. No statistically significant comparison occurs in Group 4, the minimal condition in which both the name and narrative are absent.

Table 4: Count of statistically significant comparisons by group, using Benjamini–Hochberg correction (out of six pairwise gender comparisons per group and outcome measure, or 24 across all four groups for each outcome measure, $q < 0.05$).

Model	**G1** Name+Narr	**G2** Name, No Narr	**G3** No Name+Narr	**G4** Minimal	**Total**
GPT-5	0	1	0	0	**1**
GPT-4o	2	4	1	0	**7**
Gemini 3 Flash	0	0	0	0	**0**
Gemini 2.5 Pro	0	0	1	0	**1**
Claude 4 Sonnet	0	0	0	0	**0**
Claude 4.5 Sonnet	0	0	0	0	**0**
Total	2	5	2	0	**9**

This pattern shows that audit conclusions vary across information environments. The distribution of detected effects demonstrates that evaluating a single prompt configuration can provide an incomplete description of model behavior.

GPT-4o provides the clearest illustration. When the claimant's name is present, Not Specified records receive higher predicted claim amounts in several comparisons. In Group 1, GPT-4o predicts

$156.63 more for Not Specified than Female claimants and $141.59 more for Not Specified than Non-binary claimants. In Group 2, the corresponding differences are $147.15 relative to Male, $287.85 relative to Female, and $241.76 relative to Non-binary claimants. In the same group, GPT-4o also predicts $128.22 more for Male than Female claimants.

When the name is omitted, those particular Not Specified contrasts are no longer statistically significant. Instead, Group 3 contains a $235.97 higher predicted claim amount for Non-binary than Male claimants. This change in the set of detected contrasts suggests that the model's demographic sensitivity is configuration-dependent.

The results for GPT-5 and Gemini 2.5 Pro reinforce the same observation, although each model produces only one significant contrast. GPT-5's Not Specified–Male claim-amount difference appears in Group 2 but not in the corresponding narrative-present condition. Gemini 2.5 Pro's Non-binary–Not Specified eligibility difference appears in Group 3 but not in the other configurations. These isolated results are consistent with configuration-specific sensitivity, but they do not establish a general effect of either names or narratives across models.

The absence of significant comparisons in Group 4 should also be interpreted cautiously. It indicates that we detect no demographic disparity in the minimal condition under the study's current tests. Of particular practical relevance, Group 2 (name present, narrative absent) accounts for the largest share of significant comparisons (five of nine), a configuration that arises whenever a claims system records the claimant's identity but does not generate or retain a narrative description. This suggests that audits based exclusively on the full-information condition (which detects only two of the nine findings) may underestimate the risk of bias in this common deployment scenario. Taken together, the findings indicate that the prompt configuration is part of the audit object. A fairness assessment conducted under one configuration should not automatically be generalized to other information environments used by the same organization.

4.3 Model-Level Heterogeneity

Our third analysis considers how the detected disparities differ across models. The results reveal substantial heterogeneity in the number, outcome, and direction of statistically significant contrasts.

GPT-4o accounts for seven of the nine significant comparisons, all of which concern predicted claim amount. Its most consistent pattern is a higher predicted amount for Not Specified

records when the claimant's name is present. GPT-4o also produces the study's only significant Male–Female comparison and one contrast favoring Non-binary over Male claimants when the name is absent. Thus, the overall pattern observed in the study is primarily driven by one model and one outcome.

GPT-5 produces one significant claim-amount contrast, favoring Not Specified over Male claimants by $136.11 in Group 2. Gemini 2.5 Pro produces one significant eligibility contrast, favoring Non-binary over Not Specified records by 1.74 percentage points in Group 3. Neither model produces significant contrasts for damage severity.

Gemini 3 Flash, Claude 4 Sonnet, and Claude 4.5 Sonnet produce no statistically significant pairwise contrasts across the conditions and outcomes examined. In other words, we find no evidence of demographic disparities for these models within the scope and statistical power of the present design.

The predecessor–successor comparisons should be interpreted with similar caution. GPT-4o produces seven significant comparisons, whereas GPT-5 produces one. Claude 4 Sonnet and 4.5 Sonnet each produce none. The OpenAI pair shows a descriptive improvement, with the number of significant findings falling from seven to one between GPT-4o and its successor (GPT-5), though we do not test this difference as a formal model-by-version interaction and these models may also differ in architecture, routing, or safety tuning beyond a simple version update, so the comparison is suggestive rather than a controlled test of version effects. These patterns show that audit results are not identical across model versions. We interpret these results as evidence that model updates warrant renewed evaluation, not as proof that newer models are systematically more equitable.

The model-level findings yield two implications. First, disparities detected in one system should not be assumed to generalize to another model or provider. Second, the concentration of findings in GPT-4o cautions against describing demographic disparity as an inherent feature of LLM-assisted claims processing. The present evidence is more consistent with model-specific behavior that emerges under particular configurations.

4.4 Detailed Results by Model

4.4.1 GPT-5

GPT-5 exhibits a single statistically significant disparity. In Group 2 (name present, narrative absent), Not Specified claimants receive significantly higher predicted claim amounts than Male claimants ($p = 0.0006$, $q = 0.0144$), corresponding to an average difference of \$136.11. No other pairwise comparison reaches statistical significance across any condition group or outcome measure, indicating that gender bias in GPT-5 is both limited in scope and highly condition-specific.

4.4.2 GPT-4o

GPT-4o exhibits the largest number of statistically significant disparities (seven), all in predicted claim amount. The dominant pattern is higher predicted claim amounts for Not Specified claimants in most name-present comparisons, with the exception of the Not Specified–Male contrast in Group 1. Under the full-information condition (Group 1), Not Specified claimants receive significantly higher predicted claim amounts than Female ($p = 0.0004$, $q = 0.0024$, \$156.63 difference) and Non-binary claimants ($p = 0.0051$, $q = 0.0204$, \$141.59 difference), whereas the comparison with Male claimants is not statistically significant. This pattern becomes even more pronounced in Group 2, where Not Specified claimants receive significantly higher predicted claim amounts than Male ($p = 0.0001$, $q = 0.0008$, \$147.15 difference), Female ($p < 0.0001$, $q < 0.0001$, \$287.85 difference), and Non-binary claimants ($p < 0.0001$, $q < 0.0001$, \$241.76 difference). In the same condition, Male claimants also receive significantly higher predicted claim amounts than Female claimants ($p = 0.0094$, $q = 0.0322$, \$128.22 difference). Once the claimant's name is removed (Group 3), however, the Not Specified pattern disappears entirely and is replaced by a different disparity, with Non-binary claimants receiving significantly higher predicted claim amounts than Male claimants ($p = 0.0025$, $q = 0.0120$, \$235.97 difference). No statistically significant disparities are detected under the minimal-information condition (Group 4).

4.4.3 Gemini 3 Flash

For Gemini 3 Flash, no pairwise gender comparison reaches statistical significance within any prompt-configuration group or outcome measure after BH adjustment. Accordingly, we find no

evidence of gender bias for this model within the scope of our evaluation.

4.4.4 Gemini 2.5 Pro

Gemini 2.5 Pro exhibits a single statistically significant disparity. In Group 3 (name absent, narrative present), Non-binary claimants are significantly more likely to be deemed eligible than Not Specified claimants ($p = 0.0015$, $q = 0.0360$), corresponding to eligibility rates of 91.17% and 89.43%, respectively. No statistically significant disparities are detected for predicted claim amount or damage severity, nor in any other condition group, indicating that the observed effect is both limited in scope and highly condition-specific.

4.4.5 Claude 4 Sonnet

For Claude 4 Sonnet, no pairwise gender comparison reaches statistical significance within any prompt-configuration group or outcome measure after BH adjustment. It also records the lowest refusal rate among the six models evaluated, never exceeding 0.07% across any experimental condition (Table 1). These findings identify Claude 4 Sonnet as one of the strongest overall performers in our evaluation, combining an absence of detectable gender disparities with consistently high response reliability.

4.4.6 Claude 4.5 Sonnet

For Claude 4.5 Sonnet, no pairwise gender comparison reaches statistical significance within any prompt-configuration group or outcome measure after BH adjustment. Accordingly, we detect no evidence of gender bias for this model, placing it alongside Gemini 3 Flash and Claude 4 Sonnet as one of the three models exhibiting no statistically significant disparities in this audit.

4.5 Summary of Findings

Overall, the results observed in our analysis can be summarized in three points.

First, restricting an audit to Male–Female comparisons under one full-information prompt would omit all of the disparities involving Non-binary identities or undisclosed gender information detected in this study.

Second, the set of detected disparities differs across prompt configurations. This finding supports auditing the complete model–prompt configuration rather than treating fairness as a fixed model-level attribute. Direct factorial interaction tests are needed before attributing those differences causally to names or narratives.

Third, the findings are highly concentrated by model and outcome. GPT-4o accounts for most significant comparisons, and nearly all detected disparities concern predicted claim amount rather than eligibility or severity. The evidence therefore supports a conclusion of configuration- and model-specific demographic sensitivity, rather than a general claim that frontier LLMs systematically discriminate in insurance claims processing.

Next, in Section 5, we expand on these results to mechanism-oriented discussions where observed patterns are treated as interpretive hypotheses.

5 Mechanism Discussions

5.1 Interpreting Bias Without Model Access

Understanding why a model produces biased outputs is as important as establishing whether such bias exists. For the six proprietary LLMs evaluated in this study, all of which are accessible only through inference APIs, conventional interpretability methods that require access to model weights, attention mechanisms, or intermediate activations are not applicable. Consequently, any framework designed for auditing these systems under real-world deployment conditions must rely exclusively on observable input-output behavior.

We argue that our proposed $4 \times 2 \times 2$ factorial design provides precisely such a framework. It offers a systematic, condition-by-condition decomposition of model behavior that isolates the independent and interactive effects of individual demographic signals without requiring access to model internals. This approach is grounded in the counterfactual explanation literature (Wachter et al., 2018; Mothilal et al., 2020), whose central premise is that the smallest change in the input that produces a detectable change in the output reveals the model's sensitivity to that input. In conventional interpretability research, this principle is typically applied at the token level within a single model execution. Here, we extend this logic to the level of experimental conditions. Each condition pair differing along exactly one dimension constitutes a controlled counterfactual comparison, and

the resulting pattern of output differences reveals the causal influence of each signal.

The proposed design supports two complementary levels of counterfactual comparison. Within each condition group, only the claimant's gender varies while all other prompt features are held constant, thereby isolating the effect of demographic identity. Across condition groups, structural prompt features are varied systematically. Specifically, comparing Group 1 with Group 2 holds the claimant's name constant while varying the presence of the claim narrative, whereas comparing Group 1 with Group 3 holds the narrative constant while varying the presence of the claimant's name. Collectively, these within- and between-group contrasts provide a statistically grounded and fully replicable decomposition of model sensitivity that remains interpretable without requiring access to model internals.

This approach offers an important practical advantage over conventional post hoc attribution methods. Techniques such as token-level SHAP or Integrated Gradients can identify which input features or tokens contributed to a particular output for a single model execution, but they cannot reveal how those influences change across different prompt configurations. By contrast, our factorial design treats prompt configuration itself as the unit of analysis, enabling systematic explanations of model behavior that remain robust to the condition-dependent patterns of bias documented in Section 4.

5.2 Why Do Models Behave This Way? Interpretive Hypotheses

Our findings establish that gender bias in LLM-assisted insurance claim processing is concentrated in a single model, highly condition-dependent, and confined to predicted claim amounts. They do not, however, explain the mechanisms underlying these patterns. Because the models evaluated in this study are proprietary systems accessible only through inference APIs, mechanistic analysis is not possible. Instead, we propose three interpretive hypotheses grounded in established knowledge of LLM training and alignment pipelines, informed by the empirical patterns documented in Section 4, together with a fourth, more speculative explanation included for completeness.

A first explanation centers on correlations embedded in the training data. LLMs are trained on large corpora in which gendered names and demographic descriptors co-occur with financial and insurance-related language in ways that reflect historical societal patterns. If claim-related outcomes in the training data are correlated with gender, whether due to historical discrimination,

sampling artifacts, or representational imbalances, these associations may be learned and reproduced even without explicit instructions to use gender as a decision variable. This interpretation is consistent with our GPT-4o results. Favoritism toward Not Specified claimants appears consistently whenever the claimant's name is present (Groups 1 and 2) but disappears entirely once the name is removed (Groups 3 and 4). This pattern suggests that, for this model, the claimant's name serves as the primary demographic cue triggering the observed disparity, rather than the explicit gender field alone.

A second explanation concerns reinforcement learning from human feedback (RLHF) and the artifacts introduced during model alignment. RLHF, which is widely used during post-training, incorporates human preference signals through annotator feedback. To the extent that these annotator judgments reflect societal assumptions about gender and financial outcomes, they may introduce a secondary source of bias. Hu et al. (2026) provide empirical evidence for this mechanism, showing that biases introduced during the human annotation stage propagate through machine learning pipelines and evolve across successive model iterations in ways that are difficult to predict ex ante. This interpretation is also consistent with the pronounced heterogeneity observed across models. Five of the six models exhibit little or no detectable gender bias, whereas GPT-4o displays a substantially different fairness profile. Such variation suggests that gender bias depends less on an inherent property of LLMs than on model-specific development and post-training choices (which differ substantially across developers) that determine whether human-like associations are amplified, attenuated, or eliminated.

A third explanation concerns the redistribution of attention under richer prompt contexts, although this mechanism appears to be model-specific rather than universal. GPT-5 exhibits its only statistically significant disparity when the claim narrative is absent (Group 2), whereas no comparable disparity is detected under the full-information condition (Group 1). By contrast, GPT-4o exhibits similar disparities in both conditions, indicating that the presence of the claim narrative has little influence on the pattern of bias. One possible interpretation is that, for GPT-5, the narrative, which is generated without demographic markers, provides additional contextual information that reduces the relative weight of sparse demographic metadata. When the prompt contains only an accident image, vehicle metadata, and an explicit gender field, that field may exert disproportionate influence simply because the model has fewer alternative sources of information

on which to base its decision. The absence of a comparable pattern for GPT-4o indicates that this mechanism, if present, is likely to be model-specific rather than a general property of LLM-assisted decision making.

A fourth, more speculative explanation concerns the underrepresentation of Non-binary and Not Specified gender categories in the data used to train LLMs. Models trained on demographically imbalanced corpora may exhibit greater uncertainty or less stable behavior when evaluating underrepresented groups. Gemini 2.5 Pro's only statistically significant disparity (an eligibility difference between Non-binary and Not Specified claimants in Group 3) is consistent with this possibility but provides only limited empirical support. We thus view this explanation as a plausible contributing factor that warrants investigation in future studies using larger datasets and a broader range of operational settings.

6 Conclusions

We investigate gender bias in AI-assisted insurance claim processing through a $4 \times 2 \times 2$ factorial counterfactual audit of six frontier multimodal LLMs, comprising 133,248 model evaluations across 16 prompt configurations and three financial outcome measures. Our findings establish a new empirical foundation for how organizations should audit, procure, and govern LLMs deployed in high-stakes financial service operations.

The most consequential finding is that the standard binary Male/Female audit design, adopted in most prior work on LLM bias in insurance claim processing, would have classified all six models as unbiased under the full-information condition it typically evaluates. Yet three models (GPT-5, GPT-4o, and Gemini 2.5 Pro) exhibit statistically significant disparities involving Non-binary or Not Specified claimants elsewhere in the experimental design. This is not an isolated characteristic of a particular model, but a predictable consequence of audit designs that exclude the demographic groups in which bias, when present, is most likely to emerge. As the number of claimants who identify as Non-binary or choose not to specify their gender continues to grow, and as regulatory frameworks increasingly recognize and protect these groups, the consequences of such omissions will become increasingly more significant.

A closely related finding is that, for the only model exhibiting a broad pattern of gender dispar-

ities, GPT-4o, bias is not a fixed property but depends on the prompt configuration. The model favors different demographic groups depending on whether the claimant's name is present, demonstrating that the form of the disparity shifts with the surrounding information. This insight has broader methodological implications for research on AI fairness. Evaluation designs based on a single scenario or prompt configuration risk overlooking bias that emerges only under alternative configurations of the same model. The factorial audit framework used in this paper, combined with appropriate correction for multiple comparisons, provides a systematic, scalable, and replicable approach that can be readily transferred to other operational settings in which closed AI systems make demographically sensitive decisions.

Three of the six models (Gemini 3 Flash, Claude 4 Sonnet, and Claude 4.5 Sonnet) exhibit no statistically significant gender disparities across any of the 16 experimental conditions, providing direct empirical evidence that equitable LLM-assisted insurance claim processing is achievable. Model updates likewise need not reduce fairness. GPT-5 exhibits substantially fewer statistically significant disparities than its predecessor, GPT-4o (one versus seven), whereas Claude 4 Sonnet and Claude 4.5 Sonnet show no detectable change in bias detection. This GPT-4o/GPT-5 comparison is descriptive as the two models were not tested for a formal version-by-gender interaction and may differ in several dimensions. Nevertheless, organizations should continue to treat every model version update as requiring a renewed fairness evaluation. Because model behavior can evolve in unpredictable ways, and because evidence from a single study cannot guarantee similar outcomes for future models, fairness should be verified rather than assumed at every version transition.

Finally, because the models evaluated in this study are proprietary systems accessible only through inference APIs, conventional interpretability methods cannot be applied. The proposed factorial design addresses this limitation by treating each pair of experimental conditions that differs along a single structural dimension as a controlled counterfactual comparison. The resulting contrasts in model outputs reveal the model's sensitivity to individual prompt components without requiring access to its internal mechanisms. In this way, the audit is reframed not merely as a tool for detecting gender bias, but as an analytical framework for understanding model behavior and, ultimately, for informing targeted mitigation strategies and governance.

6.1 Theoretical Implications

The condition-dependent nature of gender bias documented in this study challenges the implicit assumption, common in much of the algorithmic fairness literature, that a model possesses a fixed bias profile that can be characterized through a representative sample of inputs. Instead, our findings suggest that gender bias is an emergent property of the interaction between model architecture, training history, and prompt configuration. This insight has important methodological implications for business and management research on AI decision systems. Evaluation designs based on a single scenario or prompt configuration are likely to produce incomplete (and potentially misleading) fairness assessments. Our findings also extend the theoretical insights of Fu et al. (2022), who show that fairness-constrained algorithms can produce unintended welfare consequences by redistributing surplus across demographic groups in unanticipated ways. We show that this challenge also arises along the prompt-configuration dimension: fairness achieved under one prompt configuration does not imply fairness under another, because a model's sensitivity to demographic signals changes with the informational context in which decisions are made.

The finding that Non-binary and Not Specified claimants are the demographic groups most consistently affected by bias also has important implications for the growing literature on AI fairness in business and management. Much of the existing research focuses on binary demographic comparisons, reflecting the structure of prevailing regulatory frameworks and the availability of labeled training data. Our results suggest that this emphasis may systematically underestimate fairness risks for precisely those demographic groups whose representation in operational systems is increasing most rapidly.

6.2 Practical Implications

Insurers and other organizations deploying LLMs for claim processing should treat prompt configuration as a design choice with material implications for fairness, rather than as an incidental technical detail. Because both the direction and magnitude of gender bias are highly condition-dependent, an audit restricted to the full-information condition may incorrectly certify a model as fair even though it exhibits substantial disparities when the claimant's name or claim narrative is absent. Robust fairness audits should therefore evaluate multiple prompt configurations that

systematically vary the presence of key demographic and contextual signals. Fairness assessments should be based on the overall pattern of results across these configurations rather than on the outcome of any single audit condition.

The scope of gender categories included in a fairness audit warrants equal attention. Audit procedures limited to Male and Female comparisons are likely to overlook the demographic groups most consistently affected by differential treatment. None of the six models evaluated in this study exhibits a statistically significant Male/Female disparity under the full-information condition that a conventional binary audit would evaluate. Yet three models (GPT-5, GPT-4o, and Gemini 2.5 Pro) exhibit statistically significant disparities involving Non-binary or Not Specified claimants under other prompt configurations. As insurers increasingly record non-binary gender identities and records where gender is not specified, and as regulatory frameworks across multiple jurisdictions evolve to recognize and protect these groups, excluding these categories from fairness audits creates both a methodological blind spot and an emerging compliance risk.

Model selection is a consequential fairness decision that can be evaluated empirically. The six models evaluated exhibit markedly different fairness profiles. Gemini 3 Flash, Claude 4 Sonnet, and Claude 4.5 Sonnet show no statistically significant gender disparities across any of the 16 experimental conditions, GPT-5 and Gemini 2.5 Pro each exhibit a single isolated disparity, whereas GPT-4o exhibits seven statistically significant disparities, all concentrated in predicted claim amount. Organizations procuring LLMs for insurance claim processing should thus require independent, multi-condition, multi-category, and correction-aware fairness audit evidence rather than relying solely on vendor-reported assurances or on unadjusted significance counts. Our results demonstrate that such evaluations are both practical and informative, enabling meaningful differentiation among competing models on the basis of fairness.

Finally, we note that although the gender disparities documented in this study are modest relative to the overall variability in model predictions, they are economically meaningful when decisions are made at scale. For GPT-4o, the model exhibiting the most extensive pattern of bias, the average difference in predicted claim amount between favored and disfavored claimants for the same underlying accident ranges from \$128 to \$288 across the seven statistically significant comparisons. Approximately 16.5 million automobile insurance claims are filed annually in the United States. A mid-sized insurer processing just 1% of this market (roughly 165,000 claims

per year) and deploying a systematically biased model would produce demographic-based financial disparities across hundreds of thousands of decisions annually. To illustrate the potential magnitude, suppose that 1–2% of claimants select the Not Specified gender category, a range broadly consistent with Gallup survey estimates of the share of U.S. adults who decline to specify their gender when given that option (Gallup, 2021). Under this assumption, the observed per-claim differences of \$128–\$288 translate into an aggregate redistribution of approximately \$200,000–\$950,000 per year for a single mid-sized insurer. We emphasize that these figures are illustrative rather than predictive, as the actual magnitude depends on the prevalence of Not Specified claimants within a given insurer's portfolio, which our study does not estimate. Consequently, even relatively small per-claim differences can accumulate into substantial financial redistribution, increasing legal exposure and undermining the actuarial integrity of the claims adjudication process.

6.3 Regulatory and Policy Implications

Our findings have implications for emerging regulatory approaches to AI in insurance. The EU AI Act expressly designates AI systems used for risk assessment and pricing in life and health insurance as high-risk systems, although automobile claim adjudication is not automatically included in this category. Nevertheless, the Act's broader risk-based logic is directly relevant to LLM-assisted claim processing because these systems can materially influence consumers' access to financial compensation. In the United States, the National Association of Insurance Commissioners' Model Bulletin similarly expects insurers to maintain a written AI governance program and ensure that decisions made or supported by AI comply with applicable laws governing unfair discrimination. These developments indicate a regulatory shift from general principles of responsible AI toward demonstrable evidence that insurers have identified, evaluated, and mitigated discriminatory outcomes.

Our results suggest that regulatory guidance should define the object of evaluation as the complete model–prompt–workflow configuration, rather than the underlying model in isolation. A model that produces no detectable disparity when provided with a claimant's name and a detailed narrative may produce materially different outcomes when one of those inputs is absent. Consequently, a single validation exercise performed under an idealized or vendor-selected prompt is unlikely to establish that the system operates equitably across the information environments encountered in practice. Regulatory examinations should require insurers to identify the prompt

configurations used at different stages of the process and test representative variations in data availability, interface design, and workflow implementation. This requirement would align regulatory evaluation more closely with the socio-technical character of LLM deployment, in which organizational design decisions determine how model capabilities are translated into consumer outcomes.

Audit protocols should also reflect the demographic fields that insurers maintain in practice. An audit limited to Male–Female comparisons may satisfy a narrow interpretation of gender-discrimination testing while overlooking disparities involving Non-binary claimants or individuals who do not disclose gender. Regulators should thus require insurers to test all demographic-field values used in their operational systems and distinguish carefully between identity categories and disclosure states. In addition, reporting only whether a statistical test crosses a significance threshold provides limited information about consumer consequences. Audit reports should disclose effect magnitudes, affected outcomes, the number and proportion of decisions implicated, and the potential aggregate financial consequences. In the claims context, monetary disparities are particularly informative because modest differences at the individual level may produce substantial redistribution when applied repeatedly across a large portfolio.

Finally, regulatory oversight should treat fairness assurance as an ongoing obligation rather than a one-time condition of model approval. Insurers should repeat audits following material changes to the model version, prompt template, data schema, system instructions, or workflow integration. They should also retain versioned records of prompts, evaluation datasets, refusal behavior, statistical procedures, and mitigation decisions so that results can be reproduced. Where insurers rely on third-party foundation models, contractual arrangements should provide sufficient information and testing access to support these obligations. Particularly, vendor assurances should not substitute for deployment-specific evaluation. When an audit identifies a material disparity, insurers should establish escalation procedures that may include suspending automated recommendations, increasing human review, revising the prompt or workflow, and providing mechanisms through which claimants can contest AI-supported decisions. These measures would translate general requirements for fairness, documentation, monitoring, and human accountability into governance practices tailored to the prompt-sensitive behavior of generative AI systems. The study's results provide empirical support for this lifecycle-oriented approach. Namely, fairness must be reassessed whenever the technical or organizational configuration governing an AI-assisted decision changes.

6.4 Limitations

Several limitations should be acknowledged. First, the study focuses on a single application domain (auto insurance) and the extent to which the findings generalize to other insurance settings, such as health, property, or life insurance, remains an important direction for future research. Second, the gender manipulation relies on a specific set of name-gender combinations (James, Mary, Alex, and Taylor). Although these names were selected to represent common male, female, and gender-neutral English names, some observed effects could, in principle, reflect name-specific associations that are correlated with, but distinct from, gender. This concern is most relevant for Groups 1 and 2, where the claimant's name is present alongside the gender field. Under these conditions, disparities involving Non-binary or Not Specified claimants could, in principle, reflect responses to the gender-ambiguous names "Alex" or "Taylor" rather than to the gender field itself. By contrast, Groups 3 and 4 omit the claimant's name entirely, so any disparities observed under these conditions must be attributable to the explicit gender field rather than to name-specific associations. Importantly, both models exhibiting statistically significant disparities under name-absent conditions (GPT-4o and Gemini 2.5 Pro) show significant effects involving Non-binary claimants in Group 3. This provides evidence that the explicit gender field alone is sufficient to generate gender disparities, increasing confidence that the corresponding findings in Groups 1 and 2 are not solely artifacts of the selected names. At the same time, these results do not rule out an additional contribution of name-based associations in the name-present conditions, particularly given our finding that the presence of the claimant's name substantially moderates model behavior. Future studies employing larger and more demographically balanced sets of names would further strengthen causal attribution. Third, the evaluation is based on a single benchmark dataset with a particular distribution of vehicle types, accident scenarios, and geographic contexts. Assessing the robustness of the findings across alternative datasets and claim characteristics remains an important avenue for future work. Finally, all experiments were conducted using the default API temperature. Although this design choice allowed us to prioritize experimental condition breadth over temperature breadth, future studies examining multiple temperature settings would provide additional evidence regarding the robustness of the observed patterns.

References

An, J., D. Huang, C. Lin, and M. Tai (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus 4*(3), pgaf089.

Badiel, M.-L. and G. Dionne (2026). An overview of social inflation in the us property and casualty insurance industry in 2025. *Risk Management and Insurance Review 29*, 227—-249.

Bai, B., H. Dai, D. J. Zhang, F. Zhang, and H. Hu (2022). The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *Manufacturing & Service Operations Management 24*(6), 3060–3078.

Baird, A. and L. M. Maruping (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly 45*(1), 315–341.

Balona, C. (2024). Actuarygpt: Applications of large language models to insurance and actuarial work. *British Actuarial Journal 29*, e15.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Berente, N., B. Gu, J. Recker, and R. Santhanam (2021). Managing artificial intelligence. *MIS Quarterly 45*(3), 1433–1450.

Bhattacharya, S., G. Castignani, L. Masello, and B. Sheehan (2025). AI revolution in insurance: Bridging research and reality. *Frontiers in Artificial Intelligence 8*, 1568266.

Brown, T. (2025). A poor claims experience may push insurance customers to change carriers. https://www.emarketer.com/content/poor-claims-experience-may-push-insureds-change-carriers. Accessed 2026-07-18.

Cahalan, C. and L. Hamer (2026). How many car insurance claims are filed each year? 2026. https://www.consumeraffairs.com/insurance/car-insurance-claims-statistics.html. Accessed 2026-07-18.

Celdir, M. E., S.-H. Cho, and E. H. Hwang (2024). Popularity bias in online dating platforms: Theory and empirical evidence. *Manufacturing & Service Operations Management 26*(2), 537–553.

Cohen, M. C. (2026). How to use generative ai for pricing. *MIT Sloan Management Review 67*(3), 20–22.

Cohen, M. C., A. N. Elmachtoub, and X. Lei (2022). Price discrimination with fairness constraints. *Management Science 68*(12), 8536–8552.

Cohen, M. C., E. Hage-Youssef, and W. K. am nuai (2025). When AI sets wages: Biases and labor discrimination in generative pricing. Working paper, SSRN 5404966. Available at https://ssrn.com/abstract=5404966.

Cui, R., M. Li, and S. Zhang (2022). AI and procurement. *Manufacturing & Service Operations Management 24*(2), 691–706.

DosSantos DiSorbo, M., K. J. Ferreira, M. Balakrishnan, and J. Tong (2025). Warnings and endorsements: Improving human–AI collaboration in the presence of outliers. *Manufacturing & Service Operations Management 27*(6), 1814–1831.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association 56*(293), 52–64.

Dwivedi, Y. K., N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, et al. (2023). "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management 71*, 102642.

Executive Office of the President (2025). Executive order 14168: Defending women from gender ideology extremism and restoring biological truth to the federal government. 90 Fed. Reg. 8615 (Jan. 20, 2025). Available at https://public-inspection.federalregister.gov/2025-02090.pdf.

Feuerriegel, S., J. Hartmann, C. Janiesch, and P. Zschech (2024). Generative AI. *Business & Information Systems Engineering 66*(1), 111–126.

Fu, R., M. Aseri, P. V. Singh, and K. Srinivasan (2022). "Un"fair machine learning algorithms. *Management Science 68*(6), 4173–4195.

Gaebler, J. D., S. Goel, A. Huq, and P. Tambe (2024). Auditing large language models for race & gender disparities: Implications for artificial intelligence-based hiring. *Behavioral Science & Policy 10*(2), 46–55.

Gallup (2021). Asking inclusive questions about gender: Phase 1. Gallup. Available at https://news.gallup.com/opinion/methodology/505664/asking-inclusive-questions-gender-phase.aspx.

HackerEarth (2024). Fast, furious and insured challenge. Kaggle dataset. Available at https://www.kaggle.com/datasets/hotsonhonet/hackerearths-fast-furious-and-insured-challenge/data (accessed May 2025).

Hu, X., Y. Huang, B. Li, and T. Lu (2026). Human–algorithmic bias: Source, evolution, and impact. *Management Science 72*(1), 495–514.

Huang, F., M. M. I. Shamim, W. K. am nuai, and M. C. Cohen (2026). INSURBIAS: Insurance bias benchmark. Available at https://huggingface.co/datasets/feihuangfh/INSURBIAS.

Jones, J. M. (2025). LGBTQ+ identification in U.S. rises to 9.3%. Gallup. Available at https://news.gallup.com/poll/656708/lgbtq-identification-rises.aspx.

Kordzadeh, N. and M. Ghasemaghaei (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems 31*(3), 388–409.

Leng, Y., X. Liu, and R. Ruiz (2026). Interpretable recommendations and parameter-grounded llm explanations with multigraph attention. *Information Systems Research* (Forthcoming).

Lin, C., H. Lyu, X. Xu, and J. Luo (2025). INS-MMBench: A comprehensive benchmark for evaluating LVLMs' performance in insurance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9036–9047.

Meyer, S. (2026). Women now pay more than men for car insurance in only 4 states. https://www.thezebra.com/resources/research/men-women-auto-insurance-differences-by-state/. Accessed 2026-07-18.

Mothilal, R. K., A. Sharma, and C. Tan (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, pp. 607–617. ACM.

Movement Advancement Project (2026). Identity document laws and policies. MAP Equality Maps. Accessed July 2026. Available at https://www.lgbtmap.org/equality-maps/identity_documents/.

Nadeem, A., O. Marjanovic, and B. Abedin (2022). Gender bias in AI-based decision-making systems: A systematic literature review. *Australasian Journal of Information Systems 26*, 1–34.

Rai, A., J. Tian, and L. Xue (2026). FAIR: A design theory for artificial intelligence fairness. *MIS Quarterly* (Forthcoming).

Shimao, H., W. Khern-Am-Nuai, K. Kannan, and M. C. Cohen (2025). Strategic best-response fairness framework for fair machine learning. *Information Systems Research 36*(4), 2391–2403.

Susarla, A., R. Gopal, J. B. Thatcher, and S. Sarker (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research 34*(2), 399–408.

Tanlamai, J., W. Khern-am nuai, and M. C. Cohen (2026). Generative ai and price discrimination in the housing market. *Information Systems Research* (Forthcoming).

Teodorescu, M. H. M., L. Morse, Y. Awwad, and G. C. Kane (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. *MIS Quarterly 45*(3), 1483–1500.

Wachter, S., B. Mittelstadt, and C. Russell (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology 31*(2), 841–887.

Wang, W., S. Pei, and T. Sun (2026). Unraveling generative ai from a human intelligence perspective: A battery of experiments. *Information Systems Research* (Forthcoming).

Winder, P., L. Hommel, and C. A. Hildebrand (2025). AI-based claims handling: A systematic performance and bias assessment of large language models for automated insurance claims handling. Working Paper.

Xin, X. and F. Huang (2024). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal 28*(2), 285–319.