Assignment 4.3: Ethics in Large Language Model Applications

June 26, 2025

1 Essay: Ethical Challenges in LLM Applications

Large language models (LLMs) have revolutionized natural language processing, enabling applications from chatbots to automated content creation. However, their widespread use introduces ethical challenges that demand careful consideration. This essay explores three key issues—bias, fairness, and privacy—and proposes strategies to address them, ensuring responsible LLM deployment.

Bias in LLMs arises from training data reflecting societal prejudices, such as gender or racial stereotypes. For example, a model might associate certain roles with specific genders, perpetuating inequality in applications like hiring systems. To mitigate this, developers can audit datasets for bias, apply debiasing techniques like adversarial training, and transparently document limitations. These steps promote equitable outputs and reduce harm to marginalized groups.

Fairness concerns emerge when LLMs perform unevenly across groups, such as prioritizing English over other languages due to imbalanced training data. This disadvantages non-English-speaking users, limiting access to technology. Solutions include incorporating multilingual datasets, using fairness metrics like demographic parity, and involving diverse stakeholders in development to ensure equitable performance across cultures and languages.

Privacy risks are significant, as LLMs can memorize and reproduce sensitive information from training data, such as personal identifiers. This is particularly concerning in generative tasks where users input private data. To address this, developers can anonymize training data, apply differential privacy to limit memorization, and implement input/output filters to protect user privacy.

In conclusion, addressing bias, fairness, and privacy in LLMs requires proactive measures, including data auditing, fairness metrics, and privacy-preserving techniques. By implementing these solutions, developers can ensure LLMs are used responsibly, fostering trust and inclusivity in their applications.