Code Overview

Our implementation creates a secure, ethical AI assistant (InsuranceGPT) using Meta's Llama 3.1 model (8B) in LM Studio. The system prompt establishes ethical guardrails, permissible risk factors, and transparent calculation methodologies. The premium calculation uses a sequential multiplier approach starting with a base premium from vehicle value (BLUEBOOK), then applies risk multipliers for legally permissible factors including age, vehicle usage, car age, claims history, license status, MVR points, location, and commute time.

Each calculation is accompanied by transparent explanations breaking down how factors influenced the final premium, rounded to the nearest $25. Our testing confirmed the system successfully rejected problematic requests related to race/ethnicity adjustments, accessing others' personal information, and assistance with fraudulent activities, validating our ethical guardrail implementation. All calculation steps are documented to ensure transparency and auditability of recommendations.

Overview of LLMs for Safety and Vulnerability

LLMs introduce unique safety concerns in high-stakes domains like insurance, where recommendations directly impact financial decisions. Our analysis using ALERT and SALAD-Bench benchmarks identified several vulnerability categories.

Direct harm potential manifests through misinformation risk, where inaccurate calculations could cause financial harm. Despite explicit programming to avoid discrimination, latent biases may establish correlations between permitted variables and protected attributes, potentially resulting in indirect discrimination [1]. This concern aligns with research showing LLMs can perpetuate societal biases from their training data.

Adversarial vulnerabilities represent significant concerns, including prompt injection attempts to circumvent ethical guidelines (e.g., requests to adjust premiums based on race), privacy exploitation attempts, and requests for assistance with fraudulent activities. These vulnerabilities persist despite explicit safeguards [2].

Misuse potential extends beyond direct attacks to more subtle exploitation paths, including soliciting harmful advice and multi-step interactions where benign requests establish context for subsequent problematic queries. These findings reflect broader patterns where contextual manipulation often proves more effective than direct attacks [3].

Our Standard

InsuranceGPT upholds fairness, security, and privacy through three main principles: ethical guardrails, attack resistance, and privacy protection. It calculates premiums using only legally permissible features, explicitly rejecting discriminatory factors like gender, income, and family status. Instead, it focuses on objective risk factors such as vehicle value and driving behavior, ensuring transparency and consistency in pricing. Customers receive clear explanations of how each factor influences their premium, promoting fairness throughout the process.

The system features input validation to block unethical requests, such as race-based premium adjustments or fraudulent activity. Ethical response protocols reject such requests while offering appropriate alternatives. Testing confirmed that the model can resist manipulation and maintain ethical integrity in its operations. To ensure privacy protection, InsuranceGPT follows data minimization principles, only collecting the necessary information for premium calculation. The system complies with GDPR[4], safeguarding personal data and protecting against social engineering attempts.

Premiums are determined by the vehicle's BLUEBOOK value, ensuring pricing is based on objective metrics rather than irrelevant characteristics. Risk multipliers are applied based on age, vehicle usage, claims history, and other factors, with clear reasoning for each adjustment. These multipliers follow principles of direct harm (e.g., higher risk for young drivers) and indirect harm (e.g., increased risk for commercial vehicles).

By excluding prohibited factors like gender and income, InsuranceGPT ensures fair treatment for all customers, maintaining both ethical fairness and legal compliance. Testing with ALERT and SALAD-Bench showed that the system balances utility with safety, ensuring ethical and legal standards are met.

Potential Objections and Future Solutions

Despite robust implementation, several limitations warrant consideration. The proxy variable challenge represents a significant concern, as permitted variables (particularly location) might serve as unintentional proxies for protected attributes. To address this, we propose implementing regular fairness audits that systematically evaluate premium recommendations across demographic groups.

Model drift presents another challenge, as underlying LLM behavior changes with updates. Safety measures effective today might require recalibration as models evolve. We recommend continuous monitoring protocols that regularly assess safety performance against benchmark datasets [5].

Advanced prompting techniques represent an evolving threat, as users develop new methods to circumvent safety measures. Our implementation resists known attacks, but new techniques will emerge. To counter this, we propose establishing an adversarial testing program that continuously evaluates the system against evolving attack methods.

Finally, our current system lacks comprehensive intersectional fairness analysis. We propose implementing methods to detect and mitigate bias affecting specific subgroups defined by multiple attributes, aligning with approaches described by Kearns et al. for subgroup fairness analysis [6].

By addressing these limitations through ongoing refinement, InsuranceGPT can provide valuable, personalized insurance recommendations while upholding ethical standards for fairness, safety, and accountability.

References

[1] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Le, Q. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

[3] Perez, E., Huang, S., Song, H. F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

[4] Legal text. General Data Protection Regulation (GDPR). (2024). https://gdpr-info.eu/

[5] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

[6] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. Proceedings of the 35th International Conference on Machine Learning.