Challenges and Ethical Considerations in Developing Deep Learning-Based Text Summarization Systems

Hello, my name is Dewyn, and thank you for joining me.

This presentation outlines my research proposal on the *challenges and ethical considerations involved in developing deep learning-based text summarization systems*.

1. Introduction

We live in an age of information overload — news articles, academic papers, reports — all demanding our attention. Deep learning-based summarization systems offer a powerful solution, generating concise summaries using models like BERT, GPT, and T5 (Devlin et al., 2019; Radford et al., 2019).

But these systems raise important concerns. Can we trust their outputs? Are they unbiased? Do they expose sensitive information? And perhaps most critically — who is accountable when things go wrong?

My research addresses these questions by proposing a practical ethical evaluation framework — one that not only identifies risks but helps developers and researchers measure and manage them, through both structured criteria and a working artefact.

---

2. Research Question

The core research question guiding this project is:

How can ethical considerations such as bias, privacy, and transparency be systematically integrated into the development of deep learning-based text summarization systems without compromising summarization performance?

This question reflects a major challenge in AI: balancing innovation with responsibility. Today, state-of-the-art summarization systems are evaluated primarily using metrics like ROUGE and BLEU (Lin, 2004), which measure surface similarity but say nothing about bias, truthfulness, or fairness.

As Kryściński et al. (2019) point out, these systems may generate fluent text — yet still hallucinate facts or misrepresent source content. And as shown by Abadi et al. (2016), privacy violations can occur when models inadvertently reveal information from their training data.

My project seeks to bridge this gap — by designing an evaluation framework that captures both performance and ethics, and by giving practitioners a toolkit they can actually use in real-world development.

---

3. Aims and Objectives

The aim of this research is to develop a usable, ethically aware evaluation framework — along with a scoring artefact — to help practitioners assess summarization systems from both technical and ethical perspectives.

The project is guided by five objectives:

1. **Identify and classify key ethical challenges**
   Drawing from recent literature and known failures, I'll map common issues like bias (Mehrabi et al., 2021), privacy leakage (Abadi et al., 2016), and hallucination — building a foundation for meaningful evaluation.
2. **Evaluate model performance on both standard and ethical dimensions**
   Using a test set from CNN/DailyMail or similar, I'll compare summaries from models like BART or T5 using traditional metrics (Lin, 2004) as well as ethical markers: fairness, factual accuracy, and privacy risks.

3. **Develop a hybrid framework**
   This framework will combine technical evaluation with ethical dimensions — using inspiration from fairness surveys (Mehrabi et al., 2021) and explainability work (Doshi-Velez & Kim, 2017) to create a balanced, practical guide.
4. **Build a lightweight scoring artefact**
   The artefact will be a Python script and scoring rubric that allows developers to quickly audit summaries — giving scores and flags across four key dimensions: accuracy, bias, privacy, and transparency.
5. **Validate the framework through domain-specific testing**
   I'll apply the toolkit to summaries in domains like healthcare, legal texts, and news reporting — refining the criteria based on those real-world use cases.

---

4. Literature Review

Let's now explore the key ethical issues raised by the literature.

Bias and fairness are persistent concerns. As Mehrabi et al. (2021) explain, models trained on internet-scale data often replicate societal bias — leading to summaries that amplify stereotypes or marginalize certain voices.

Privacy is another red flag. Abadi et al. (2016) demonstrated that deep learning models can memorize portions of their training data — risking the exposure of personal or sensitive information, particularly in medical or legal domains.

Evaluation metrics are another issue. ROUGE and BLEU (Lin, 2004) remain the default standards, but they don't reflect whether a summary is *true*, *coherent*, or *fair*. Kryściński et al. (2019) and others have pushed for metrics like FactCC and QuestEval to fill that gap — but no standard has yet emerged.

Transparency is the final piece. Transformer models like GPT or T5 operate as black boxes. As Lipton (2018) and Doshi-Velez and Kim (2017) point out, this lack of interpretability makes it hard to trace decisions — and harder still to build accountability when errors or bias occur.

These challenges aren't just technical — they're deeply human. And they demand a structured, practical response. That's what this research proposes to deliver.

---

5. Methodology

This project will follow a four-phase methodology:

Phase 1: Thematic Literature Analysis
I'll conduct a structured review of summarization research, ethical AI, and real-world case studies to define the evaluation categories.

Phase 2: Model Evaluation
Using a dataset like CNN/DailyMail, I'll run common summarization models and score them across standard and ethical metrics — incorporating benchmarks like ROUGE (Lin, 2004), FactCC (Kryściński et al., 2019), and fairness indicators drawn from Mehrabi et al. (2021).

Phase 3: Framework and Artefact Development
The framework will include detailed scoring categories and thresholds. I'll implement it as a Python script for automated scoring, paired with a rubric and guidebook for human review.

Phase 4: Case-Based Validation
To evaluate its scalability, I'll apply the framework to several distinct domains — such as summarizing patient notes, legal briefs, or news headlines — and refine the scoring rules based on what works in practice.

---

6. Ethical Considerations

Although this study uses public datasets, it engages heavily with systems that can impact people.

These ethical considerations aren't just technical details — they reflect real-world consequences. A biased summary in a legal context could misrepresent evidence. A privacy breach in a medical document could compromise patient confidentiality. And a lack of transparency could erode trust in automated systems altogether. That's why this project treats ethics not as an afterthought, but as a core part of model evaluation.

Ethical priorities include:

- **Avoiding datasets** that contain sensitive or personally identifiable information
- **Framing the artefact** as a decision-support tool — not a definitive ethics validator
- **Encouraging transparency** and traceability in model evaluation

The framework will be designed to raise flags rather than produce blanket judgments — empowering human oversight rather than replacing it. By making ethical risks more visible and structured, the goal is to help developers make more informed, responsible decisions when deploying summarization systems in high-stakes domains.

---

7. Risk Assessment

While this project is considered low risk in terms of data handling and implementation, the implications of its application in sensitive areas require thoughtful attention. Summarization tools, when used in domains like healthcare or law, can influence how information is interpreted — and by extension, how decisions are made. With that in mind, the project includes a proactive approach to identifying and managing potential risks.

Key areas of focus include:

- **Data misuse**: All datasets will be reviewed to ensure they are ethically sourced and properly anonymized, with a preference for well-established public corpora.
- **Scoring bias**: Since ethical dimensions like fairness and transparency involve some level of subjectivity, the rubric will include clear definitions, real-world examples, and undergo small-scale pilot testing to promote consistency.
- **Artefact misuse**: The tool could be misunderstood as a definitive ethics checker. To address this, it will include clear documentation that positions it as a decision-support tool — not a substitute for human judgment.
- **Generalization limitations**: The framework may not be equally effective in every context. This will be acknowledged during testing, with domain-specific results used to refine and clarify its intended scope.

Overall, these measures aim to ensure the framework is both practical and responsible — highlighting its usefulness while also recognizing its boundaries.

---

8. Artefact Description

The artefact will consist of two parts:

1. **Scoring Tool**
   A Python script that accepts summary outputs and returns scores across four axes: technical quality, bias detection, privacy exposure, and interpretability. It will also highlight key areas for review — similar to a model audit.
2. **Rubric and Guidebook**
   A PDF rubric will define the scoring logic with examples. This will help developers, researchers, or auditors apply the framework consistently — and adapt it to their own use cases.

The goal is not to create a heavy academic framework, but a *lightweight practical toolkit* — something teams could use during development sprints or model evaluation phases.

---

9. Timeline

Week	Goal

Week 1-2	Finalize research question and literature review
Weeks 3–4	Model selection and summary output collection
Weeks 5–6	Define evaluation metrics and begin framework design
Weeks 7–8	Build artefact and conduct domain testing
Week 9	Refine outputs, test usability
[illegible]	[illegible]

10. Conclusion

In conclusion, deep learning has transformed how we process and condense information. Text summarization systems can make knowledge more accessible and efficient — but without oversight, they can also amplify bias, reveal private data, or distort the truth, all while scoring highly on traditional metrics.

This project doesn't aim to halt innovation — it embraces it, while adding a layer of reflection and responsibility. By creating a practical, scalable framework and artefact for ethical evaluation, I aim to help developers and researchers make ethics an integrated part of the process — something measurable, understandable, and actionable. Because if summarization systems are to be trusted in high-stakes areas like health, law, or education, we need to ensure they're not just smart — but also fair, transparent, and safe.

Ultimately, this work aspires to support a future where artificial intelligence respects both the complexity of language and the dignity of the people behind it.

Thank you very much for listening.

References

- **Abadi, M.**, Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L. (2016) 'Deep learning with differential privacy', *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*
- **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)*
- **Devlin, J.**, Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv preprint*, arXiv:1810.04805.
- **Doshi-Velez, F.** and **Kim, B.** (2017) 'Towards a rigorous science of interpretable machine learning', *arXiv preprint*, arXiv:1702.08608.
- **Kryściński, W.**, McCann, B., Xiong, C. and Socher, R. (2019) 'Evaluating the factual consistency of abstractive text summarization', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- **Lin, C.Y.** (2004) 'ROUGE: A Package for Automatic Evaluation of Summaries', *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.
- **Lipton, Z.C.** (2018) 'The Mythos of Model Interpretability', *Communications of the ACM*, 61(10).
- **Mehrabi, N.**, Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*.
- **Radford, A.**, Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) 'Language models are unsupervised multitask learners', *OpenAI Technical Report*, Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- **Vaswani, A.**, Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems (NeurIPS)*