AI-Based Claims Handling: A Systematic Performance and Bias Assessment of Large Language Models for Automated Insurance Claims Handling

Abstract

The integration of Large Language Models (LLMs) into insurance operations is expected to transform traditional insurance claims handling. Despite anecdotal evidence from single cases by insurance providers and agencies, a systematic assessment of the performance and potential biases of LLMs to perform claims handling across a wide variety of insurance domains is absent. This paper presents a systematic evaluation of LLM-powered claims handling across three studies. First, we benchmark LLM performance against human insurance clerks using standardized exam cases. Second, we assess the performance and potential biases of LLM-based claims handling in simulated cases across four major insurance domains and tasks. Third, we examine the applicability of LLMs by comparing their claim assessments to those of an expert third-party administrator based on real-world claim cases. Our findings show that LLMs process claims not only efficiently (seconds of processing image data, processing and matching text data from policy documents and damage reports, etc.) but also highly accurately (close to perfect assessments for simple claims and up to 94% accuracy for more complex claims involving multiple parties and policies), often exceeding human benchmarks and reducing claims leakage. These findings have important implications for the future of automated claims handling and the potential to not only substantially improve the speed but also the accuracy of claims handling in industry practice.

Keywords: LLM, Generative AI, Algorithmic Bias, Claims Handling, Claim Settlement, Claim Leakage

Introduction

The insurance industry is undergoing a paradigm shift. LLMs are starting to fully automate the underwriting, contract change, billing, cancellation, and claims handling process (Koetter et al., 2019). Recent developments in LLMs are predicted to have a productivity impact of up to USD 100 Billion for Property & Casualty claims handling alone (Donnelly et al., 2024). This is equivalent to up to 5% of the global Property & Casualty premium, driven by an estimated 20% to 25% decrease in loss of administrative expenses and a 30% to 50% reduction in claims leakage[1] (Donnelly et al., 2024; Kimura et al., 2024). Major insurers across the globe invest in their LLM capabilities, such as AXA developing its "AXA Secure GPT" to optimize both internal and customer-facing processes (e.g., automated claims handling), RGA leveraging a third-party LLM to summarize medical reports to speed up the claims process, or Allstate automatically drafting all 50,000 daily claims-related policyholder communications with a customized version of OpenAI's GPT (AXA, 2023; Bousquette, 2025; RGA, 2023). These productivity gains are important not just from a business operations perspective but also to address customer satisfaction and churn rates by streamlining the claims handling process for clients, which is often reported as a major source of frustration for policyholders (NAIC, 2024b).

One example with high media coverage is how insurance carrier Lemonade used their conversational bot "AI Jim" to offer a fully integrated end-to-end claims handling bot that settled the theft of a client's stolen luxury jacket within three seconds, including submitting the claim, cross-referencing it against a customer's policy, approving and settling the claim (Schreiber, 2016). While such examples are single observations based on company-reported cases, they signal the value of how LLMs can efficiently automate customer-facing processes in the insurance industry at scale and potentially reduce widespread frustration with the claims handling process (de Andrade & Tumelero, 2022; Riikkinen et al., 2018).

[1] Claims leakage refers to the financial loss insurers experience due to non-policy-consistent claims handling e.g., overpayment.

Although LLMs are expected to create an upside potential for policyholders and insurers through improved operational efficiency and customer satisfaction, a systematic and large-scale assessment of the performance and biases of LLMs for insurance claims handling is absent. Such a systematic assessment is important as recent work suggests that even drafting a basic reference letter, an LLM can exhibit gender bias by using stereotypically gendered descriptors such as "respectful" for males and "warm" for females, causing potential backlash by customers (Wan et al., 2023).

The current paper provides such a first, systematic, and large-scale assessment of using LLMs for automated claims handling. In Study 1, we benchmark LLM performance against human insurance clerks using standardized, professional exam cases. Study 2 extends this analysis by examining both performance and potential algorithmic biases across claim cases of varying complexity in four major insurance domains: property, casualty, life, and health. Finally, Study 3 evaluates the real-world implementation potential of LLMs by comparing their assessments to those made by a third-party claim administrator. Our findings demonstrate that proprietary LLMs consistently perform at or above human benchmarks in claims handling accuracy while reducing the claims handling time down to seconds and the costs down to cents per claim. Importantly, the current results reveal minimal algorithmic bias, highlighting that insurers can simultaneously enhance operational productivity and elevate customer satisfaction through faster, more transparent, and consistently accurate (and arguably fair) claim settlements. The implications extend beyond a mere cost reduction, pointing toward a more fundamental shift of future claims handling that benefits both insurers and policyholders in a competitive marketplace.

In what follows, we first provide a focused review of existing research on the challenges of human compared to LLM-based claims handling. We then report evidence across three studies on the performance and biases of LLMs in claims handling and close with a discussion and future research for key stakeholders in the insurance value chain (policyholders, insurers, InsurTech companies, and regulators).

Theoretical background

Current challenges in insurance claims handling

Improving performance and unbiased claims handling is crucial, as it directly impacts profitability (NAIC, 2024d), customer satisfaction, including trust in the insurer (Courbage & Nicolas, 2021; Eckert et al., 2022), and regulatory compliance. Performance in claims handling can be defined as the efficient ("promptly") and accurate ("fairly") evaluation of coverage and reimbursement amounts (FCA, 2023, ICOBS 8.1.1).

Efficiency in claims handling can be assessed from the insurer perspective through claims handling expenses, which have risen in the Property & Casualty sector by an average of 4% annually from 2016 to 2024 and fluctuated between 4% and 12% of premiums, depending on the year and region (EIOPA, 2024; NAIC, 2024e). In some retail lines, such as motor insurance, where combined ratios often hover around 100% in a highly competitive environment (NAIC, 2024a), this represents a substantial and strategically important cost component that must be managed efficiently. From a customer perspective, efficiency also remains a concern, as reflected in consumer complaint data, where delays in claims processing are consistently cited among the top reasons for dissatisfaction (EIOPA, 2023; NAIC, 2024c). Beyond efficiency, accuracy is equally critical in claims handling. It entails the fair and unbiased determination of both coverage and reimbursement amounts. Inaccuracies in this process give rise to claims leakage, the gap between the amount actually paid by the insurer and the lower amount that should have been paid, which is estimated to account for 5% to 10% of total U.S. personal auto claims annually (Cognizant, 2018). Moreover, addressing claims leakage requires not only process improvements but also an active response to systematic biases in claim assessments, including those based on gender, ethnicity, age, and socio-economic status (see Table 1). For example, gender and age biases were detected as female, young drivers, and elderly were exposed to a higher assessment of fault based on automobile bodily injury liability claims (Doerpinghaus et al., 2003). Likewise, prior work also documented gender biases as female compared to male claimants received less auto liability insurance payouts (Doerpinghaus et al., 2008). The literature also

documented the presence of an ethnicity bias as Hispanic claimants got insurance reimbursements later compared to non-Hispanic (Baker & McElrath, 1996a), while black claimants got fewer amount of approved payments (Lin et al., 2022). Furthermore, socio-economic bias was revealed in health insurance and financial loan services settings in which low-income individuals were treated worse (e.g., longer waiting times, more negative service experiences) (Becker & Newsom, 2003a; Scott et al., 2023a). The downstream consequences of biased claims handling can be severe, including violations of anti-discrimination laws, and consequently costly lawsuits, as well as damage to brand reputation through negative press coverage (Flitter, 2022).

Potential of LLM-based claims handling

Current LLMs could effectively address the above performance and bias challenges due to their demonstrated capabilities in information assessment, complex reasoning, and consistent decision-making across diverse contexts (Anthropic, 2024; Gunasekar et al., 2023; Jiang et al., 2023; OpenAI, 2023a; Srivastava et al., 2022; Touvron et al., 2023). For instance, GPT-3 performs at a level equivalent to freelance writers in news summarization tasks (Zhang et al., 2024) and surpasses human experts in predicting outcomes of neuroscience experiments based on the assessment of past experiments (Luo et al., 2024), which could be helpful during the initial phase of an accurate claim assessment, summarizing and matching an existing policy document to the damage report filed by a customer. Similarly, GPT-4 ranks in the top 10% on the Uniform Bar Examination, a test evaluating the reasoning skills of prospective lawyers (OpenAI, 2023a), which could support tasks related to coverage determination after the initial assessment phase. Additionally, GPT-4 achieves a top 20% score in the Quantitative section of the Graduate Record Examination (GRE), indicating its potential utility in calculating potentially accurate reimbursement amounts after assessing and matching the existing policy of a client, accounting for potential deductions, and finally suggesting an appropriate reimbursement for the settlement of a claim (OpenAI, 2023a).

In applied insurance settings, LLMs have already been tested for use in claim document classification and inconsistency detection (Li et al., 2025). Initial case studies have started to demonstrate how LLMs can

process claim descriptions, extract structured information, summarize complex documents, and assist in fraud detection (Balona, 2024). These functionalities could lead to more efficient and accurate claims processing, thereby reducing claims handling expenses and potentially improving customer satisfaction.

However, although LLM providers have aggressively implemented a range of debiasing interventions from pre-training data filtering, algorithm debiasing measures, and response filtering (Anil et al., 2024; Anthropic, 2024; OpenAI, 2023b; Touvron et al., 2023), LLMs also show biases analogous to the biases in human-based insurance claims handling (Bender et al., 2021; Ferrara, 2023; Weidinger et al., 2021). For example, prior work indicates that LLMs may possess a gender bias implicitly assuming that "doctors" are males and "nurses" are females (Kotek et al., 2023). Ethnicity bias in LLMs was revealed by responses to medicine-related queries, such as those concerning differences in skin thickness between Black and White individuals, where several LLMs provided inaccurate and scientifically unsupported information, suggesting differences where none exist (Omiye et al., 2023). GPT-4 also tended to overrepresent stereotypes of diseases, such as a greater prominence of sarcoidosis (i.e., an inflammatory disease) in Black and hepatitis B in Asian patients (Zack et al., 2024), potentially leading to a systematically unfair treatment of clients.

Given that LLMs often achieve top-percentile performance in key skills required for claims handling (see preceding section), we posit and test whether and to which extent LLM-powered claims handling bots can match or even exceed human claim handlers in accurately determining coverage and reimbursement amounts, and the presence of potential biases (e.g., varying reimbursements based on gender, ethnicity, age, or socio-economic status).

Article	Task executer	Context	Research Design	Type of Bias	Main IVs	Main DVs	Main Finding
(Baker & McElrath, 1996a)	Human	Home insurance claims handling	Correlational & qualitative	Ethnicity	Ethnicity	Insurance payment within 1 month	Hispanics (vs. white respondents) are only 60% as likely to get a payment within one month
(Becker & Newsom, 2003a)	Human	Health insurance	Qualitative	Socio-economic status	-	-	Low-income participants are likely to identify behavior as discrimination
(Doerpinghaus et al., 2003)	Human	Motor claims handling	Correlational	Age & gender	Age, gender	Policyholder fault rate	Female, elderly, and young drivers assessed a higher fault percentage
(Doerpinghaus et al., 2008)	Human	Motor claims handling	Correlational	Age & gender	Age, gender	Settlement amount	Female claimants receive lower payments
(Lin et al., 2022)	Human	Earthquake claims handling	Correlational	Ethnicity	Ethnicity	Settlement rate & amount	Blacks are less likely to be paid & are paid less
(Scott et al., 2023a)	Human	Loan Seeking	Experimental	Ethnicity, socio-economic status	Ethnicity, socio-economic status	Employees' warmth, competence & service quality	Black business owners, with lower SES experience less favorable treatment from service employees
Current paper	LLM	Life, Health, Property and Casualty (Study 2)	Experimental	Age, gender, ethnicity, socio-economic status	Age, gender, ethnicity, socio economic status;	Relative excess reimbursement amount, reimbursement letter structure & language style	LLMs exceeds human experts in industry exam benchmarks (Study 1), providing unbiased claim assessments across a broad range of claim cases and policyholder characteristics (Study 2), and match third-party administrator experts in 78% of the cases while offering novel ways to reduce claim leakage when deviating from their assessment (Study 3).

Table 1: Summary of Related Studies

Study 1

As LLMs increasingly enter insurance operations, a critical first question is whether they can match or exceed human performance in core insurance claim assessment tasks. Study 1 addresses this question by benchmarking LLM-based claims handling against human insurance professionals in a standardized, real-world assessment environment. This evaluation serves two essential purposes: (1) The study establishes a validated performance baseline for LLM capabilities relative to industry standards. (2) The exams cover a range of insurance claim domains, allowing the identification of potential domain-specific strengths and limitations across insurance tasks. We utilize standardized insurance clerk examination materials—the same tests used to certify human professionals—to create a direct, objective comparison of performance.

Method

In this study, we evaluated the performance of LLM-based claims handling bots against a professional industry benchmark. We utilized the claim cases from the insurance clerk final examination administered by the German Chamber of Industry and Commerce, which serves as a qualification standard for roles in insurance sales and claims handling in Germany (for a case overview, see Web Appendix A1). The examination consists of two parts: (1) Economics and Social Studies, and (2) Customer Needs Analysis and Insurance Claims Handling. For Part 2, examinees select one of six insurance domains with three cases each: accident, health, liability, life, motor, or property. We focused specifically on the claim cases from Part 2 of the Summer 2024 examination, excluding one health case that involved a customer needs analysis rather than claims processing, resulting in 17 claim-related exam cases. We obtained the examination materials, which are available for purchase, from the German Chamber of Industry and Commerce. The examination materials comprised the original exam questions, sample solutions, and examination statistics. First, we automatically translated all 17 claim-related exam cases and material (including customer data, customer request, claim event details, and policy documentation) from German to English using the DeepL API. We performed this translation to ensure consistency and comparability across studies in this research,

and as reasoning-ability has been benchmarked especially based on English language corpora (OpenAI, 2023a). Second, we developed a LLM-based claims handling bot using the "messages" endpoint of the OpenAI API with GPT 4o (gpt-4o-2024-08-06) and GPT o1 (o1-preview-2024-09-12), given their strong demonstrated reasoning capabilities and broad market coverage (Wei et al., 2022). Our prompts included a structured task description, policyholder and claim data, and relevant insurance policy contents, requesting the LLM to develop comprehensive sample solutions with their reasoning process and a recommended reimbursement amount or a list of amounts in case of several losses (see Web Appendix A2 for the complete prompt skeleton). To address potential selection issues associated with analyzing single LLM responses (e.g., Chen et al., 2023), we sent every prompt five times to the API and retrieved valid responses for all prompts.

Next, we evaluated the responses generated by the LLM-powered claims handling bots using the official sample solutions provided by the German Chamber of Industry and Commerce. These sample solutions specify which aspects must be addressed in each claim case and how points are allocated. While examiners can award additional points for relevant content not explicitly listed in the sample solution, we performed a more conservative assessment without such additional points and limited the point distribution only to the predefined criteria. It is worth noting that the Chamber's predefined criteria and the associated points, as well as additional point-eligible answers, allow examinees to achieve points beyond the maximum points specified in each sub-task, but are still limited to the maximum point amount.

Finally, we compared the performance of the LLM-powered claims handling bots with the exam performance of all examinees (i.e., insurance clerks) taking the exam in Germany in Summer 2024 (N = 1,778 examinees). Following open science practices (Fišar et al., 2024), we made all sharable data and code available on our Open Science Framework repository (see https://osf.io/2ufr5/?view_only=53699dc9a4a544a68aaf3057757a1526). Due to copyright and confidentiality restrictions of this particular study, we share only the translated claim cases and expected solution steps provided by the German Chamber of Industry and Commerce.

Results

As shown in Figure 1, GPT o1 significantly outperformed the human benchmark (indicated by a horizontal dashed line), achieving an average score of 82.44% (SD = 12.05%) compared to the mean human examinee score of 73.98% ($t(29) = 3.85$, $p < .001$). This translates to a grade B performance for GPT o1, placing it in the upper quartile of human examinees. In contrast, GPT 4o with lower reasoning capabilities scored 65.56% (SD = 15.69%), which corresponds to a grade D performance and falls slightly below the human benchmark ($t(29) = -2.94$, $p < .001$).

Performance varied across insurance domains, with both models showing a strong performance in liability and accident cases. These findings suggest that LLMs excel in domains that are more structured, text-driven, and rule-based, such as liability and accident claims. These cases often rely on written evidence and involve standardized documents, established legal procedures, and preset compensation rules, making them easier to process algorithmically.

In contrast, claims in health, life, motor, or property insurance often require deeper domain expertise to evaluate medical histories, technical damage, or complex causality, making them harder to evaluate correctly for LLMs. GPT o1 demonstrated particularly strong performance in motor and property insurance claims, consistently exceeding human benchmarks in these technically complex domains. Health insurance claims presented challenges for GPT 4o, with some responses falling into the F grade range, while GPT o1 maintained a less varied, more consistent performance across all domains. The significant performance advantage of GPT o1 likely stems from its enhanced reasoning capabilities, which enables the model to follow and evaluate more complex causality chains.

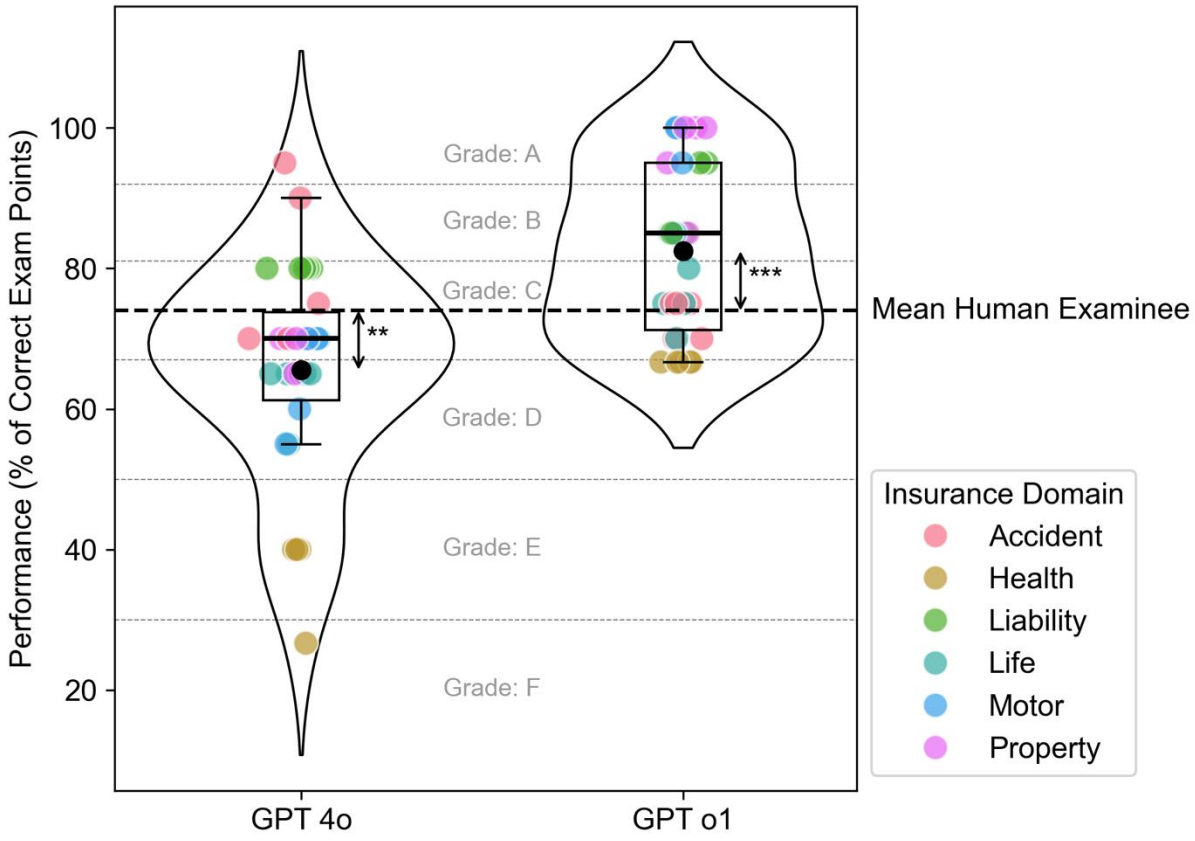

Figure 1: LLM-powered Claims Handling Bots German Insurance Clerk Final Exam Performance

From an efficiency perspective, it is worth highlighting that both models processed claims with remarkable speed (M_{4o} = 9.51s/claim, M_{o1} = 62.14s/claim) and cost-efficiency (M_{4o} = \$.10/claim, M_{o1} = \$1.05/claim). While the original exam did not track the completion time for single questions, the average time allocated by the Chamber to solve one exercise was 11 minutes and 40 seconds and would have costed \$2.0[2] per examinee, assuming the salary for an insurance clerk in the final year of apprenticeship.

[2] Assuming 52 weeks/year, 35h/week, USD/EUR 1.08, and a gross salary of €1,369 per month in the final apprenticeship year (Bundesinstitut für Berufsbildung, 2024)

Discussion

The results of Study 1 demonstrate that LLM-powered claims handling bots can evaluate insurance claim cases well enough to pass the German insurance clerk final exam and can even outperform the average score of insurance clerks (i.e., a GPT o1-powered claims handling bot with proper reasoning abilities). Given this performance, LLM-powered claims handling bots have significant potential to augment human insurance clerks, as they are not only fairly accurate but also efficient with respect to their speed and cost-efficiency.

It is worthwhile noting that the performance results reported in Study 1 are conservative as human examinees received additional points not only for the correct final reimbursement decision but also for the correctness of the solution steps. In contrast, the LLM-powered bot often reported the correct final decision with fewer decision steps and without following the claims-processing blueprint that insurance clerks learn during their apprenticeship.

In the next Study, we systematically investigate whether LLMs can consistently make accurate final reimbursement decisions regardless of intermediate reasoning steps. Ensuring that these decisions are both correct and free from biases toward the policyholder is a fundamental requirement for deploying LLMs in claims handling.

Study 2

The objectives of Study 2 were twofold: First, this study evaluates LLMs' ability to process insurance claims "end-to-end" (i.e., from initial claim submission through evaluation, decision-making, and communication of the settlement), using simulated claim cases that reflect real-world industry settings. Each case includes detailed information on the policyholder, the claim event, and the relevant insurance policy. Second, the current study assesses the presence of potential biases by claims handling LLMs as a function of a policyholder's gender, ethnicity, age, and socioeconomic status. Varying these experimental factors independently is important, given that females (vs. males), minorities (especially non-white population), and people with lower (vs. higher) socioeconomic status have been shown to receive

systematically lower reimbursement offers in traditional claims handling (as summarized in our theoretical background section).

Method and procedure

We prompted two LLMs (Claude Sonnet 3.5, GPT 4o) to evaluate simulated (but realistic) claim cases across four insurance domains (casualty, health, life, and property). We selected these LLMs since they are the best performing models on benchmark tests and frequently used in practice at the time conducting the study (we didn't use GPT o1 for conversative testing and cost-efficiency reasons). For each domain we had one claim case of low and one claim case of high complexity (see Web Appendix B1 for the cases). Low complexity cases are characterized by the following criteria: a single claim item to be processed (e.g., a stolen gold necklace or a single medical service), involvement of only one individual for casualty cases, and the inclusion of 1 to 3 supporting documents (such as the policy file, original receipt) that the LLM must reference to handle the claim. High complexity cases are defined by the presence of multiple claim items, involvement of several individuals for casualty cases, and the need to analyze 3 to 6 supporting documents for the LLM to effectively handle the claim. Additionally, we also varied the policyholder characteristics of gender (female vs. male), ethnicity (Black, Hispanic, White), age (21 yrs, 36 yrs, 51 yrs, 66 yrs, 81yrs), and socioeconomic status (low, medium, high) through the policyholder's name, birthdate, and address for each claim case.[3] This resulted in 90 different experimental conditions for each claim case (8), which we each prompted five times to every LLM (4) to control for single response selection bias. In total, we therefore generated a total of 7,200 prompts (8 claim cases × 2 LLMs × 5 repetitions × 90 policyholder variations).

We built an LLM-powered claims handling bot for each technology provider using the "messages" endpoint of the Claude API (claude-3-5-sonnet-20241022) and OpenAI API (gpt-4o-2024-08-06). We

[3] While inferring attributes such as ethnicity solely from a name is inherently biased, our approach mirrors established human decision-making processes as documented in prior research (Bertrand & Mullainathan, 2004; Lockhart et al., 2023).

automatically queried each LLM to evaluate a policyholder's claim, determine a claim reimbursement amount, and write a corresponding reimbursement letter for the policyholder in our prompt (see Web Appendix B2 for the prompt skeleton). Analogous to Study 1, we employed chain-of-thought prompting. The prompt provided the task description, policyholder and claim details, and the relevant insurance policy content to the bots to ensure an informed claim evaluation was possible. All claim cases required the claims handling bot to consider all policyholder data (i.e., insured person and insured risk details), claim data (including the policyholder's claim request, invoices, and optional photos), and the corresponding policy file contents to solve the case correctly. We asked the LLMs to respond with the reimbursement amount and a customer-facing reimbursement settlement letter. The claims handling bots responded in all 7,200 cases with a valid response. Next, the LLM-powered claims handling bot responses were analyzed (for an exemplary response, see Web Appendix B3). The methodological procedure and data wrangling processes are summarized in Figure 2.

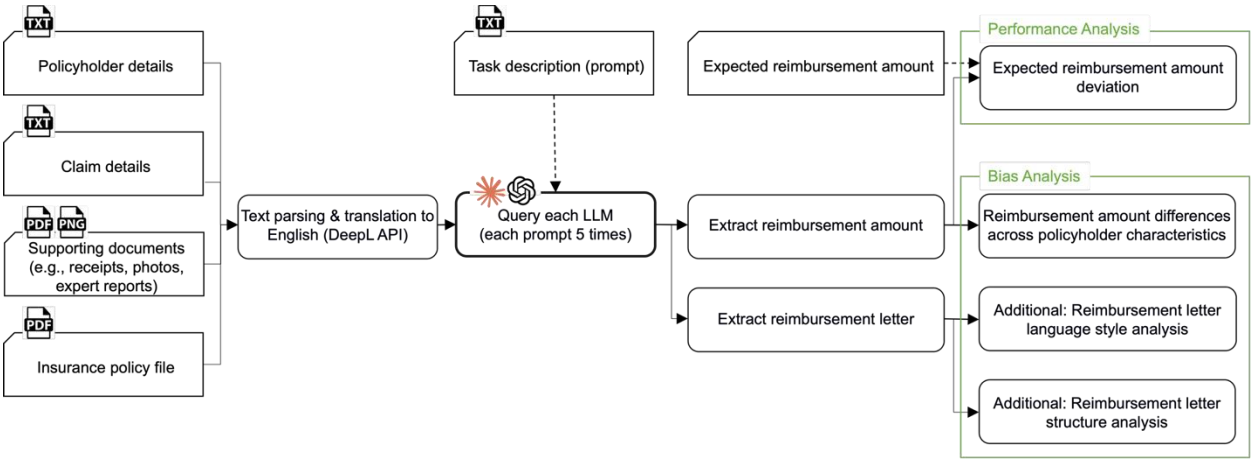

Figure 2: Summary of Methodological Approach (Study 2)

First, we conducted a performance analysis comparing the LLM-powered claims handling bots' reimbursement amount proposals with the expected reimbursement amount, which was determined by an

insurance professional, for the corresponding claim case. Invalid bot responses were classified as offering incorrect reimbursement amounts.

Second, we conducted a bias analysis testing whether the claim reimbursement amount offered by the LLM to the policyholder differed depending on policyholder characteristics (i.e., gender, ethnicity, age, and socioeconomic status). Therefore, we calculated the relative excess reimbursement amount (i.e., the absolute claims handling bot's reimbursement amount subtracted by the expected reimbursement amount divided by the expected reimbursement amount) for each claim case. Additionally, we tested whether LLMs exhibit biases through structural (i.e., length of individual reimbursement letter parts, letter length) and language style (i.e., readability, perspective taking, presence of providing a rationale) claim reimbursement letter differences across policyholder characteristics. Specifically, for the language style analysis, we automatically queried GPT 4o (gpt-4o-2024-08-06; via the OpenAI API; for the prompt skeleton see Web Appendix B7) to rate the emotionality (1: low emotional responding - 7: high emotional responding), perspective taking (1: low perspective taking - 7: high perspective taking), rational (1: low explainability - 7: high explainability), and readability (1: low readability - 7: high readability) of the proposed reimbursement letters. Additionally, we calculated the word count of the responses and the Flesch Kincaid readability score. The structure analysis involved splitting the letters into sentences, calculating the cosine similarity of the sentence embeddings using the SentenceTransformer (Reimers & Gurevych, 2019 and clustering similar sentences with k-means to structural parts. To gain a deeper understanding of the topic of each identified structural part, we ran a topic model on each cluster. For each identified structural part, we also counted the words in the reimbursement letter.

Additionally, we applied exactly the same methodological procedures and analyses for the best-performing open-weight LLMs at the time conducting this study, Llama 3.1 405B (llama-3.1-405) and Pixtral 12B (pixtral-12b-2409). However, the results showed that these open-weight LLMs cannot offer consistent performance in such complex insurance tasks and cannot keep up with the best proprietary LLMs GPT 4o and Sonnet 3.5, which is why we will only focus on these models for the remainder of this paper (for results and details on the open-weight models see Web Appendix B4 - B9).

Pre-test

We first conducted a pre-test to evaluate whether the LLM-powered claims handling bots can correctly infer gender, ethnicity, and socio-economic status merely from the policyholder's name, age from the policyholder's birthdate, and socio-economic status from the home address. This is crucial because claim cases rely solely on basic policyholder information (besides other, name, birthdate, and home address) from the contract information. As in the main experiment, we automatically queried the LLMs (claude-3-5-sonnet-20241022 & gpt-4o-2024-08-06) and asked to classify the person's gender (male vs. female), ethnicity (white vs. black vs. Hispanic), age, and socioeconomic status (1: low socio-economic status - 7: high socio-economic status) based solely on the contract information (see Web Appendix B9). Our results revealed that both LLMs can correctly infer the gender and ethnicity from the name in 100% of the cases. For age, GPT was off by only one year and Claude Sonnet 3.5 by two years, depending on the training data cutoff date. Both models can correctly infer the socio-economic status from the home address in 100% of the cases.

These findings provide evidence that LLM-powered claims handling bots can accurately infer gender and ethnicity from the name, but minimal errors for inferring age from the birthdate. Further, we find evidence that LLMs can infer the socioeconomic status merely from a person's home address.

Results

Performance Analysis. Our findings indicate that LLM-powered claims handling bots correctly determine the reimbursement amount in over 92% of cases across all insurance domains (see Figure 3, Panel A). The accuracy varies by domain, with the highest performance in life (100%) and health insurance (98.8%), while slightly more assessment errors in property (94.5%) and casualty insurance (92.8%). Additionally, our results show that error rates increase with claim complexity, with accuracy declining slightly from 97.9% for low complexity claims to 95.1% for high-complexity claims (see Figure 3, Panel B).

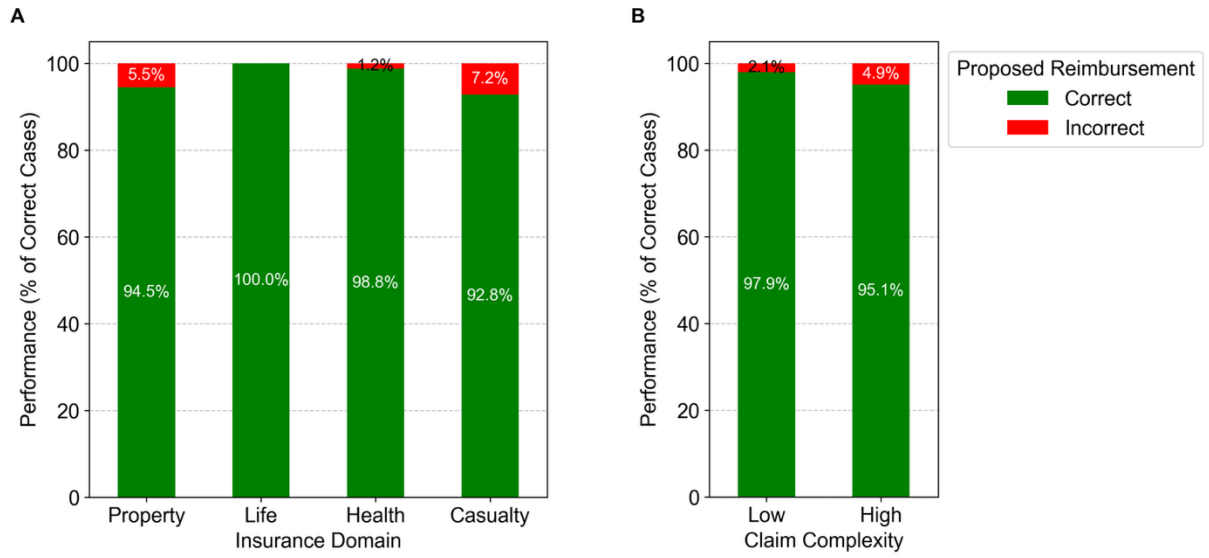

Figure 3: Proprietary LLM-powered Claims Handling Bots Performance by Insurance Domain (Panel A) and Claim Complexity (Panel B)

Additionally, we find that the LLM-powered claims handling bots performed both fast ($M_{ClaudeSonnet3.5}$ = 16.56 s/prompt, $M_{GPT\ 4o}$ = 8.14 s/prompt) and cost-efficient ($M_{ClaudeSonnet3.5}$ = .12 $/prompt, $M_{GPT\ 4o}$ = .08 $/prompt).

Bias Analysis: Relative Excess Reimbursement Amount. Holm-corrected contrasts for multiple comparison revealed no or negligible small evidence of bias in LLM-powered claims bots based on policyholder characteristics (for detailed results and contrasts, see Web Appendix B6).

Bias Analysis: Structure & Language Style. We conducted ancillary structure and language style analyses on the reimbursement letters to better understand whether LLM-powered claims handling bots change the structure or language in terms of length (word count), readability, emotionality, perspective taking, and rationale, depending on the policyholder's gender, ethnicity, age, and socio-economic status. We performed these additional analyses to assess to what extent a bot adapts its language style to the person receiving a letter.

Overall, we find a very high level of readability (i.e., simple and easy to proceed language structure; 6.56 out of 7), low emotionality (i.e., low arousal and more controlled language style; 3.15 out of 7),

medium perspective taking (i.e., signals of mutual understanding; 4.28 out of 7), and a high presence of offering a rationale for the reimbursement decision (i.e., explaining in the letter the basis for the reimbursement amount; 6.15 out of 7). Our results show that LLMs are professional in their language when responding to policyholders. As detailed in Web Appendix B7, we find no significant or negligibly small differences in terms of word count, readability, emotionality, perspective taking, and offering a rationale across the policyholder characteristics for the LLMs. However, in the proposed reimbursement letters, LLMs adjusted their language for high-complexity claims compared to low-complexity ones by enhancing readability while reducing the emotionality, perspective-taking, and rational content. While Sonnet 3.5 slightly increased the word count for more complex claims, GPT 4o slightly reduced it.

When it comes to the reimbursement letter structure, we identified structural parts describing the *reimbursement calculation* (e.g., "This amount reflects the application of your $1,500 deductible and 25% coinsurance for covered services"), offering *claim support* (e.g., "If you have any further questions or require additional assistance, please do not hesitate to contact us."), describing the *claim process* (e.g., "Please note that this claim will be recorded on your policy history."), and the *reimbursement payment* (e.g., "We will reimburse you for the damages to your vehicle in the amount of $950."). We find that LLMs use most of their words for describing the *claim process* (43.23%), followed by offering *claim support* (24.13%), providing a *reimbursement calculation* (21.64%), and finally offering a *reimbursement payment* (11%) (for details, see Web Appendix B8). We only find a minor ethnicity bias in the reimbursement letter structure of letters generated by Claude Sonnet 3.5. Specifically, our results show that Claude Sonnet 3.5 described the *claim process* less (fewer words) for Black policyholders compared to White and Hispanic policyholders. All other effects were non-significant or small.

Discussion

Study 2 provides causal evidence that LLM-powered claims handling bots are efficient at handling insurance claims based on a task description, policyholder and claim details, and the corresponding policy file across a large number of tasks and complexity levels. Our results show that LLMs, do not show typical

(human) biases previously reported in the literature (see Table 1 for a summary, such as implicit biases based on a consumer's gender, ethnicity, age, or socioeconomic status; Baker & McElrath, 1996; Becker & Newsom, 2003; Doerpinghaus et al., 2003, 2008; Lin et al., 2022; Scott et al., 2023) except for a negligible ethnicity bias in the reimbursement letter structure of Claude Sonnet 3.5.

Study 3

Study 1 established that LLMs can match or exceed human performance on standardized insurance clerk examination cases, providing an initial performance benchmark against industry certification standards. Study 2 extended the current results through a large-scale multi-factorial experiment, demonstrating that LLMs deliver consistently accurate and relatively unbiased claim assessments across a broad range of claim tasks and policyholder demographics. The objective of Study 3 is to directly compare LLM performance relative to professional claims handling experts who have already assessed real insurance cases. The study examines precisely where LLM assessments align with those of experienced claims third-party administrators (TPAs) and—perhaps more importantly—where and why they differ. The implications of this study and comparisons have therefore direct implications for the efficiency of insurance operations, as insurers could auto-approve claims in cases where the assessments of the claims TPA and the LLM-powered bot align. This study also illuminates which specific factors drive assessment differences between the human claims handling experts and the LLMs.

Method

In this study, we prompted the LLM-powered claims handling bot relying on GPT 4o (gpt-4o-2024-08-06) to evaluate 53 real claim cases across the insurance domains, personal liability, professional liability, household, and homeowner insurances provided by a claims TPA (see Web Appendix C1 for an overview of the case structure). These actual claims happened in Germany and have been covered by different insurance companies. The claims TPA processed and evaluated these claim cases as an outsourcing provider for insurance companies providing a reimbursement decision proposal, which is a common practice in the

insurance industry. The reimbursement proposal of a TPA consists of five key assessments: (1) whether the claim is plausible, (2) whether the policyholder is liable (for liability insurances only), (3) whether gross negligence by the policyholder is present, (4) whether the claim should be paid out, which includes an insurance policy coverage check, and (5) whether recourse is possible. Whether the claim should be paid out (4) depends on the assessments (1) to (3) and the insurance policy coverage check. The determination of recourse does not influence the claim reimbursement decision, it serves as an additional assessment typically included in the reimbursement decision proposal conducted by a claim handler (in this case, from the TPA). The claim handler performs these assessments based on the detailed claim case data (including the insured object, residential structure, damage event, cause of damage, and extent of damage) and the corresponding wording (see Figure 4 for an overview of the methodological approach).

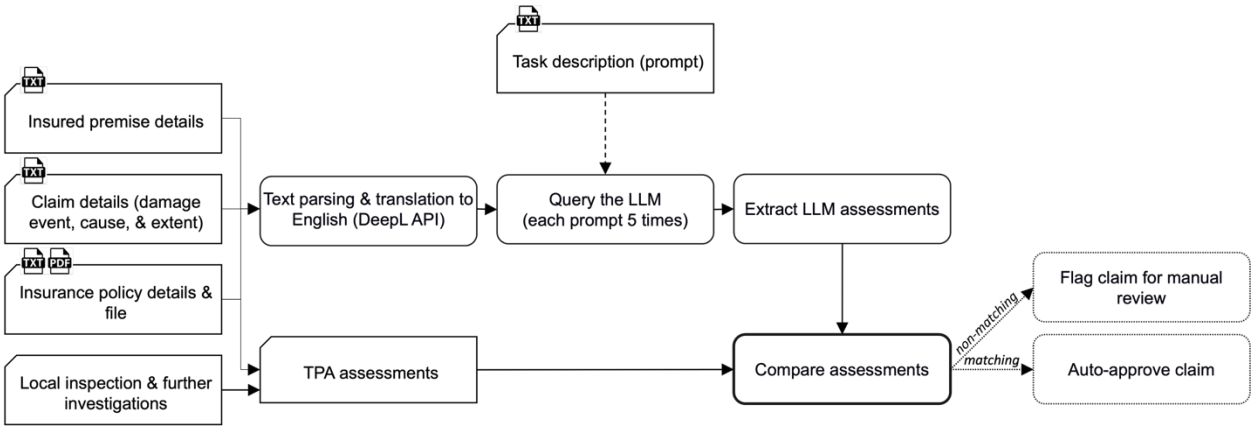

Figure 4: Summary of the Methodological Approach (Study 3)

As in Study 1, we automatically translated all 53 claim cases and material from German to English using the DeepL API. We performed this translation to ensure consistency and comparability across studies in this research. Analogous to the claim handler, we then prompted the LLM-powered claims handling bot with a task description, the detailed claim case data, and the corresponding policy file to conduct the same five reimbursement proposal assessments (see Web Appendix C2 for the prompt skeleton). As in Studies 1 and 2, we employed line-of-thought prompting and sent every claim case five times for assessment to the claims handling bot to control for single response selection bias. We received a valid response for all

prompts. Next, we compared the assessments of the reimbursement decision proposals between the LLM-powered claims handling bot and the claim handler. We also flagged and further analyzed also non-matching assessments. To gain a deeper understanding of why claims have been flagged, we identified and coded the assessment error source and assessment error reason. The identification of the error source and reason was done by an insurance professional conducting a case-by-case analysis analyzing the wording and the assessment of the professional claims handling expert and the LLM-powered claims handling bot as it should be done for flagged claims. Due to copyright and confidentiality restrictions, we are only allowed to share the evaluations but no claim case details of the insurers. All shareable code and data are available in our OSF directory.

Results

The claims handling bot closely aligned with the professional human claim handler assessments in 94.3% of cases for claim plausibility, 90.6% for liability, and 88.7% for gross negligence (see Figure 5). Since the final claim reimbursement decision (i.e., is the claim reimbursed: yes/no) depends on these three evaluations (plausibility, liability, gross negligence), besides the policy coverage, its maximum claim handler matching rate is constrained by the lowest match (88.7%). Ultimately, the bot aligned with the claim handler's final reimbursement decision in 77.7% of claim cases, indicating greater deviation from the claim handler's final reimbursement decision. As a side result, we find that the recourse assessments aligned in 86.4% of the cases.

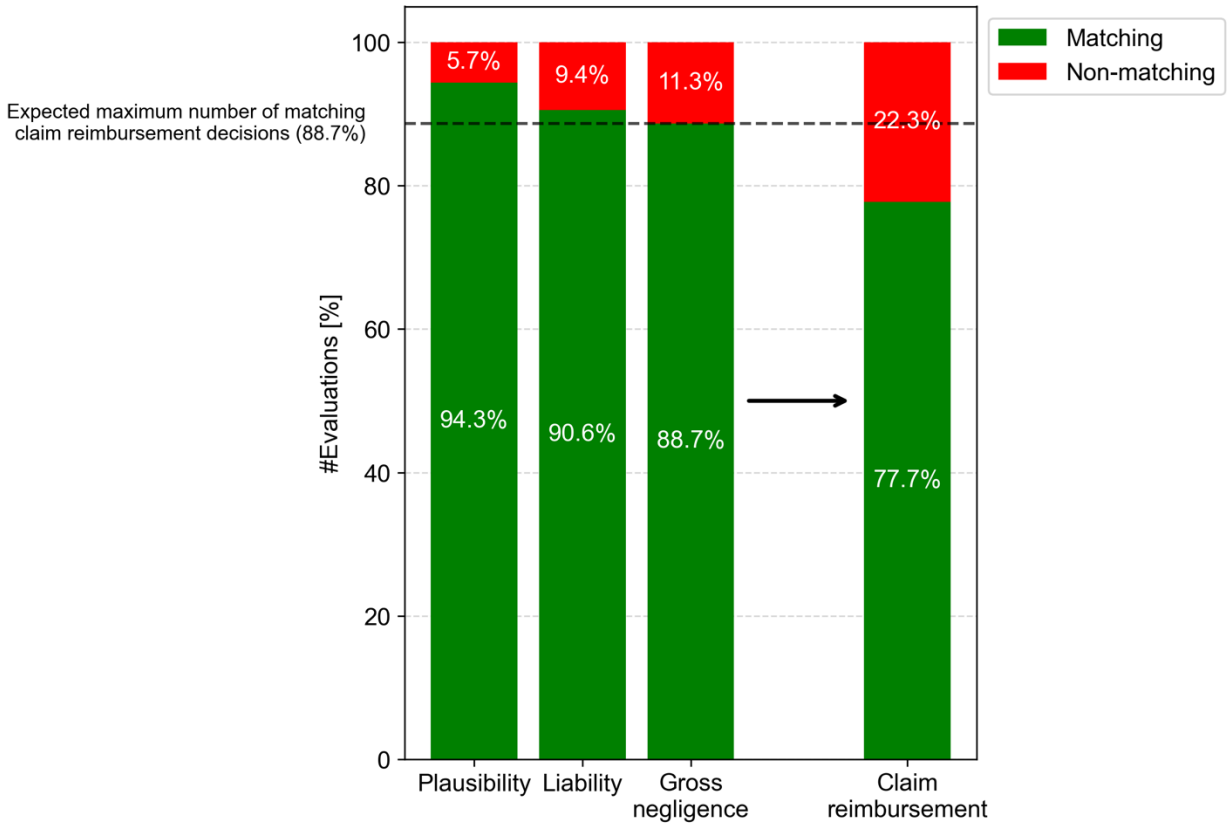

Figure 5: GPT 4o-powered Claims Handling Bots Claim Case Evaluations

After coding each discrepancy based on the available data and with the expertise of an insurance professional, we found that 57.76% of assessment errors were attributable to the LLM, 30.43% to the claim handler, and 11.80% could not be clearly assigned (see Figure 5). Specifically, for the final claim reimbursement decision, our results show that 50.91% of assessment errors can be attributed to the LLM, 40% to the claim handler, and 9.09% could not be clearly assigned. When the LLM erred, it *wrongly denied* a reimbursement payment in 82% of cases, whereas the human claim handlers *wrongly granted* a reimbursement payment in 100% of their assessment errors, indicating the potential cost implications and relative risks of human compared to LLM-based claims handling.

As depicted in Figure 6, among all LLM-related discrepancies, 29.81% resulted from a lack of context: in many cases, the claim handler included crucial information only in the assessments (i.e., information that is only available to the human claims TPA based on local inspection and further

investigation of the claim that is only shared by the TPA with the insurer in the assessments and therefore not available to the LLM for its assessments). The remaining 27.95% stemmed from either misinterpretation of the situation or ambiguous wording. For example, one policy excluded only losses from "flowing or standing waters," and although the actual damage was caused by rain, the LLM denied payment, likely based on a misreading of this clause. As for the claim handler, 8.7% of total errors were due to inconsistencies in its recommendations. One such case involved household damage from backflow caused by adverse weather, where the insured had deactivated the backflow prevention device, directly contributing to the loss. While both the LLM and the claim handler agreed this constituted gross negligence, the claim handler still recommended payment without providing a clear justification, whereas the LLM (correctly) denied the claim on the grounds of insufficient care. The remaining 21.74% of claim handler discrepancies were due to misinterpretation of wording. These assessment errors often stem from misinterpreting insurance conditions, leading to incorrect coverage decisions. In one case where damage occurred to a house due to flooding, although the policy clearly excluded damages caused by flooding or precipitation, the human claim handler recommended approval, while the LLM again correctly denied the claim. In 11.8% of all discrepancy cases, no clear attribution was possible—either because the data did not allow for a conclusive interpretation (e.g., vague terms like "gross negligence") or because the policy wording itself was unclear.

Taken together, these results suggest that if claims TPAs provided the insurer with full documentation from local inspections and investigations, rather than only including this information in their final assessments, the error rate attributable to the LLM would fall substantially and would place it below the error rate of the human claims handler (30.43%) (unlike the human claims TPA, note that the LLMs didn't produce errors caused by inconsistencies in their assessment based on existing policies). Further, the assessment errors due to misinterpretation could be decreased for LLMs by providing the policies as text directly (in the current work, we had to parse the text from PDF documents, which could have resulted in more misinterpretations since breaks cannot always be restored perfectly).

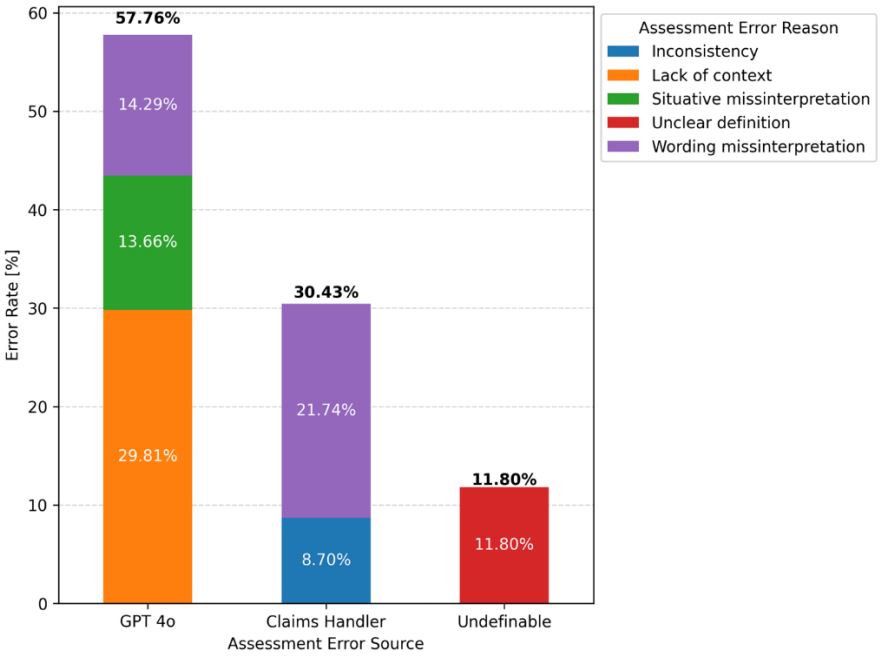

Figure 6: Claim Assessment Error Rate by Assessment Error Reason and Source

Finally, it is worth noting again that our LLM-powered claims handling bot was both fast ($M_{GPT\ 4o}$ = 8.59s/claim case) and cost-efficient ($M_{GPT\ 4o}$ = $.04/claim case) to perform the reimbursement assessment. By comparison, a human claims handling TPA typically might take between 15-30 minutes depending on the complexity of a case to assess a similar claim case, with an estimated minimum labor cost of $5.3[4] per claim based on current salary agreements in the German insurance industry (excluding the time and costs incurred in each case for the local inspection and further investigation of the claim). For complex claims requiring multiple reviews or specialized expertise, human processing times can also easily extend to several hours, of course, further widening the efficiency gap between LLM and human-based assessment.

[4] Assuming 52 weeks/year, 35h/week, USD/EUR 1.08, and a gross salary of €2,974 per month in the first working year (Bundesinstitut für Berufsbildung, 2024)

Discussion

Study 3 demonstrates that LLM-powered claims handling bots can effectively automate insurance claims processing. When the bot's evaluation matches the claim handler's reimbursement decision proposal, which occurs in 77.7% of cases in our dataset, the claim could be automatically approved. In cases of discrepancy, the claim is flagged for manual review by a human (senior) claim handler from the insurer. This hybrid approach would have the potential to massively reduce processing time and costs for insurers and policyholders but also minimizes claim leakage by preventing erroneous claim handler decisions that might otherwise have been approved.

General Discussion

The current research is the first systematic, causal assessment of the actual performance and biases of LLMs for automated insurance claims handling. Across three studies, we demonstrate that LLM-based claims handling bots consistently deliver a competitive or superior performance relative to human claims handling professionals, exceeding in industry exam benchmarks without being specifically trained for them (Study 1), providing unbiased claim assessments across a broad range of claim cases and policyholder characteristics (Study 2), and matching the assessment of around 78% of the cases of extern claims handling experts, while offering novel ways to reduce claim leakage when deviating from their assessment (Study 3). These results suggest significant opportunities for enhancing both operational efficiency across the insurance value chain, with processing times of just seconds (or around a minute for very complex cases involving multiple pdf documents and images as in Study 2) versus hours or days (in the case of humans) and costs of under $1 per claim. The current results also demonstrate that reimbursement proposals are independent and largely unbiased from policyholder demographics or socioeconomic status, in contrast to biases revealed with human claims handling experts in prior work (Baker & McElrath, 1996b; Becker & Newsom, 2003b; Doerpinghaus et al., 2003, 2008; Lin et al., 2022; Scott et al., 2023b). In what follows, we discuss the theoretical, practical, and methodological implications of this research and its impact on

policyholders, insurers, InsurTech firms, and public policy for future research on the impact of LLM-powered claims handling in the insurance ecosystem.

Theoretical, practical, and methodological implications

The findings of this research offer a new theoretical, practical, and methodological perspective and contribution assessing the impact of LLMs on insurers, policyholders, and the broader insurance ecosystem. The key insights of the research highlight the applicability of LLMs in insurance practice, how they exhibit biases, and perform efficiently and accurately in claims handling.

From a theoretical standpoint, the current findings contribute to research on both LLM performance and biases in comparison to human claims handling experts. Our research shows that current LLMs can solve complex, not specifically trained tasks in the claims handling context at least as good as humans. Moreover, our findings significantly advance AI-related research in claims handling by extending it to the potential of fully end-to-end applications of LLMs (from claim assessments to actual reimbursement proposals). While earlier studies demonstrated the potential of AI systems to outperform traditional methods in tasks like classifying claims (Dimri et al., 2019) or of LLMs to conduct basic claims inconsistency detection (Li et al., 2025), the current work provides a systematic and causal assessment across major claims handling tasks and against both industry benchmarks as well as human claims handing professionals.

From a practical perspective, this research offers a more nuanced lens for insurance companies, InsurTech firms, and regulators to rethink claims handling processes and solve current challenges in the insurance industry practice. The findings demonstrate that LLMs perform on par with, if not better than, human claim handlers in terms of efficiency and accuracy while exhibiting only minimal bias. Therefore, LLMs could augment human claim handlers by generating claim assessment drafts and prioritizing claims provided by other claim handlers or TPAs for a second analysis. Consequently, insurers stand to achieve substantial profitability gains by lowering loss administrative costs (average costs per claim in Study 2: \$.07) and increasing the speed (time per claim in Study 2 for GPT 4o: 8.14s) as well as a reduction in claim leakage by accurate claims handling. Besides increased profitability, adopting LLMs enables insurers to

improve customer satisfaction by delivering prompt and consistent responses. In Study 2, GPT 4o needed 8.14 seconds to handle the claim, a speed unmatched by human handlers. Beyond operational efficiency and customer satisfaction, integrating LLMs into claims handling also addresses long-standing issues of human biases. Prior research (see Table 1) has highlighted various biases in human claims handling. By minimizing subjectivity, LLM systems promise fairer and more regulatory-compliant claims handling, ensuring that policyholders are treated mostly equitably regardless of their demographics or background.

Finally, from a methodological standpoint, this research integrates exams, simulations, and real-world cases to create an approach that can be replicated by others for evaluating LLMs for potential biases and their performance in claims handling. By employing factorial experiments, it systematically investigates the impact of key policyholder features—such as gender, ethnicity, age, and socioeconomic status, previously studied in other works—on claim outcomes. This approach offers a thorough method for assessing algorithmic bias. Furthermore, the research includes a systematic comparison to a large-scale insurance clerk exam, providing a solid human benchmark for performance evaluation.

Future research directions

In what follows, we outline directions for future research on how LLMs may impact automated claims handling across various stakeholders, including academic researchers, policyholders, insurers, InsurTech companies, and regulators (see Table 2). From an academic research perspective, we believe more work is needed to better understand the effects of LLMs on attributions towards the insurer (e.g., in terms of perceived welfare, level of trust) as well as the willingness to accept the reimbursement offer using LLMs in claims handling. From a policyholder perspective, future work may zoom in further to better understand how the specific language style and responses of these systems can produce positive effects on reimbursement acceptance, claims handling satisfaction, and ultimately, more positive firm evaluation. As insurers along with InsurTech firms must balance multiple systems across the insurance value chain, the question remains how to best integrate these systems to maximize firm and policyholder value. Additionally, insurers and InsurTechs need to test the possibility of full automation of the claims handling

process with LLMs and then ensure accountability when such systems start to make decisions fully autonomously (as in the case of Lemonade outlined at the outset of the paper). Future research could therefore further assess the optimal combination of human-AI collaboration throughout the claim's lifecycle, identifying specific handoff points where human judgment adds the most value while allowing automation to handle routine assessments. Finally, regulators face the difficult task of developing policies that safeguard policyholders and offer high levels of consumer protection, but at the same time can shape an environment that ensures—or even fosters—innovation and competitiveness in the insurance industry. While some regions, such as the European Union and the United States, have begun drafting relevant guidelines (e.g., The European Parliament, 2024; The White House, 2022), a collaborative approach integrating industry, public policy, and academic research will arguably be most fruitful in creating a regulatory landscape that ensures public trust and encourages (instead of hinders) the responsible adoption of AI across the insurance value chain.

Stakeholder	*Exemplary research questions*
Academic research	• How does the use of LLMs in insurance claims handling influence the attribution towards the insurer and the claim acceptance rate over time? • Where do policyholder appreciate vs. object algorithmic decisions, and which factors moderate such acceptance or rejection decisions?
Policyholder	• Which factors, such as language style and response time, influence whether consumers accept a reimbursement offer? • How does the use of LLMs impact long-term customer satisfaction, loyalty, and behavior?
Insurer	• How can insurers assure the ethical use of LLMs in their claims handling? • How can insurers ensure that LLMs are transparent and accountable in their decision-making processes?
InsurTech firm	• What are the key barriers to the adoption of LLM-based claims handling solutions among traditional insurance firms, and how can they be overcome? • What metrics and KPIs should be developed to evaluate the performance and fairness of LLM-driven claims handling systems?
Regulator	• What regulatory frameworks are needed to protect the consumer from potentially biased LLM-powered claims handling bot responses? • How can the use of LLMs be effectively monitored?

Table 2: Future research directions based on the stakeholder perspective

References

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., … Vinyals, O. (2024). *Gemini: A Family of Highly Capable Multimodal Models*. http://arxiv.org/pdf/2312.11805v2

Anthropic. (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku*.

AXA. (2023). *AXA offers secure Generative AI to employees: Press release*. https://www.axa.com/en/press/press-releases/axa-offers-securegenerative-ai-to-employees

Baker, T., & McElrath, K. (1996a). Whose Safety Net? Home Insurance and Inequality. *Law & Social Inquiry*, *21*(02), 229–264. https://doi.org/10.1111/j.1747-4469.1996.tb00080.x

Baker, T., & McElrath, K. (1996b). Whose Safety Net? Home Insurance and Inequality. *Law & Social Inquiry*, *21*(02), 229–264. https://doi.org/10.1111/j.1747-4469.1996.tb00080.x

Balona, C. (2024). ActuaryGPT: Applications of large language models to insurance and actuarial work. *British Actuarial Journal*, *29*, e15. https://doi.org/10.1017/S1357321724000102

Becker, G., & Newsom, E. (2003a). Socioeconomic status and dissatisfaction with health care among chronically ill African Americans. *American Journal of Public Health*, *93*(5), 742–748. https://doi.org/10.2105/ajph.93.5.742

Becker, G., & Newsom, E. (2003b). Socioeconomic status and dissatisfaction with health care among chronically ill African Americans. *American Journal of Public Health*, *93*(5), 742–748. https://doi.org/10.2105/ajph.93.5.742

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bousquette, Isabelle. (2025). Turns Out AI Is More Empathetic Than Allstate's Insurance Reps. *The Wall Street Journal*. https://www.wsj.com/articles/turns-out-ai-is-more-empathetic-than-allstates-insurance-reps-cf5f7c98?utm_source=chatgpt.com

Bundesinstitut für Berufsbildung. (2024). *Tarifliche Ausbildungsvergütungen 2024 in Westdeutschland*. https://www.ihk-muenchen.de/ihk/documents/Berufliche-Bildung/Ausbildungsberatung/1_BiBB-Tabelle_2024-2.pdf

Chen, Y., Andiappan, M., Jenkin, T., & Ovchinnikov, A. (2023). *A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?* (SSRN Scholarly Paper 4380365). https://doi.org/10.2139/ssrn.4380365

Cognizant. (2018). *Property & Casualty: Deterring Claims Leakage in the Digital Age*. https://www.cognizant.com/whitepapers/property-and-casualty-deterring-claims-leakage-in-the-digital-age-codex3332.pdf

Courbage, C., & Nicolas, C. (2021). Trust in insurance: The importance of experiences. *Journal of Risk and Insurance*, *88*(2), 263–291. https://doi.org/10.1111/jori.12324

de Andrade, I. M., & Tumelero, C. (2022). Increasing customer service efficiency through artificial intelligence chatbot. *Revista de Gestão*, *29*(3), 238–251. https://doi.org/10.1108/REGE-07-2021-0120

Dimri, A., Yerramilli, S., Lee, P., Afra, S., & Jakubowski, A. (2019). Enhancing Claims Handling Processes with Insurance Based Language Models. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1750–1755. https://doi.org/10.1109/ICMLA.2019.00284

Doerpinghaus, H., Schmit, J. T., & Yeh, J. J. (2008). Age and Gender Effects on Auto Liability Insurance Payouts. *Journal of Risk and Insurance*, *75*(3), 527–550. https://doi.org/10.1111/j.1539-6975.2008.00273.x

Doerpinghaus, H., Schmit, J., & Yeh, J. J. (2003). Personal Bias in Automobile Claims Settlement. *Journal of Risk and Insurance*, *70*(2), 185–205. https://doi.org/10.1111/1539-6975.00055

Donnelly, K., O`Neill, S., Brettel, T., & Bargallo, M. (2024). *The $100 Billion Opportunity for Generative AI in P&C Claims Handling*. https://www.bain.com/insights/100-billion-dollar-opportunity-for-generative-ai-in-p-and-c-claims-handling/

Eckert, C., Neunsinger, C., & Osterrieder, K. (2022). Managing customer satisfaction: Digital applications for insurance companies. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *47*(3), 569–602. https://doi.org/10.1057/s41288-021-00257-z

EIOPA. (2023, December 8). *Consumer Trends Report 2023 – Statistical Annex*. https://www.eiopa.europa.eu/document/download/00126f13-7707-40b9-b033-674de0c65986_en?filename=EIOPA-BoS-23-471-Consumer-Trends-Report-2023-Annex.pdf

EIOPA. (2024). *Premiums, claims and expenses [Extract from S.05.01/Annual/Group/EUR million/Ratios]*. https://register.eiopa.europa.eu/_layouts/15/download.aspx?SourceUrl=https://register.eiopa.europa.eu/Publications/Insurance%20Statistics/GA_Premiums_Claims_Expenses.xlsx

FCA. (2023). *Insurance Conduct of Business Sourcebook*.

Ferrara, E. (2023). *Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models*. https://doi.org/10.48550/ARXIV.2304.03738

Fišar, M., Greiner, B., Huber, C., Katok, E., & Ozkes, A. I. (2024). Reproducibility in Management Science. *Management Science*, *70*(3), 1343–1356. https://doi.org/10.1287/mnsc.2023.03556

Flitter, E. (2022). New Suit Uses Data to Back Racial Bias Claims Against State Farm. *New York Times*.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S.,

Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks Are All You Need* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2306.11644

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.06825

Kimura, A., Cultu, D., Larrea Tamayo, E., Tumas Dienstag, G., Novo Sánchez, J. M., Serra Montolí, B., Martini, F., Kohls, S., Wilms, H., Diganci, N., Polyblank, J., Bokhari, A., Varney, S., Ebert, S., Danielson, J., & Khaitan, S. (2024). *Global Insurance Report 2025: The pursuit of growth*. https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-report

Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S. (2019). Motivations, Classification and Model Trial of Conversational Agents for Insurance Companies. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 19–30. https://doi.org/10.5220/0007252100190030

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. In M. Bernstein, S. Savage, & A. Bozzon (Eds.), *Proceedings of The ACM Collective Intelligence Conference* (pp. 12–24). ACM. https://doi.org/10.1145/3582269.3615599

Li, D., Jin, Z., Qian, L., & Yang, H. (2025). Textual analysis of insurance claims with large language models. *Journal of Risk and Insurance*, jori.70004. https://doi.org/10.1111/jori.70004

Lin, X., Browne, M. J., & Hofmann, A. (2022). Race discrimination in the adjudication of claims: Evidence from earthquake insurance. *Journal of Risk and Insurance*, *89*(3), 553–580. https://doi.org/10.1111/jori.12386

Lockhart, J. W., King, M. M., & Munsch, C. (2023). Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, *7*(7), 1084–1095. https://doi.org/10.1038/s41562-023-01587-9

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., … Love, B. C. (2024). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-02046-9

NAIC. (2024a). *2023 Property and Casualty Market Share*.

NAIC. (2024b). *Closed confirmed consumer complaints by reason*. https://content.naic.org/cis_agg_reason.htm

NAIC. (2024c). *Closed confirmed consumer complaints by reason*. https://content.naic.org/cis_agg_reason.htm

NAIC. (2024d). *Report on Profitability by Line by State in 2022*. https://content.naic.org/sites/default/files/publication-pbl-pb-profitability-line-state.pdf

NAIC. (2024e). *U.S. Property & Casualty and Title Insurance Industries – 2023 Full Year Results*. https://content.naic.org/sites/default/files/2023-annual-property-and-casualty-insurance-industries-analysis-report.pdf

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, *6*(1), 195. https://doi.org/10.1038/s41746-023-00939-z

OpenAI. (2023a). *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2023b). *Our approach to AI safety*. https://openai.com/blog/our-approach-to-ai-safety

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

RGA. (2023). *ChatGPT: A conversation about underwriting and life insurance*. https://www.rgare.com/knowledge-center/article/chatgpt-a-conversation-about-underwriting-and-life-insurance

Riikkinen, M., Saarijärvi, H., Sarlin, P., & Lähteenmäki, I. (2018). Using artificial intelligence to create value in insurance. *International Journal of Bank Marketing*, *36*(6), 1145–1168. https://doi.org/10.1108/IJBM-01-2017-0015

Schreiber, D. (2016). *Lemonade Sets a New World Record: How A.I. Jim broke a world record without breaking a sweat*. https://www.lemonade.com/blog/lemonade-sets-new-world-record/

Scott, M. L., Bone, S. A., Christensen, G. L., Lederer, A., Mende, M., Christensen, B. G., & Cozac, M. (2023a). Revealing and Mitigating Racial Bias and Discrimination in Financial Services. *Journal of Marketing Research*. https://doi.org/10.1177/00222437231176470

Scott, M. L., Bone, S. A., Christensen, G. L., Lederer, A., Mende, M., Christensen, B. G., & Cozac, M. (2023b). Revealing and Mitigating Racial Bias and Discrimination in Financial Services. *Journal of Marketing Research*. https://doi.org/10.1177/00222437231176470

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. https://doi.org/10.48550/ARXIV.2206.04615

The European Parliament. (2024). *Artificial Intelligence Act*. https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf

The White House. (2022). *Blueprint for an AI Bill of Rights: Making automated systems work for the American people*. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). *"Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2310.09219

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, brian, Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, *6*(1), e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57. https://doi.org/10.1162/tacl_a_00632