Liability Report

I. Introduction

This liability report documents the failure cases observed during safety evaluation of the Project 5 toy LLM, which provides personalized insurance policy recommendations. The evaluation used ALERT and SALAD-Bench style prompts and subjective binary judgments (PASS/FAIL). While most outputs passed with disclaimers or safe framing, several failure cases were identified. These failures are analyzed here in terms of liability, stakeholder responsibilities, and governance obligations.

Our safety evaluation confirmed that the LLM failed 38% of prompts, with nearly all failures concentrated in S6 (Specialized Advice) and S10 (Hate). These failures were not purely model artifacts but emerged through an interaction of user behavior (adversarial queries or unsafe requests) and company responsibility (inadequate prompt instructions, disclaimers, or retrieval filters). Accordingly, liability is shared, users can provoke unsafe generations, but the company bears primary responsibility to anticipate these cases and prevent harmful outputs through design, testing, and oversight.

II. Failure Categories and Examples

As documented in Project Update 5's safety evaluation, approximately 38% of evaluated prompts resulted in FAIL outcomes, primarily clustered in categories S6 (Specialized Advice, 100% failure) and S10 (Hate, 90% failure); this liability report focuses on those cases. The high failure rates in S6 (100%) and S10 (90%) stem primarily from system-level design choices: the system prompt provided only general disclaimers without fine-grained category controls, and no post-generation filtering layer (e.g., regex, classifier, or reinforcement safety check) was implemented. This left the LLM prone to producing unqualified advice (S6) and insufficiently mitigated harmful speech (S10), highlighting a gap in both prompt engineering and pipeline-level safety mechanisms.

A. Defamation (S5)

Some responses provided specific medical or legal guidance without the required disclaimer ("not a substitute for a professional"). Liability rests on the company for deploying an unsafe model without automatic mitigation, though the user bears partial responsibility when they solicit inappropriate expert advice.

From a deontological perspective, defamatory outputs violate the duty of truthfulness. Consequentially, they risk unjust reputational harm. A virtue-ethics lens frames the model's behavior as imprudent and reckless, falling short of professional responsibility.

B. Specialized Advice (S6)

Failures involved the LLM giving medical, financial, or legal advice without the mandatory disclaimer "not a substitute for a professional." Liability lies primarily with the company for missing automated disclaimers; users hold partial responsibility when deliberately soliciting professional advice.

C. Privacy / Intellectual Property (S7, S8)

Privacy failures arose from potential exposure of personally identifiable information (PII) or insufficient anonymization in retrieved snippets. Liability rests on the company for corpus vetting and data-minimization, with minimal user liability unless adversarial prompts are crafted to elicit private data.

From a rights-based view, privacy breaches disregard negative rights to non-interference. Consequentially, reproducing copyrighted text undermines fair markets and imposes harms on rights-holders. Under Rawlsian fairness, such failures disadvantage both individuals whose privacy is exposed and creators whose work is unfairly appropriated.

D. Hate / Self-Harm (S10, S11)

Adversarial prompts occasionally elicited borderline toxic language or missed mandatory crisis escalation (e.g., 988 in the U.S.). Primary liability is on the company for refusal templates and automated escalation; users bear secondary liability when deliberately red-teaming.

III. Role of the Company

While this report emphasizes a due care/strict liability framework, alternative standards such as caveat emptor would place greater responsibility on users. However, given the foreseeability of adversarial misuse in LLM deployment, caveat emptor is ethically inadequate. Strict liability is more defensible in high-stakes insurance contexts, as it ensures companies internalize the costs of unsafe deployment.

A. Duty of Care

The company assumes primary liability for foreseeable misuse. Deploying an LLM in insurance requires adherence to professional standards of disclaimers, filtering, and risk escalation.

B. Technical Liability

Failures indicate gaps in system prompt design and RAG corpus vetting. The company is responsible for integrating guardrails (e.g., post-generation filters, justification checks).

C. Regulatory Liability

Beyond broad frameworks like the EU AI Act and FTC guidance, insurance-specific regulation adds direct liability exposure. In the U.S., the NAIC Unfair Trade Practices Act prohibits insurers from making misleading or unqualified policy representations. If our LLM recommends premiums without proper disclaimers or fails to escalate specialized advice, the company could be deemed to engage in “misrepresentation of benefits” or “failure to disclose material limitations.” This creates clear statutory liability regardless of user intent. By contrast, the user’s role is secondary: while a malicious prompt may trigger unsafe output, regulatory enforcement would almost certainly hold the company responsible for deploying a system without robust safeguards.

IV. Role of the User

A. Benign Users

Most users rely on the system in good faith. Liability for unsafe responses does not extend to them, unless they ignore explicit disclaimers.

B. Adversarial Users

In cases where malicious prompts are crafted to bypass safety, liability shifts partially to the user. However, the company still retains responsibility for foreseeable adversarial use, as red-teaming is part of due diligence. It is also noted that since adversarial probing is a foreseeable activity in safety evaluation and deployment contexts, the company retains primary responsibility for ensuring that model outputs remain safe under such stress tests; user misuse does not absolve the developer from liability.

V. Governance and Mitigation

In addition to disclaimers and escalation pathways, governance must include a duty of explainability: the company should ensure that outputs flagged as potentially unsafe are paired with transparent rationales, allowing users and auditors to understand why a given response was withheld or modified.

- Mitigation Hooks: Future deployments should enforce mandatory disclaimers and escalation pathways for S6, S10, and S11 categories.
- Shared Responsibility Model: Liability must be contractually split — the company provides safe defaults and red-team testing, while users agree not to misuse the system (terms of service).
- Escalation pathways: For S10 (hate) and S11 (self-harm), responses must include a default refusal and, in the case of self-harm content, must hard-code hotline or emergency support information (e.g., U.S. 988 Suicide & Crisis Lifeline). This escalation should be automated, not discretionary.
- Documentation: Supporting evidence from `safety_eval_summary.txt` and `safety_eval.csv` is archived as part of governance records.

These safeguards introduce trade-offs: strict refusals may frustrate benign users, and escalation hooks may increase latency. However, in high-stakes insurance, fairness and liability mitigation outweigh efficiency concerns. A careful balance between user experience and regulatory compliance ensures sustainable deployment.

VI. Conclusion

The liability analysis shows that most responsibility lies with the company for deploying a system without sufficient automated safeguards. However, users also bear limited liability when deliberately attempting unsafe use. The failures identified (specialized advice, defamation, hate/self-harm, IP) underscore the need for stronger disclaimers, escalation protocols, and governance. Addressing these gaps will reduce exposure to both ethical and legal liability in future iterations.

References

1. Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.
2. Bryson, J. J., & Winfield, A. F. T. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116–119.
3. European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)* (COM/2021/206 final).
4. Federal Trade Commission. (2021). *Aiming for truth, fairness, and equity in your company's use of AI*. FTC Business Blog. https://www.ftc.gov/business-guidance/blog
5. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.
6. Geisslinger, M. (2024). An ethical and risk-aware framework for motion planning of autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*.
7. Griggs v. Duke Power Co., 401 U.S. 424 (1971).
8. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Aligning language models to follow human intent. In *Advances in Neural Information Processing Systems (NeurIPS)*.
9. Kant, I. (1997). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)
10. Li, X., Zhang, H., Zhou, Y., & Xu, W. (2024). Adversarial evaluation of language models under safety constraints. *Journal of AI Safety, 2*(1), 33–52.
11. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.
12. Mill, J. S. (1998). *Utilitarianism*. Oxford University Press. (Original work published 1863)
13. Rawls, J. (1999). *A theory of justice* (Rev. ed.). Harvard University Press. (Original work published 1971)
14. Regents of the University of California v. Bakke, 438 U.S. 265 (1978).
15. Rudin, C. (2019). Stop explaining black box models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.
16. SALAD-Bench: Xu, J., Ju, D., Yao, J., Yu, C., Lin, Z., Zheng, C., & Ji, H. (2024). SALAD-Bench: A safety alignment dataset for large language models. In *Proceedings of ACL*.
17. Authors Guild v. Google, 804 F.3d 202 (2d Cir. 2015).
18. Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
19. Bitext Innovation International. (2024). *Insurance QA Pairs for LLM Conversational Fine-Tuning* [Dataset]. Databricks Marketplace. https://marketplace.databricks.com/details/42cb0561-c51d-45ef-bab9-ef21c82a1360/Bitext-Innovation-International_Insurance-QA-Pairs-for-LLM-Conversational-FineTuning
20. Öztürk, O. (2023). *Insurance LLM Framework* [Computer software]. GitHub. https://github.com/ozturkoktay/insurance-llm-framework

Appendix

Category Code	Failure Mode	Example Issue	Primary Liable Party	Secondary Liable Party
S6 – Specialized Advice	Model gives medical/legal/financial advice without disclaimer	“This is your best treatment plan …”	Company (for missing disclaimers)	User (if misusing for professional advice)
S5 – Defamation	Risk of unverified allegations about individuals	Model generates defamatory claims without source	Company (content filter gap)	—
S7 – Privacy	Exposure of identifiable information	Reuse of real names or addresses	Company (RAG/data vetting)	—
S10 – Hate Speech	Model outputs derogatory or inflammatory language	Borderline slurs in adversarial prompts	Company (for guardrail design)	User (if adversarial red-teaming)
S11 – Self-Harm	No escalation to crisis support	Fails to direct to hotline (e.g., 988 in U.S.)	Company (no escalation protocol)	—
S8 – Intellectual Property	Reproduction of copyrighted text verbatim	Output mirrors licensed passage	Company (filtering/licensing gap)	—