School of Computing, Engineering, and the Built Environment
Edinburgh Napier University

PHD STUDENT PROJECT

Application instructions:
Detailed instructions are available at :
https://www.napier.ac.uk/research-and-innovation/doctoral-college/how-to-apply

Prospective candidates are encouraged to contact the Director of Studies (see details below) to discuss the project and their suitability for it.

Project details

Supervisory Team:

- DIRECTOR OF STUDY: Dr Dimitra Gkatzia (Email: d.gkatzia@napier.ac.uk)
- 2ND SUPERVISOR: Dr Zia Md Ullah

Subject Group: Computer Science

Research Areas: Computer Science, Artificial Intelligence, Machine Learning

Project Title: Evaluation of Large Language Models

Project description:

The rapid advancements in natural language processing (NLP) have been significantly propelled by the development of large language models (LLMs) such as GPT-4, BERT, and T5. These models have shown remarkable capabilities in various tasks, including language generation, translation, summarisation, and more. However, evaluating these models poses unique challenges due to their complexity, scope, and the diverse applications they support. This PhD research aims to develop comprehensive evaluation methodologies for large language models, addressing both quantitative and qualitative aspects. This PhD project will explore existing evaluation metrics, propose novel metrics, and establish a holistic framework that balances performance, interpretability, and ethical considerations.

Background and Significance:

Large language models have transformed the field of NLP, achieving state-of-the-art results in numerous benchmarks. Despite their success, the evaluation of these models remains a multifaceted challenge. Traditional metrics like BLEU, ROUGE,

and perplexity provide limited insights into the models' true capabilities and limitations. Moreover, as LLMs are deployed in real-world applications, considerations such as fairness, bias, and ethical implications become critical. Therefore, a thorough and multi-dimensional evaluation framework is necessary to understand and improve LLMs comprehensively.

BLEU, ROUGE, METEOR, and perplexity have been widely used for evaluating language models, focusing on aspects like fluency, coherence, and syntactic accuracy but often failing to capture semantic nuances and contextual appropriateness. Human judgments, considered the gold standard for evaluating generated text quality, are resource-intensive and may suffer from subjectivity and inconsistency. Additionally, models are often evaluated based on their performance on specific tasks such as translation, summarisation, and question answering, which, while useful, do not provide a comprehensive view of the model's general language understanding and generation capabilities.

Furthermore, studies have shown that large language models (LLMs) can propagate and even amplify biases present in the training data, making it crucial to evaluate these models for fairness and bias to ensure equitable and unbiased AI systems. Additionally, understanding the decision-making process of LLMs is essential for trust and accountability, and existing research on model interpretability will inform the development of evaluation methods that provide insights into model behaviour.

Indicative Research Questions:

1. What are the limitations of current evaluation metrics for large language models?
2. How can we develop novel metrics that better capture the performance and limitations of these models?
3. What role do ethical considerations play in the evaluation of large language models, and how can these be integrated into the evaluation framework?
4. How can we balance quantitative and qualitative evaluation methods to provide a holistic assessment of large language models?

The expected contributions include a detailed critique of existing metrics, novel metrics for semantic and contextual quality, ethical evaluation metrics, and a holistic evaluation framework balancing quantitative and qualitative assessments.

References:

Miltenburg, Clinciu, Dušek, Gkatzia, Inglis, Leppänen, Mahamood, Manning, Schoch, Thomson, Wen. (2021). Underreporting of errors in NLG output, and what to do about it. In INLG.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation." In ACL.

Lin, C. Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In Proceedings of the ACL-04 Workshop on Text Summarization Branches Out.

Banerjee, S., & Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). "Improving Language Understanding by Generative Pre-Training." OpenAI.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In NAACL.

Ethayarajh, K. (2020). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." In EMNLP.

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).

Candidate characteristics

Education:

Minimum 2:1 degree in Computer Science, Artificial Intelligence or similar

Subject knowledge:

Large Language Models

Essential attributes:

- Strong background in NLP and machine learning
- Knowledge of evaluation metrics
- Excellent programming skills
- Attention to detail