Transcript: Research Proposal Presentation

Slide 1 – Title
Good morning, my name is Marie Levesque, and I am presenting my MSc Data Science research proposal titled Bias and Fairness in Large Language Models for Financial Compliance: Risks and Mitigation.

This project investigates the behaviour of Large Language Models (LLMs) when applied to compliance tasks in the financial services sector, where accuracy, impartiality, and fairness are paramount. The focus is on identifying biases, testing mitigation strategies, and developing a framework for fairness auditing.

———

Slide 2 – Introduction and Significance
Financial institutions operate within complex regulatory frameworks, including Solvency II for insurers, IFRS 17 for financial reporting, Basel III for banking stability, and successive Anti-Money Laundering Directives in the EU. Compliance is costly, labour-intensive, and central to maintaining both trust and financial stability.

Regulatory technology, or RegTech, has emerged as a response to this. By applying AI and machine learning, firms aim to reduce compliance costs while improving consistency. Large Language Models are especially attractive because they can summarise complex texts, extract obligations, and classify risks.

However, this promise comes with serious risks. Models trained on broad, internet-scale corpora may reproduce or amplify bias. In compliance, biased outputs could mean unfair treatment of clients, uneven enforcement of obligations, or reputational and legal risks. As the European Securities and Markets Authority and the Alan Turing Institute (2025) stress, trustworthy AI in finance requires fairness, transparency, and accountability.

This project is therefore significant because it addresses both a professional need — reliable compliance automation — and an academic gap — fairness in financial AI.

———

Slide 3 – Research Problem
The research problem can be summarised as follows:

Bias in AI is well-documented in various domains, including hiring, healthcare, and criminal justice (Bender et al., 2021; Mehrabi et al., 2021). Yet in financial compliance, despite its high stakes, systematic research on fairness in LLMs is lacking.

Meanwhile, most financial AI research has focused on investment, sentiment analysis, and trading. Compliance tasks, although crucial, have received far less attention. Without investigation, institutions risk deploying biased tools into legally sensitive areas.

Thus, the problem lies in the lack of empirical evidence on how LLMs perform in compliance contexts and whether fairness can be meaningfully evaluated and improved.

———

Slide 4 – Research Question
The guiding research question is:

How can bias and fairness be evaluated and mitigated when Large Language Models are applied to financial compliance?

Supporting questions are:

- What forms of bias emerge in compliance-related tasks such as obligation extraction, risk classification, and regulatory summarisation?
- Do domain-specific models such as FinGPT behave differently from general-purpose ones like GPT-4?
- Which mitigation approaches — prompt engineering, retrieval-augmented generation, or fine-tuning — are most effective at reducing bias while maintaining accuracy?

———

Slide 5 – Aims and Objectives
The aim is to critically evaluate fairness in LLMs for financial compliance and propose an auditing framework.

The objectives are:
1. To identify and categorise sources of bias in compliance-related LLM outputs.
2. To compare fairness and consistency between general-purpose and finance-specific models.
3. To experiment with mitigation strategies.
4. To design a practical auditing framework to help compliance teams monitor AI outputs.

———

Slide 6 – Literature Review: Bias and Fairness in AI
Bias in AI has been extensively analysed. Bender et al. (2021) argue that LLMs risk amplifying harmful stereotypes because of opaque training data, while Blodgett et al. (2020) survey how "bias" is conceptualised in natural language processing and stress that fairness is context-dependent.

Mehrabi et al. (2021) provide a comprehensive survey of fairness metrics, highlighting quantitative measures such as demographic parity, equalised odds, and calibration. Although designed for classification, these can be adapted to compliance tasks.

Jobin, Ienca and Vayena (2019) identify fairness, accountability, and transparency as consistent principles across global AI ethics guidelines. This aligns with Leslie's (2019) guidance from the Alan Turing Institute, which stresses responsible deployment of AI in public and regulated sectors.

Taken together, these works establish the centrality of fairness but also reveal that methodologies remain fragmented and rarely tested in compliance domains.

———

Slide 7 – Literature Review: Governance and Regulation
AI in financial services sits within an evolving governance landscape. Goodman and Flaxman (2017) examine the EU's "right to explanation," which requires transparency in algorithmic decision-making under GDPR. Veale and Edwards (2018) extend this analysis to profiling guidance, showing how accountability expectations are tightening.

At a higher level, the EU AI Act identifies finance as a "high-risk" sector, demanding enhanced auditing and explainability. Zetzsche et al. (2020) discuss how RegTech adoption can support regulatory efficiency but also highlight governance challenges.

These works show that regulators expect AI to be both effective and fair. Yet, practical frameworks for auditing LLMs in compliance are still underdeveloped.

———

Slide 8 – Literature Review: LLMs in Finance

In finance-specific contexts, LLMs have demonstrated strong performance on tasks such as sentiment analysis and information retrieval. Bloomberg (2023) presented BloombergGPT, a 50-billion-parameter LLM trained on financial data, which achieves superior accuracy on financial NLP benchmarks. FinGPT, launched by the AI4Finance community, provides an open-source alternative fine-tuned on financial corpora.

Empirical studies have shown that LLMs outperform traditional models in trading strategies and sentiment extraction (Kirtac and Germano, 2025). However, as Blodgett et al. (2020) caution, accuracy does not guarantee fairness.

The ESMA and Alan Turing Institute (2025) report explicitly warns that fairness and explainability must not be overlooked in finance; however, they remain largely conceptual.

Thus, the literature demonstrates capability but leaves fairness in compliance as an open problem — one that this project directly addresses.

———

Slide 9 – Methodology: Research Design
The project will use an experimental, mixed-methods design.

Data will consist of publicly available regulatory texts, including Solvency II, IFRS 17, Basel III, and AML Directives. These provide rich material for compliance tasks.

Models tested will include GPT-4 (general-purpose), FinGPT (open-source, finance-specific), and benchmarks from BloombergGPT.

Tasks will be obligation extraction, compliance risk classification, and regulatory summarisation.

———

Slide 10 – Methodology: Evaluation and Fairness Metrics
Evaluation will address both accuracy and fairness.

Accuracy will be measured using standard NLP metrics, including precision, recall, and F1 scores for extraction and classification tasks, as well as ROUGE or BERTScore for summarisation.

Fairness will be operationalised through a dual approach:

- Quantitative metrics, adapted from surveys by Mehrabi et al. (2021) and Blodgett et al. (2020). These include consistency (identical inputs should yield consistent outputs), calibration (predicted risk levels should reflect actual distributions), and parity (outputs should not disproportionately target particular groups or jurisdictions).
- Qualitative analysis, inspired by Raji et al. (2020), will examine outputs for systematic skew — for example, whether specific industries are over-flagged.

Trade-offs between accuracy and fairness will be explicitly analysed, since mitigation techniques may reduce one while improving the other. This critical balance is central to the research design.

———

Slide 11 – Mitigation Strategies
The study will test three approaches to reducing bias:

1. Prompt engineering, to phrase inputs neutrally and reduce unintended framing.
2. Retrieval-augmented generation (RAG), grounding outputs in curated regulatory corpora to reduce hallucinations and bias.
3. Fine-tuning, using compliance-specific subsets with FinGPT, to align outputs more closely with regulatory contexts.

Performance before and after mitigation will be compared to determine which strategies are most effective.

———

Slide 12 – Artefact and Framework
The primary artefact will be a fairness auditing framework tailored to LLM applications in compliance.

It will consist of:

- A structured framework document, outlining indicators and guidelines that compliance teams can apply in practice.
- A prototype Python notebook, automating fairness checks such as consistency and calibration tests.

Validation will follow Raji et al. (2020), who propose end-to-end auditing workflows. Ideally, the framework will be reviewed with input from compliance practitioners, ensuring it is both academically robust and practically usable.

———

Slide 13 – Ethical Considerations
Ethical safeguards include using only publicly available regulatory data and synthetic scenarios, while avoiding the use of personal data entirely.

The project acknowledges risks of unfair outcomes and will maintain a human-in-the-loop perspective, stressing that AI should support rather than replace compliance officers.

Environmental concerns are also relevant, given the energy intensity of large models. These will be reflected in the ethical evaluation.

———

Slide 14 – Timeline
The project spans six months.

Months one and two: literature review and dataset preparation.
Months three and four: baseline experiments and bias evaluation.
Month five: mitigation testing and framework development.
Month six: validation, analysis, and dissertation write-up.

This staged plan ensures feasibility within the scope of the MSc.

———

Slide 15 – Contribution
The contribution will be threefold.

Academically, it extends fairness research into financial compliance, a domain where empirical work is scarce.

Practically, it delivers a framework and prototype tool for auditing fairness in LLM outputs.

At policy level, it supports regulators' emphasis on trustworthy AI, aligning with GDPR, the EU AI Act, and ESMA's fairness principles.

In summary, the project bridges the fields of data science, finance, and regulation, laying the groundwork for both academic research and practical implementation.

———

Slide 16 – Conclusion
To conclude, compliance is a high-stakes and resource-intensive endeavour. LLMs offer transformative potential, but without fairness auditing, they risk introducing bias into one of the most sensitive domains in finance.

This project will critically evaluate bias, test mitigation strategies, and propose a practical auditing framework for addressing bias. It therefore addresses a clear academic gap, meets regulatory priorities, and offers real-world impact.

Thank you for your attention. I welcome your questions and feedback.

———

References

Bender, E.M., Gebru, T., McMillan-Major, A. and 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 (FAccT '21), pp. 610–623.

Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020) 'Language (technology) is power: A critical survey of "bias" in NLP', Transactions of the Association for Computational Linguistics, 8, pp. 545–565.

Bloomberg (2023) BloombergGPT: A large language model for finance. Available at: https://arxiv.org/abs/2303.17564 (Accessed: 9 October 2025).

European Securities and Markets Authority (ESMA) and The Alan Turing Institute (2025) Large language models in finance. Paris: ESMA. Available at: https://www.esma.europa.eu/sites/default/files/2025-06/LLMs_in_finance_-_ILB_ESMA_Turing_Report.pdf (Accessed: 9 October 2025).

Goodman, B. and Flaxman, S. (2017) 'European Union regulations on algorithmic decision-making and a "right to explanation"', AI Magazine, 38(3), pp. 50–57.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1(9), pp. 389–399.

Leslie, D. (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. London: The Alan Turing Institute.

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

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing', Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20), pp. 33–44.

Veale, M. and Edwards, L. (2018) 'Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling', Computer Law & Security Review, 34(2), pp. 398–404.

Zetzsche, D.A., Buckley, R.P., Arner, D.W. and Barberis, J.N. (2020) 'Regulating FinTech: Lessons from Africa', Law and Contemporary Problems, 83(2), pp. 43–79.