iCIIT CONCLAVE 2024

LARGE LANGUAGE MODELS AND GENERATIVE AI

BRINGING TOGETHER PRACTITIONERS,
RESEARCHERS AND LEARNERS FROM SRI LANKA AND THE REGION

PROCEEDINGS

30TH & 31ST MAY 2024 AT THE INFORMATICS INSTITUTE OF TECHNOLOGY,
SPENCER BUILDING, 435 GALLE ROAD, COLOMBO 03, SRI LANKA.

INFORMATICS INSTITUTE OF TECHNOLOGY
57, RAMAKRISHNA ROAD,
COLOMBO 06,
SRI LANKA.
WWW.ICIIT.IIT.AC.LK

Disclaimer

The responsibility for opinions expressed in articles, in studies and other contributions in this publication rests solely with the respective authors. The iCIIT conclave 2024 on Large Language Models and Generative AI of Informatics Institute of Technology shall have no liability or responsibility to any person or entity regarding any loss or damage incurred, or alleged to have incurred, directly or indirectly, by the information contained in this book.

Published By:

Informatics Institute of Technology

57, Ramakrishna Road,

Colombo 06,

Sri Lanka.

www.iciit.iit.ac.lk

ISSN: 3051-5483

iCIIT Conclave on Large Language Models and Generative AI

The ICIIT Conclave 2024 was organized to bring together academia, researchers, industry professionals, and practitioners on a single platform to discuss a research topic in depth. Our objectives were to attract IIT researchers, other university and international researchers, industry professionals, and practitioners; to ensure the continuity of research ideas and their development; to build interest among like-minded individuals; and to disseminate knowledge across different segments of people. To achieve these goals, we decided to change the format of our traditional research meetups and introduced a new format, calling it a research conclave.

The conclave featured a shared task with several sub-tasks tackled by different groups, tutorial sessions to impart relevant knowledge, a keynote address, and two invited speeches discussing the subject area at a high level. Additionally, there were abstract presentations from researchers, academics, practitioners, and industry experts, followed by a panel discussion that allowed the audience to clarify doubts with an eminent panel of experts.

Post-event, we observed that we successfully attracted researchers from IIT and other universities, with representations from state universities, private universities, international universities, and the industry. We initiated the building of a Sinhala Language LLM through our shared task, where various groups presented their approaches to implementing sub-tasks. One group of participants committed to following up on the research area in their conference in November, ensuring continuity of the discussion. Furthermore, we organized three tutorials for beginners and practitioners, providing adequate knowledge on generative AI and LLMs.

Overall, we believe the conclave met several of our objectives and has laid a foundation for moving forward to the next level. The proceedings of the abstracts, along with the keynote address and invited speeches, are attached for your perusal. We hope you enjoy reading this research work, gain valuable insights, and continue the good work in research.

Table of Contents

Agenda

Start Time	End Time	Workshop Initiation	
8.00 AM	9.00 AM	Registrations of Participants	
9.00 AM	9.00 AM	Audience: call to order	
9.00 AM	9.05 AM	Welcoming the Audience	
9.05 AM	9.15 AM	Welcome & Intro by the Chair of the Conclave: Dr. Ruvan Weerasinghe :	
9.15 AM	9.25 AM	Address by the CEO: Mr. Mohan Fernando	
9.25 AM	10.10AM	**Keynote speech by: Dr. Romesh Ranawana**	
10.10AM	10.55AM	**Invited Talk by: Prof Nirmalee Wiratunga and Dr. Stuart Massie (RGU)**	
10.55 AM	**11.25 AM**	**Tea Break**	
11.25 AM	*11.30AM*	*Breaking out to two sessions*	
		Track 1: Industry Applications (Morning Session)	**Track 2: NLP (Morning Session)**
11.30 AM	11.45AM	**Theekshana Samaradiwakara**: Paper ID **9**: *Enhancing Stock Trading Education and Accessibility Through a Chatbot Application*	**KelaniyangodaGamage S Danoja**: Paper ID 12: *Identifying the Authenticity of Misleading Information in Social Media Using Large Language Models*
11.45 AM	12.00Noon	**Samudra N Kanankearachchi**: Paper ID 16: *Revolutionizing Analytics Task Augmentation with LLM and Retrieval Augmented Generation: A Case Study of OptiMaxer AI*	**Nabeelah Faumi**: Paper ID 21: *Enhance Authorship Attribution through Machine Learning and Stylometry in English and Romanized Sinhala*
12 Noon	12.15 PM	**Anton Erantha Jayakody**: Paper ID 8: *Zero to Hero: Enhancing Zero-Shot Accuracy in Low Parameter LLMs Through Prompt Engineering and User Sentiment Integration*	**Kyle Martin**: Paper ID 11: *Fine-Tuning Language Models for Error-Correction of Automatic Speech Recognition Applied to Clinical Dialogues*
12.15 PM	12.30 PM	**Sandaru R Tissera**: Paper ID 13: *Enhancing Cryptocurrency Disinformation Detection through Fine-Tuning and RAG-Based Data Labelling with LLM*	**Theekshana Samaradiwakara**: Paper ID 10: *Long Text Summarization Using Open Source Models*
12.30 PM	12.45 PM	**Muljayan S Jalangan**: Paper ID 7: *Utilizing RAG and Prompt Engineering for Categorization & Summarization of Judgments in the Sri Lankan Jurisprudence*	**Sahan Wewelwala**: Paper ID 6: *Enhancing Audio Transcription Accuracy with Large Language Models*
12.45 PM	1.00 PM	**Sandares Dhanujaya**: on behalf of **Randil Pushpananda:** Paper ID 15: *LLM-driven Sinhala voice-enabled Banking chatbot*	**Deshan K Sumanathilaka**: Paper ID 17: *Can AI Tackle Ambiguity in Natural Text: Quantitative Evaluation on LLMs for WSD*
1.00 PM	**2.00 PM**	**Lunch Break**	

		Track 3 - Educational Technology (Evening Session)	Track 4: Other Applications (Evening Session)
2.00 PM	2.15 PM	**Damayanthi K Herath**: Paper ID 3: *The role of AI based tools in academic research: insights from an engineering undergraduate research project course*	**Ransika Costa**: Paper ID 14: *A Novel Approach to Ingredient Substitution in Food Recipes*
2.15 PM	2.30 PM	**Nirhoshan Sivaroopan**: Paper ID 26: *Generative AI-Powered Intelligent Assistant for Internal Developer Platform Users*	**Ishara Neranjana**: Paper ID 18: *Assessing the Effectiveness and Challenges of RAG Applications: A Comprehensive Testing Framework*
2.30 PM	2.45 PM	**Harshani Madhushani Bandara**: Paper ID 20: *A Novel Approach to Generating Customised Math Word Problems*	**Chanuka R Algama**: Paper ID 23: *Conversational RAG with Memory-Based Context Enhancement*
2.45 PM	*3.00 PM*	*Relocating to Main room - Anounced by Comperes*	
3.00 PM	4.00 PM	Panel Discussion - Prof. Nirmalee Wiratunga, Dr. Srinath Perera, Dr. Kyle Martin, Dr. Stuart Massie - Moderator Dr. Ruvan Weerasinghe	
4.00 PM	4.10 PM	Closing Remarks by Chair	
4.10 PM	4.25 PM	Awards and Certificates - *Anounced by Comperes*	
4.25 PM	4.30 PM	Vote of Thanks - Head of Research, Dr. Dinesh Arunatileka	
4.30 PM	**5.30 PM**	**Tea and Fellowship**	

Committee Members

Conclave Chair:

- Dr. Ruvan Weerasinghe

Organizing Committee

- Prof. Nalaka Wickramasinghe, Business School, Informatics Institute of Technology, Sri Lanka
- Dr. Dinesh Arunatileka, Research Unit, Informatics Institute of Technology, Sri Lanka
- Mr. Sudarshan Welihinda, Informatics Institute of Technology, Sri Lanka
- Mr. Ashwaq Ahmed, Informatics Institute of Technology, Sri Lanka

Programme Committee

- Ruvan Weerasinghe – School of Computing, Informatics Institute of Technology, Sri Lanka
- Nirmalie Wiratunga — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Stewart Massie — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Kyle Martin — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Ike Nkisi-Orji — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Nipuna Senanayake – School of Computing, Informatics Institute of Technology, Sri Lanka
- Prasan Yapa – Ubiquitous and personal computing lab, Graduate School of Engineering, Kyoto University of Advanced Science, Japan
- Isuri Anuradha – UCREL NLP Group, School of Computing and Communications, Lancaster University, UK
- Iresh Bandara – Perception Engineering Research Group, Faculty of ITEE, UBICOMP, University of Oulu, Finland
- Deshan Sumanthilaka – Department of Computer Science, Swansea University, Wales, UK

Review Committee

- Ruvan Weerasinghe – School of Computing, Informatics Institute of Technology, Sri Lanka
- Nirmalie Wiratunga — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Stewart Massie — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Kyle Martin — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Ike Nkisi-Orji — Artificial Intelligence and Reasoning Group, Robert Gordon University
- Nipuna Senanayake – School of Computing, Informatics Institute of Technology, Sri Lanka
- Prasan Yapa – Ubiquitous and personal computing lab, Graduate School of Engineering, Kyoto University of Advanced Science, Japan
- Isuri Anuradha – UCREL NLP Group, School of Computing and Communications, Lancaster University, UK

- Iresh Bandara – Perception Engineering Research Group, Faculty of ITEE, UBICOMP, University of Oulu, Finland
- Deshan Sumanthilaka – Department of Computer Science, Swansea University, Wales, UK

Keynote Speaker

Dr. Romesh Rananwana

Group Chief Analytics and AI Officer at Dialog Axiata PLC

Dr. Ranawana is the Chairman of the National Committee to Formulate an AI Policy and Strategy for Sri Lanka, established by the President in 2023. He held the position of Chairman at the SLASSCOM AI Center of Excellence, where he contributed to the development and adoption of AI in Sri Lanka. His expertise has led to his involvement in various boards and committees, including the University of Colombo School of Computing and the Open University of Sri Lanka. Currently, he holds the position of Group Chief Analytics and AI Officer at Dialog Axiata PLC, where he provides strategic guidance and leadership for all Analytics and AI initiatives. His career journey began with his role as a Co-founder, CTO, and Managing Director of SimCentric Technologies (Pvt) Ltd, a company that emerged as a prominent player in the industry under his guidance. He then served as the CTO of Tengri UAV, driving significant advancements in unmanned aerial vehicle technology. Recognizing the growing importance of AI, he founded Enterprise Machine Learning (Pvt) Ltd, a consultancy firm that offers strategic guidance on AI transformation to multinational organizations. He passed his BSc with First Class Honours at University of Peradeniya and completed his PhD at the University of Oxford, UK.

Title: Enhancing Generative AI with Contextual Intelligence: The Role of Case-Based Reasoning in LLMs

Romesh Ranawana, Dialog Axiata PLC

Since the seminal 2017 'Attention Is All You Need' paper, the usability of Large Language Models (LLMs) utilising transformer architectures have surged, requiring progressively less data, time, and expertise due to advances in fine-tuning and prompted modelling techniques. Recent human-in-the-loop methodologies have further democratised their use. Despite their fluency, LLMs often struggle with content accuracy in precision-critical domains. This limitation largely stems from the models' lack of constructed memory and inability to track task completion across interactions. To address this, integrating LLMs with case-based reasoning could prove beneficial. This methodology could help structure, index, and retain memories of contextually relevant knowledge, enhancing the LLMs' ability to provide accurate responses. In this talk, we will present two case studies demonstrating how Case-Based Reasoning (CBR) can be used to capture context for reasoning and improve content accuracy in precision critical applications. Specifically, we will explore 1) a Legal Q&A system enhanced with Retrieval Augmented Generation, which uses retrieved legal cases to inform current responses, and 2) a data-to-text Natural Language Generation (NLG) system for producing accurate, context-aware summaries for sports commentaries.

Invited speaker

Prof. Nirmalie Wiratunga

Professor in Intelligent Systems at Robert Gordon University, United Kingdom

Nirmalie Wiratunga is a Professor in Intelligent Systems at RGU's School of Computing, and the Associate Dean for Research in the school, with over two decades of experience in computer science and AI research. She has held positions such as post-doctoral researcher on EPSRC-funded projects and was appointed Readership in 2009, and Professorship in 2016. Nirmalie is also an adjunct IDUN professor at the Norwegian University of Science and Technology.

Nirmalie leads the Artificial Intelligence & Reasoning Research Group (AIR) in the School of Computing. She has been involved in numerous funded AIR projects, including the development of human-centered AI platforms for explanation strategy recommendation; and the use of Case-Based Reasoning for Retrieval Augmented Q&A systems using LLMs. Additionally, she co-founded the Attendr.app, a spinout for mobile attendance tracking of students at the university and attendees at conferences.

Title: CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs

Invited speaker

Dr. Stewart Massie

Stewart received his PhD (2006) on Knowledge Management for Case-Based Reasoning Systems from RGU, where he continues to work as a Reader with the Artificial Intelligence (AI) and Reasoning research group. He has more than 15 years of research experience in AI developing improved machine learning, information retrieval, recommendation, and data mining technologies. His research applies these technologies to the development of applied solutions for text and multi-media applications, as well as more recently for sensor networks. Stewart has published over 90 peer-reviewed papers in leading journals and conferences including Artificial Intelligence, IJCAI and AAAI; and serves as PC member for several international conferences including KSEM, ICCBR, ECIR, EANN and IJCAI.

Title: Case-Based Reasoning Approaches to Data-to-Text Generation

iCIIT LLM Tutorials

Hands-On Low Code/No Code LLM Tutorial

Resource Persons:

- Prasan Ratnayake, Data Science Engineer at Zone24x7
- Umair Ramzaan, Senior Data Science Engineer at Zone24x7

Hands-On LLM Tutorial: Building LLM Apps with Langchain

Resource Persons:

- Ishara Neranjana, Machine Learning Engineer at Zone24x7
- Nisal Mihiranga, Head of AI and Data Science at Zone24x7

MLOps/LLMOps Tutorial for Practitioners

Resource Persons:

- Chamika Ramanayake, Head – AI Platforms at Dialog Axiata PLC
- Praveen Dhananjaya, Machine Learning Engineer, Dialog Axiata PLC

Workshop Presentations

Track 1: Industry Applications

Enhancing Stock Trading Education and Accessibility Through a Chatbot Application

Theekshana Samaradiwakara, Sandaruth Siriwardana

IronOne Technologies, Sri Lanka

Stock trading within the Colombo Stock Exchange (CSE) is a complex domain, often requiring access to real-time information, understanding of market dynamics, and adherence to trading regulations. To facilitate access to educational resources and streamline the trading process, a chatbot application has been developed. Leveraging the open-source Llama, Falcon and Mistral models with the RAG architecture (Lewis et al., 2020) and Langchain framework, the chatbot integrates with stock trading applications and provides real-time stock price monitoring, comparison, and self-service functionalities.

Data sources include user manuals, articles, CSE rules, and stock trading website documentation. Integration with a database allows users to access real-time stock prices and perform complex queries, such as comparing stock prices or calculating the net growth of prices.

Efforts to enhance response quality involved experimenting with different data cleaning methods, text splitting methods, vector databases and software architectures in order to reduce costs and overhead, along with to ensure accurate and controllable outputs.

The system is initially run on a local server, employing different model formats like GPTQ, GGUF , AWQ and techniques (Dice, D. and Kogan, A. 2021) like Huggingface Accelerate, Optimum and Llama.cpp to minimize overall overhead.

Future plans include fine-tuning the model for the stock exchange domain and adapting it for local language use, aiming to provide a more tailored and efficient user experience.

Revolutionizing Analytics Task Augmentation with LLM and Retrieval Augmented Generation: A Case Study of OptiMaxer AI

Samudra Kanankearachchi, Charuka Rathnayake, Srilal Siriwardhana, Skenos Peniel

99X, Sri Lanka

99X research labs product OptiMaxer AI has consistently pioneered the integration of LLM base architectures into practical business applications, significantly enhancing operational efficiencies across multiple sectors. With a robust portfolio ranging from natural language-driven ERP systems to sophisticated recommendation engines and data extraction from both structured and unstructured data, our latest innovation involves the deployment of Retrieval Augmented Generation (RAG) within our analytics and Business Intelligence (BI) architecture.

This transformative approach leverages the synergy between retrieval-based methods and the generative capabilities of Large Language Models (LLMs) to furnish a dynamic, scalable, and highly accurate analytics framework. By embedding RAG into our systems, OptiMaxer AI not only refines data analysis processes but also ushers in a new era of data-driven decision-making.

In the realms of sales, retail, healthcare, finance, education, and technology, RAG's impact is profoundly felt. It introduces unprecedented accuracy and depth to data interpretation by continuously integrating the most relevant and recent information into its analysis. This capability is particularly crucial in fast-paced environments where timely and context-aware insights are essential for maintaining competitive advantages.

The architecture of our RAG-based OptiMaxer AI Analytics system is designed to be inherently adaptive, supporting real-time data processing and complex pattern recognition. This facilitates not just routine decision-making but also enables our clients to proactively respond to emerging trends and challenges effectively.

Through practical deployments and controlled experiments, we demonstrate how RAG-based task augmentation has led to a marked improvement in predictive accuracy, operational efficiency, and customer satisfaction across all sectors engaged. Our findings underscore the potential of RAG to not only keep pace with but also anticipate and adapt to the rapidly evolving landscape of global data.

Zero to Hero: Enhancing **Zero-Shot Accuracy in Low Parameter LLMs Through Prompt Engineering and User Sentiment Integration**

Anton Jayakody

National School of Business Management

Artificial Intelligence (AI) along with Natural Language Processing (NLP), have made a revolutionizing impact in technology in recent years. The usability of Large Language Models (LLM) has become a primary focus amongst enthusiasts. However, without sophisticated computational power and sufficient finances LLMs are restricted from being used by individual developers. This study examines how a prompt engineering pipeline enhances zero-shot accuracy of open-source low parameter LLMs. The proposed approach involves customizing and automating prompts to facilitate prompt-based learning opportunities for the model. Additionally, it evaluates the impact of the integration of sentiment-state of the user to the prompt. The paper presents a detailed analysis of comparisons and experimental processes conducted to evaluate the effectiveness of this approach. The findings demonstrate that strategically integrating prompt engineering techniques can lead to greater success even in lower-tier LLM such as the BLOOM-500m, offering new perspectives in the field of NLP.

Enhancing Cryptocurrency Disinformation Detection through Fine-Tuning and RAG-Based Data Labelling with LLM

Sandaru Tissera, Deshan Sumanathilaka*

Informatics Institute of Technology, Sri Lanka

*Swansea University, United Kingdom

Cryptocurrency Disinformation is a problematic concern in social media nowadays, especially in platform X, one of the most popular social media platforms utilized by crypto enthusiasts (Mirtaheri et al., 2021). Being a trending topic for the past decade, this has given rise to various forms of disinformation and manipulation, which significantly impact both investors and the global cryptocurrency market (Rangapur, Wang and Shu, 2023). Addressing this prevailing issue is vital to prevent market manipulations, protect investors and maintain market integrity throughout the cryptocurrency space.

This work involved developing a computational approach to automatically detect disinformation in cryptocurrency tweets by classifying tweet data based on the context provided. To overcome the challenge of the scarce of relevant labelled data, data annotation and labelling were carried out using few-shot prompting utilizing Retrieval Augmented Generation (RAG) (Lewis et al., 2021). Figure 1 illustrates the overview of the RAG pipeline used for the annotation process.

The proposed system was implemented by fine-tuning a GPT 3.5 Turbo model focusing on the edge cases where the base model underperformed in the provided context. An evaluation was conducted on a sample benchmark dataset of human-annotated data. It revealed an accuracy of 0.85 for the GPT-3.5 Turbo Base Model. An iterative human-in-loop mechanism was employed alongside the prompt augmentation approach to identify the optimal prompt, achieving 0.90. The fine-tuned model achieved an accuracy of 0.92, surpassing the other models in terms of accuracy.

This advancement not only contributes to a more precise detection of cryptocurrency disinformation but also lays the groundwork for deeper authenticity in the social media context.

Utilizing RAG and Prompt Engineering for Categorization & Summarization of Judgments in the Sri Lankan Jurisprudence

Muljayan Jalangan, Deshan Sumanathilaka*

Informatics Institute of Technology, Sri Lanka

*Swansea University, United Kingdom

Efficient searching of Sri Lankan court judgments, poses significant challenges due to lack of structured categorization and these documents being dispersed across various platforms further exacerbate the problem. Moreover, legal professionals grapple with the time-consuming task of reviewing these lengthy documents. To address these challenges, this research endeavour delves into the innovative integration of Retrieval-Augmented Generation(RAG) and Prompt Engineering methodologies for systematically categorizing and summarizing judgments within the Sri Lankan Jurisprudence by harnessing the power of LLMs. The data sourced from PDF files on the official websites of Sri Lanka's Court of Appeal and Supreme Court undergoes processing through a custom pipeline to extract and store relevant information. Further processing includes chunking, embedding, and storing the embeddings for efficient context-driven retrieval. Central to our approach is the human-in-the-loop step, utilizing prompt engineering and a sample list of categories to facilitate zero-shot prompting to enhance output accuracy and provide the output in JSON format. Our LLM-agnostic approach ensures compatibility with various models accessible through freely and inexpensively available APIs offered by OpenAI, Google, and TogetherAI, allowing experimentation to optimize outputs effectively. Preliminary findings show promising accuracy and summarization efficiency, with over 5000 judgments categorized and summarized within 100-250 words. By contextualizing our research within the Sri Lankan legal framework and leveraging RAG and Prompt Engineering techniques, we have observed enhanced adaptability to its legal language intricacies, facilitating more nuanced and contextually relevant analyses.

LLM-driven Sinhala voice-enabled Banking chatbot

P. L. A. Shermi Maleesha, Randil Pushpananda.

University of Colombo, School of Computing, Sri Lanka

The language inclusion of banking services is still a crucial concern in an era where digital banking is becoming more and more common, especially for people who do not understand English. To bridge the linguistic gap in Sri Lanka's banking industry, this paper presents a novel voice-enabled banking chatbot system in the Sinhala language. Through the utilization of available Large Language Models, cutting-edge translation APIs and the Rasa framework, this research leads the way in creating a chatbot that can comprehend and reply to user inquiries in Sinhala, thereby greatly improving the accessibility of banking services for Sinhala speakers.

The comprehensive integration of natural language processing techniques used in this research's approach entails converting Sinhala voice to Sinhala text (questions) through the developed Sinhala speech recognition system, translating Sinhala questions into English, processing those queries to provide pertinent responses, and then translating those responses back into Sinhala and finally, converting Sinhala text to Sinhala voice through the developed Sinhala Text to Speech system.

The results of this study could completely change the way banking services are provided to Sinhala-speaking communities by offering a model of linguistic inclusion that can be applied to a variety of languages, especially Tamil and English. This research provides a roadmap for future developments in multilingual chatbot systems and offers insightful information about how language-specific LLMs might be integrated into chatbot development.

Track 2: Natural Language Processing

Identifying the Authenticity of Misleading Information in Social Media Using Large Language Models

Kelaniyangoda Gamage Semini Danoja, Deshan Koshala Sumanathilaka*

Informatics Institute of Technology

*Swansea University, United Kingdom

In recent years, the rapid development and widespread adoption of social media platforms have changed the way information is distributed and consumed. Studies show that about 65% of content shared on platforms such as Facebook, Twitter, and Instagram is false. Social media has also raised concerns about the spread of various forms of misleading information, such as misinformation, disinformation and fake news, as well as their impact on public discourse and the possible spread of harmful content. To address the critical problem of unregulated misleading information that poses significant risks to public discourse and democratic integrity, this project has developed a deep learning model that utilizes the capabilities of advanced large language models (LLMs). By using state-of-the-art LLMs such as BERT and GPT, our model increases the precision in distinguishing authentic from misleading content, significantly improving traditional detection methods. In this methodology, we have integrated two distinct parameters derived from a Bi-directional GRU trained on Fake News detection and subjectivity analysis from the TextBlob library. The classification output, in conjunction with the subjectivity score, is employed to prompt the inference process of the Large Language Model. Implementing LLMs based approach enables nuanced analysis of text data and captures subtle clues and patterns characteristic of false information. Our approach not only improves the precision in distinguishing misinformation and disinformation from false content but also represents a methodological advance and provides practical solutions essential to supporting informed decision-making and preserving the integrity of processes. This research highlights the need for robust tools to combat the pervasive spread of false information in the digital age and highlights their potential to strengthen the media landscape and contribute to both academic and public discourse.

Enhance Authorship Attribution through Machine Learning and Stylometry in English and Romanized Sinhala

Nabeelah Faumi , Adeepa Gunathilaka, Benura Wickramanayake, Deelaka Dias, Deshan Sumanathilaka*

Informatics Institute of Technology, Sri Lanka

*Swansea University, United Kingdom

The rise of Web 2.0 and global communication has indeed left its mark on society, yielding positive and negative consequences. However, the widespread dissemination of misinformation through counterfeit channels has unfortunately underrated some of these positive impacts (Marulli et al. 2021). Authorship attribution is a crucial factor for content verification in academia, cybercrime, and copyright infringement, requiring significant improvement due to its diverse nature (Wilson 2015). To cater for this existing gap, this study proposes a novel architecture for English and Romanized Sinhala Authorship attribution using computational linguistics techniques.

The proposed author attribution system offers a unique approach by comparing two sets of text, suspect author and anonymous text, a departure from conventional approaches reliant on extensive data corpus. Based on the identified feature set, a random forest classifier (RFC) determines if the texts belong to the same author. It also pioneers authorship attribution for Romanized Sinhala, potentially paving the way for similar studies in Romanized versions of other languages. Unlike ordinary languages, Romanized text lacks a standardised vocabulary. The Large Language Model (LLM) reasoning capabilities have been employed to identify the pattern of English words mixing in Romanized Text of the suspect author's content. This task directly contributes to the model by emphasising the writing patterns and the character usage in the Romanized Sinhala context. This task was performed using few-shot learning, which was designed using a human-in-loop approach. The study's principal researcher used prompt augmentation to identify the required prompt. Further feature extractions, like determining the type of English the user uses, were explored using prompting. Identified feature sets from the LLMs are shared with the RFC to perform the final classification of the content authorship.

Expanding the scope of authorship attribution to diverse linguistic contexts, this research pioneers author attribution in English and Romanized Sinhala, which is crucial in Sri Lanka. It advances computational linguistics, enhancing understanding of cross-linguistic authorship patterns and promoting trust and ethical practices in digital communication.

Fine-Tuning Language Models for Error-Correction of Automatic Speech Recognition Applied to Clinical Dialogues

Gayani Nanayakkara, Kyle Martin, Nirmalie Wiratunga, David Corsar

Robert Gordon University

Clinical dialogue is a conversation between health practitioners and their patients, with the explicit goal of obtaining and sharing medical information for the purposes of diagnoses and treatment. Current practice is heavily reliant on manual scribing, which is time-consuming and error-prone, thus leading to inefficiencies in the healthcare system. Automatic Speech Recognition (ASR) systems show potential for non-intrusive recording of clinical dialogue. However, there exists a number of general and domain-specific challenges that must be resolved before ASR could be applied for this purpose. In particular, there is significant risk of mis-transcription leading to medical errors. While post-hoc methods for error correction are growing more common, performance suffers from the lack of domain-specific vocabulary and the mismatch between error correction and pre-training objectives.

In this talk, we will describe our work using language models for error-correction of medical transcripts within the Gastrointestinal specialism. We will discuss a comparative evaluation and analysis of existing commercial ASR systems, underpinning a discussion on ASR error types. Finally, we describe the development of novel finetuning methods for sequence-to-sequence language models, leading to reduced word error rates across 3 public and 1 private dataset. Ultimately, we demonstrate that our mask-filling objective specialised for the medical domain (med-mask-filling) outperforms the best performing commercial ASR system by 10.27%.

Long Text Summarization Using Open Source Models

Theekshana Samaradiwakara, Ravindi Weerasinghe

IronOne Technologies, Sri Lanka

Annual reports and board papers of business organizations contain crucial information for stakeholders. However, due to the sensitive and confidential nature of this information and the extensive length of these documents, we cannot rely solely on third-party LLM endpoints such as ChatGPT despite the capability of providing accurate and high-quality summaries. This project aims to create a transformer-based application that can effectively summarize lengthy business reports while operating on organizations' local servers to ensure data privacy. In this work, we prepared a dataset for the project by collecting publicly available annual reports from the Internet and segmenting these reports into distinct sections (e.g., financials, operations, management discussions). Then, we utilized the open-source Llama, Falcon and Mistral models to generate label summaries for supervised fine-tuning. The design decision for used models was based on the quality of the generated summaries and the cost-effectiveness of open-source resources.

For this project, we selected Google's Long T5 (Guo et al., 2022), BART (Mike et al., 2019) models for summarization, considering its ability to create good-quality summaries, increased context with an optimized attention mechanism and the model size. We fine-tuned the model using a subset of our dataset, and the performance evaluation indicates an improvement in model performance compared to the base models.

Future steps will involve further fine-tuning the model utilizing the entire dataset and optimizing the model to generate accurate and high-quality summaries adhering to business use cases.
This initiative empowers organizations to harness advanced NLP technologies to make informed decisions while upholding strict control over proprietary corporate data.

Enhancing Audio Transcription Accuracy with Large Language Models

Sahan Hewage Wewelwala, Deshan Sumanathilaka*

Informatics Institute of Technology, Sri Lanka

*Swansea University, United Kingdom

The predominant obstacle in the field of automated audio transcription is attaining a high level of precision, particularly in varied and cacophonous settings. Precise transcriptions are essential for efficient communication in several fields, including legal procedures, medical paperwork, and real-time communication aids for the hearing impaired. This study investigates the use of large language models (LLMs) to improve the accuracy of transcribing. LLMs, along with their advanced deep learning capabilities, present intriguing opportunities for comprehending and manipulating intricate language patterns and audio information across diverse circumstances. The methodology it employs is instructing a LLM using an extensive dataset that includes various dialects, accents, and background sounds. The objective is to enhance the model's resilience and flexibility. In addition, we incorporate sophisticated noise reduction techniques to preprocess audio streams, resulting in improved clarity and precision of model inputs. The results suggest a notable enhancement in the accuracy of transcribing speech. The model has shown improved skills in understanding speech in difficult acoustic conditions when compared to current transcription technologies. The ramifications of these developments go beyond simple technological improvements. They pledge significant enhancements in accessibility for those with hearing impairments, increased effectiveness in professional environments that necessitate word-for-word documentation, and more dependable resources for language interpretation. The incorporation of Language Models (LLMs) has the potential to greatly enhance the accuracy and reliability of automated transcription services, which are in high demand. This advancement might transform these technologies into essential tools in our digital society.

Can AI Tackles Ambiguity in Natural Text: Quantitative Evaluation on LLMs for WSD

T G Deshan K Sumanathilaka, Nicholas Micallef, Julian Hough

Swansea University, United Kingdom

Social media posts and other digital communications are widespread with ambiguous words that can hold multiple meanings. Solving "lexical ambiguity" is a challenging task with traditional computation methods (Yadav, Patel and Shah, 2021). These methods struggle with limited data and a lack of contextual understanding, leading to the consequences of poor Word Sense Disambiguation (WSD). As a result, translation tools, information retrieval systems, and question-answering technology all suffer, hindering the potential of AI in the social media sphere (Mente, Aland and Chendage, 2022).

This research explores how contextual understanding power of Large Language Models (LLMs) can be used to improve WSD. The study proposes a new approach that combines two key elements namely Prompt Augmentation and Knowledge Base inferencing approach which contains different instances of ambiguous words. During prompt augmentation phase, humans play a major role in identifying the optimal prompt for the study. Augmented prompts are supported by part-of-speech tags, synonyms, and aspect-based sense filters to guide the LLM towards the intended meaning. This "human-in-loop" approach combined with "few-shot prompting" (providing just a few examples) improves the performance of disambiguation process. The proposed approach was evaluated on FEWS few-shot test set using both commercial and Open source LLMs (Blevins, Joshi and Zettlemoyer, 2021). The results on GPT-4-Turbo yield significantly improved accuracy, opening a new avenue for research in word sense disambiguation.

This research paves the way for more accurate understanding of language in social media and beyond. With this new method, translation systems can become more nuanced, search results more relevant, and question-answering technology more insightful. Ultimately, it contributes to better communication in our increasingly digital world.

Track 1: Educational Applications

The role of AI based tools in academic research: insights from an engineering undergraduate research project course

Damayanthi Herath

University of Peradeniya, Sri Lanka

Large Language Models (LLMs) are expected to improve productivity in multiple tasks in research. Subsequent to the release of LLMs, multiple studies have been conducted looking at student perceptions and the ethical considerations on the use of Artificial Intelligence (AI) based tools in teaching and learning of courses such as medicine, arts and psychology. However, the same relevant to undergraduate engineering research projects have not been explored yet. In this study, we conduct an empirical analysis on the students' usage of AI based tools during the delivery of an undergraduate research project course in engineering. In the delivery of the course, the students were allowed to use AI based tools, and assessments included consideration of the same. Written feedback was collected from the students on their usage via 4 main questions: What were the tools used, how they were used, the opportunities and challenges observed and how they optimized their time. A qualitative analysis of the feedback demonstrates that the grammar checking tool stands out as the mostly used tool. The tools are used to summarize, improve the grammar and code in students' research work. The main concern among the students is identified as the reliability of the tools. This will be combined with a quantitative analysis of the student feedback conducted by adopting a unified theory of acceptance and use of the technology model. It will enable us to gain valuable insights into the integration of AI based tools in students research work and improve the teaching and learning of research.

Generative AI-Powered Intelligent Assistant for Internal Developer Platform Users

Nadheesh Jihan Jayawickramage, Nirhoshan Sivaroopan, Sachini Ranasinghe, Isuru Ruhunage, Ayesh Weerasinghe, Malith Jayasinghe

WSO2

The advent of Generative AI, especially Large Language Models (LLMs), has fundamentally transformed natural language processing (NLP). By amalgamating the capabilities of LLMs, external knowledge, and actions, we can develop robust tools. Conversely, by integrating a proficient assistant, the user experience and productivity within an internal developer platform (IDevP) can be significantly augmented. Acknowledging the necessity for expert agents to aid users across diverse domains and tasks—ranging from user onboarding to efficiently exploring runtime metrics and logs, automated debugging, diagnostics, and troubleshooting of user deployments—we employ Retrieval-Augmented Generation (RAG) and (Reason and Act) techniques to construct expert agents tailored to each domain, capable of engaging with users through natural language interactions. To seamlessly integrate domain expert agents, we adopt a multi-layered agent framework, incorporating a manager agent to coordinate these expert agents in responding to user inquiries while maintaining a conversational history (memory). This augmentation ensures a unified user experience, simplifying the interaction process compared to direct engagement with expert agents. Furthermore, we employ advanced prompting techniques and architectural strategies, particularly focused on managing memory, enhancing reasoning, efficiently injecting knowledge and minimizing hallucination. Finally, we are implementing a feedback pipeline to gather user feedback continually, thus refining the assistant iteratively. This strategic methodology has empowered us to develop a lifelike conversational assistant capable of significantly enhancing the productivity of IDevP users.

A NOVEL APPROACH TO GENERATE CUSTOMIZED MATH WORD PROBLEMS

Harshani Bandara, Nimesh Ariyarathne, Yasith Heshan, Surangika Ranathunga, Omega Gamage*

University of Moratuwa, *ACCELR Logic

Mathematics is often perceived as a complex subject by students, leading to high failure rates in exams. One of the ways to improve mathematics skills is to provide sample questions for students to practice problem-solving. In this research, we present a novel approach to generating customized Math Word Problems(MWPs) across different grades and question types using Large Language Models (LLMs). Our system is designed to operate efficiently on low hardware resources while maintaining high-quality problem generation. We chose Llama 2 as our baseline which is capable of generating MWPs of higher quality among open-source LLMs. We created a dataset named “Mathwizards” that contains 4k MWPs covering all the question types available from grades 1 to 6 by following the USA math syllabus named “Common Core Standards for Mathematics”(Mathematics standards). We employed various prompt engineering techniques and finetuned the LLM following the instruction tuning technique to improve the quality of the generated MWPs. We applied diversity improvement techniques to generate more diverse MWPs. Additionally, we integrate few-shot prompting and Direct Preference Optimization(DPO) methods to enhance the system's adaptability and performance.

The quality evaluation of the generated MWPs is conducted through a hybrid approach, combining human evaluation with LLM-based assessments. This comprehensive evaluation methodology ensures the quality and relevance of the generated problems, aligning with educational standards. At the end of our experiments, we achieved an accuracy of 93.10%. Our research contributes to the advancement of personalised educational tools by providing a resource-efficient solution for generating customised MWP.

Track 2: Other Applications

A NOVEL APPROACH TO INGREDIENT SUBSTITUTION IN FOOD RECIPES

Kumuthu Athukorala, Thevin Senath, Ransika Costa, Surangika Ranathunga, Rishemjit Kaur*

University of Moratuwa, Sri Lanka

*CSIR-Central Scientific Instruments Organisation, India

Recipe personalization through ingredient substitution holds promise in meeting diverse dietary needs, accommodating preferences, avoiding potential allergens, and navigating culinary exploration. Using Large Language Models (LLMs) with prompting has become the current de-facto solution in Natural Language Processing (NLP), yet the domain of ingredient substitution remains largely unexplored. In this study, we introduce a novel approach utilizing prompt engineering techniques to facilitate ingredient substitution within recipes. Our methodology employs LLMs including Llama 2 7B base and chat models, Mistral 7B, and Gemma 7B, both base and instruct models. Through prompt-based strategies such as in-context learning, instruction tuning, and prompt patterns, we address ingredient substitution considering recipe context. Fine-tuning these selected LLMs with the Recipe1MSubs dataset (Fatemi et al, 2023), which consists of substitution pairs with standardized splits, using different Parameter Efficient Fine Tuning (PEFT) methods like LoRA & QLoRA, we highlight Mistral 7B as optimal. Further refinement of the Mistral 7B model through techniques such as multi-task learning, direct preference optimization, 2-phased fine-tuning, and full fine-tuning without PEFT aims to enhance model performance. So far, our experimental validation demonstrates that our approach surpasses the baseline GISMO (Fatemi et al, 2023), currently the best method in ingredient substitution, particularly in terms of hit@1 accuracy. This research not only advances the field of recipe personalization but also underscores the efficacy of prompt-based strategies in tackling challenges within NLP.

Assessing the Effectiveness and Challenges of RAG Applications: A Comprehensive Testing Framework

Ishara Neranjana, Nisal Mihiranga, Kanishka Wijayasekara

Zone24x7

Retrieval Augmented Generation (RAG) applications have emerged as powerful tools for generating text based on retrieved information. As these applications become increasingly integral to various domains, there arises a need to rigorously assess their effectiveness and address challenges inherent in their deployment. This study investigates methods for assessing Retrieval Augmented Generation (RAG) applications. It focuses on evaluating their effectiveness, addressing challenges, and ensuring production readiness. A comprehensive testing framework analyzes application performance in text generation tasks, emphasizing response quality and overall performance metrics. While benchmarks exist for Large Language Models (LLMs), assessing custom applications for real-world scenarios requires a specific framework. This framework systematically evaluates the ability of the application based on LLMs to comprehend, generate, and manipulate natural language within the context of retrieval augmentation. It delves into challenges encountered in the retrieval process, covering aspects such as response truthfulness, relevancy, repeatability, and coherence, alongside context recall, and context precision of retrieved chunks Additionally, performance metrics like throughput, latency, and request rates are considered. The study proposes a comprehensive testing framework and strategies to mitigate these challenges and enhance LLM-based application performance. Evaluation results from the proposed framework, coupled with actionable recommendations, offer valuable insights for researchers, developers, and practitioners in generative AI. This study contributes to advancing the understanding of LLM application testing and provides a foundation for adapting existing LLM based applications to future advancements efficiently in this rapidly evolving domain.

Conversational RAG with Memory-Based Context Enhancement

Chanuka Algama

University of Kelaniya, Sri Lanka

Retrieval-augmented generation (RAG) which integrates retrieval capability into an LLM's text generation process, fetching relevant document snippets from a large corpus which the LLM then uses to produce answers, has demonstrated remarkable success in tasks requiring external information and contextual understanding. However, RAG also faces limitations, especially when applied to follow-up queries in a conversational setting, particularly in capturing context and understanding complex queries. The challenge lies in the difficulty of ensuring that the retrieved information accurately reflects the user's intent, as the proximity of text chunks in the embedding space does not guarantee a meaningful question-and-answer pair. At the same time, there lacks deployment solutions for organizational use other than foundations of RAG such as LangChain, LLAMAIndex, and PipeRAG. In this work, using Mistral-7B-Instruct-v0.1 as a base model, specifically its 4-bit quantized version, a plug-and-play conversational querying architecture is presented. The pipeline is divided into three primary phases: indexing, retrieval & generation, and saving conversation history into memory. Initially, documents are segmented into text chunks, and their embeddings are stored in a vector database, facilitated by the LlamaIndex framework. Subsequently, user queries are matched against these embeddings, prompting the LLM to generate responses based on the retrieved contexts. Finally, leveraging conversation memory with LLM temperature 0.0 to generate standalone questions for follow-up queries. Evaluation of the architecture performance is currently underway.

iCIIT 2024 CONCLAVE

LARGE LANGUAGE MODELS AND GENERATIVE AI

BRINGING TOGETHER PRACTITIONERS,
RESEARCHERS AND LEARNERS FROM SRI LANKA AND THE REGION

30TH & 31ST MAY 2024 AT THE INFORMATICS INSTITUTE OF TECHNOLOGY,
SPENCER BUILDING, 435 GALLE ROAD, COLOMBO 03, SRI LANKA.

9 773051 548005