Natural Language Processing Projects 2026-27

Contents

Exploring the Diversity of Data Storytelling: A Comparative Analysis Across Cultures, Disciplines, and Media

Supervisor: Alfie Abdul-Rahman

Areas: Human-centred computing (human-computer interaction), Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Data storytelling has become a vital practice for interpreting and communicating complex information, combining data analysis, narrative techniques, and visual representation. Yet the diversity of data storytelling approaches across cultures, disciplines, and media remains underexplored. This PhD project will investigate how cultural, disciplinary, and media-related factors—alongside technological mediation through natural language processing (NLP)—shape the design, ethics, and reception of data stories. It will integrate feminist and decolonial perspectives to critically examine questions of power, inclusivity, and representation in the creation and communication of data narratives.

Current research often isolates technical, narrative, or communicative dimensions of data storytelling, with limited attention to cross-cultural and cross-disciplinary comparisons. As NLP techniques such as summarisation, translation, and generative storytelling are increasingly deployed, they raise pressing questions about whose voices are amplified or silenced, and how automation may reproduce or resist systemic biases. Incorporating feminist and decolonial approaches makes it possible to interrogate epistemic injustices in data storytelling—asking not only how stories are told, but also whose knowledge counts.

This project will offer both theoretical and practical contributions. Empirically, it will conduct comparative analyses of data storytelling practices across cultural and disciplinary contexts, and across multiple media platforms (digital journalism, scientific publishing, social media, public exhibitions). Practically, it will develop a Data Storytelling Diversity Toolkit: a set of methodological guidelines, design heuristics, and computational tools (leveraging NLP) to support the creation of inclusive, ethical, and context-sensitive data narratives.

Possible Research Questions
- How do cultural and disciplinary contexts influence the form, ethics, and reception of data stories?
- How do different media platforms (e.g., news outlets, academic publishing, social media, public art) mediate the communication and engagement of data narratives?
- In what ways can NLP techniques (summarisation, sentiment analysis, translation, generative text) both support and constrain inclusive approaches to data storytelling?
- How can feminist and decolonial perspectives reveal power dynamics in whose stories are told, who tells them, and how they are received?
- What frameworks can evaluate the inclusivity, ethical responsibility, and audience impact of data stories?
- How can these insights be consolidated into a practical toolkit that guides researchers, journalists, and practitioners in designing equitable and culturally sensitive data storytelling practices?

Expected Contributions
- A cross-cultural and cross-disciplinary framework for analysing data storytelling.
- A critical integration of NLP, feminist, and decolonial approaches to data narrative design and evaluation.
- Case studies of data storytelling across multiple cultural and media settings.
- A Data Storytelling Diversity Toolkit, combining:
o methodological guidelines for comparative analysis,
o design heuristics informed by feminist and decolonial theory, and
o prototype NLP-based tools for inclusive and multilingual data narratives.
- Practical recommendations for creating ethical, inclusive, and context-sensitive data stories.

Understanding the Complexity of Negotiations

Supervisor: Alfie Abdul-Rahman

Areas: Human-centred computing (human-computer interaction), Natural Language Processing, Systems (software engineering, programming)

(Back to Scholarship Not Allocated)

Project Description

Negotiated texts are the outcome of formal decision-making processes in which groups of actors deliberate, draft, and refine documents over time. Many foundational texts of the modern world—including international treaties such as the Universal Declaration of Human Rights or the Treaty of Versailles, and constitutions such as that drafted by the American states in 1787—were not authored by individuals but produced through complex collective negotiation. Even relatively short documents of this kind often embody thousands of individual proposals, revisions, and decisions, reflecting a dense history of interactions among negotiating parties.

To understand such processes, it is essential to track the involvement of delegations and assess their influence on the development of negotiated texts. Traditional close reading can provide detailed insights, while distance reading and computational methods such as machine learning enable the analysis of patterns across large corpora. However, to fully capture the complexity of negotiations, these approaches need to be integrated with innovative visual analytics methods that make it possible to explore the interplay of events, actors, and evolving drafts at scale.

This project will apply a visual analytics approach to illuminate the complexity of negotiations and the influence of delegations throughout the drafting process. It will combine computational text analysis, natural language processing, and advanced visualization to bridge formal records (such as meeting minutes and draft revisions) with informal archival materials (e.g., private diaries, correspondence, or social media records in contemporary contexts).

Possible Research Questions
- How can new static and interactive visualizations be designed to support data discovery and insight generation in large datasets of negotiation events represented as interacting timelines?
- What novel approaches can reveal the evolution of complex, technical documents over extended periods of negotiation?
- How can indexing and retrieval methods for negotiation datasets be improved, and how can their results be presented in more intuitive and interpretable displays?
- In what ways can natural language processing be used to link information from informal archives (e.g., diaries, letters, or social media feeds) with the formal records of a negotiation, thereby enriching our understanding of the decision-making process?

Expected Contributions
- A visual analytics framework for analysing negotiation processes at multiple scales.
- New visualization techniques for representing evolving texts, timelines, and actor interactions.
- Computational tools for indexing, linking, and exploring negotiation-related datasets.
- Methodological insights into integrating close reading, distance reading, and machine learning with visual exploration.
- Case studies demonstrating how these methods enhance our understanding of historical and contemporary negotiated texts.

Exploring Interactive Multi-Dimensional Approaches to Patient Communication in Oral Health Education

Supervisor: Informatics: Dr Alfie Abdul-Rahman & Dr Lin Gui FoDOCS: Dr Melanie Nasseripour & Dr Ana Angelova

Areas: Artificial Intelligence, Machine learning / Deep learning, Human-centred computing (human-computer interaction), Natural Language Processing, Systems (software engineering, programming)

(Back to Scholarship Not Allocated)

Project Description

This joint project between the Department of Informatics and the Faculty of Dentistry, Oral & Craniofacial Sciences (FoDOCS) seeks to transform how patient communication is taught in oral health education. Effective patient-clinician communication is central to encouraging oral hygiene practices, supporting behaviour change, and demonstrating professionalism. However, conventional training methods are resource-intensive, requiring significant time, staff, and cost to deliver authentic clinical practice scenarios.

To address these challenges, the project proposes the design and evaluation of interactive, multi-dimensional learning environments, with a particular emphasis on immersive Virtual Reality (VR). VR offers an unparalleled ability to create engaging, realistic, and repeatable patient encounters where students can practice communication skills in a safe but lifelike clinical setting. Complementary approaches—including text-to-text and voice-to-voice communication systems—will expand accessibility, providing adaptable tools for different learners and educational contexts.

Generative Language Models (GLMs) will be harnessed to generate customised patient cases and conversational responses, ensuring that scenarios adapt dynamically to student inputs. These responsive simulations allow students to repeatedly practice communication tasks such as explaining oral hygiene, negotiating dietary changes, or discussing treatment options, thereby reinforcing learning and professional skill development.

Research Questions
- How can immersive VR be designed to simulate realistic patient-clinician interactions that enhance students' communication, empathy, and professionalism?
- What are the comparative benefits and limitations of VR, text-based, and voice-based simulation approaches in oral health education?
- In what ways can GLMs be integrated to generate personalised, contextually relevant, and adaptive patient scenarios that respond in real time to learners' actions?
- How do students perceive and engage with VR-based versus non-immersive approaches, and how does this affect learning outcomes?
- What metrics and evaluation frameworks can be developed to measure improvements in communication competence, confidence, and behaviour change skills using immersive and multi-dimensional tools?
- How can these approaches be scaled sustainably across curricula and adapted to different cultural and institutional contexts in oral health education?

Expected Contributions
- Development of a VR-based immersive training system for patient communication in oral health education.
- Comparative evaluation of VR, text, and voice-based simulations, identifying best practices for different learning settings.
- Integration of GLMs for adaptive, customised case study generation.
- A multi-dimensional toolkit for cost-effective, scalable communication training.
- Evidence-based recommendations for embedding immersive and interactive communication training within oral health curricula.

Computational Argumentation for Interactive Explainable AI

Supervisor: Antonio Rago

Areas: Artificial Intelligence, Machine learning / Deep learning, Human-centred computing (human-computer interaction), Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Context:
Computational argumentation has been shown to be an effective formalism for extracting the relevant information from AI models and providing faithful, transparent and interactive explanations to users in a variety of formats, such as those which are conversational [1]. Argumentative explanations have been shown to be useful when applied different AI models, including large language models [2], classification methods [3,4] and recommender systems [1,5]. Recently, a general framework based on argumentation has been introduced [6] that is claimed to be able to not only model interactive explanations provided by AI models, but also the reasoning processes and cognitive biases in humans. However, these claims have currently been verified only by theoretical analysis and evaluations based on simulations, without the user studies which would provide conclusive evidence of the approach's capabilities.

Aims and Methodology:
The aim of this project is to investigate the ability of argumentative approaches, perhaps inspired by [6], to model human reasoning in interactive explanations for a range of AI models in different applications. The student will define instances of argumentative technologies as methods for explaining AI models, e.g. large language models or image classifiers. The methods will then be implemented, giving novel tools for interactive explanations which shifts the decision making towards the user, as in novel Evaluative AI paradigm [7]. The method will then be evaluated by means of machine-centric properties, such as faithfulness, using publicly available datasets in domains such as engineering and healthcare. The project will then explore the evaluation by user-centric properties of explainable AI, such as user satisfaction or trust, as in [1]. The student will then undertake user studies, which may be crowdsourced via Prolific (https://www.prolific.com). The study of interactive explanation for AI models is fertile ground, not least with the rise of large language models, and we believe that this project could have significant impact on the research landscape, particularly given the cutting-edge proposals of [7].

Requirements:
An interest (or, preferably, experience in) in the theory of computational logic and argumentation will be essential, as well as the implementation and analysis of machine learning models and explainable AI methods. This work will follow much of the argumentation-related work shown here: https://antoniorago.github.io/publications/

References

References:
[1] Argumentative Explanations for Interactive Recommendations; Rago et al.; AIJ 2021. https://www.sciencedirect.com/science/article/pii/S0004370221000576
[2] Argumentative Large Language Models for Explainable and Contestable Claim Verification; Freedman et al.; AAAI 2025. https://doi.org/10.1609/aaai.v39i14.33637
[3] Data-Empowered Argumentation for Dialectically Explainable Predictions; Cocarascu et al.; ECAI 2020. http://ebooks.iospress.nl/publication/55172
[4] A Little of That Human Touch: Achieving Human-Centric Explainable AI via Argumentation; Rago; IJCAI 2024. https://www.ijcai.org/proceedings/2024/983
[5] Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them; Rago et al.; IJCAI 2018. https://www.ijcai.org/proceedings/2018/269
[6] Interactive Explanations by Conflict Resolution via Argumentative Exchanges; Rago et al.; KR 2023. https://proceedings.kr.org/2023/57/
[7] Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI; Miller; FAccT 2023. https://dl.acm.org/doi/10.1145/3593013.3594001

Verbal reasoning with Bayesian networks

Supervisor: Jeroen Keppens

Areas: Artificial Intelligence, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Bayesian networks (BNs) are probabilistic graphical models that allow one to represent uncertain knowledge in domains where variables are interdependent [4]. They are often employed as decision support tools because they combine domain knowledge (often from experts) with probabilistic (statistical) information to compute beliefs (posterior probabilities) about hypotheses or outcomes given observed evidence. This makes them well suited to reasoning under uncertainty: e.g. in medical diagnosis (where symptoms are evidence, diseases are hypotheses), risk assessment, fault diagnosis, and ecological management. They enable not only predictive inference (from causes to effects) but also diagnostic inference (from observed effects back to likely causes), and inter-causal reasoning (how evidence bearing on one cause affects the probabilities of other competing causes).

In decision support, BNs help decision makers by quantifying uncertainty, enabling "what if" analyses (e.g. what if certain evidence is observed or what if certain interventions are made), and computing posterior probabilities that feed into utility or cost-benefit analyses or influence diagrams (an extension of BNs to include decision and utility nodes) to choose among actions.

Application of BNs to decision support problems is not trivial. Construction and validation of BNs typically requires expert knowledge of the modelling technique as well as expert domain knowledge. Moreover, even though BNs compute the rational implications as specified by the models, the results can be counterintuitive to human decision makers due to cognitive biases. A range of explanation techniques have been proposed, but these too require a solid background in BN models [1,3].

This project aims to develop hybrid symbolic and sub-symbolic techniques to produce explainable decision support information using BNs via a natural language interface. To achieve this, existing techniques to generate BN models from first principles [2] will be adapted and extended to enable the construction of BNs from first principles from natural language input via an LLM. BN explanation techniques will be adapted and extended to generate explanation via an LLM.

References

[1] Derks, I. P., & de Waal, A. (2021). A Taxonomy of Explainable Bayesian Networks. arXiv preprint arXiv:2101.1184.
[2] Keppens, J., Shen, Q., & Price, C. (2011). Compositional Bayesian modelling for computation of evidence collection strategies. Applied Intelligence, 35(1), 134-161.
[3] Lacave, C., & Diez, F. J. (2002). A Review of Explanation Methods for Bayesian Networks. The Knowledge Engineering Review, 17(2), 107-127.
[4] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann.

Conflict Detection and Reconciliation Across Diverse Clinical Knowledge Sources in Pregnancy Care

Supervisor: Yulan He / Runcong Zhao

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Pregnancy care draws on diverse guidelines, hospital protocols, and patient information, but this wealth of evidence also creates challenges for women seeking clear, reliable answers. Guidance can be fragmented, inconsistent, or even contradictory, leaving patients uncertain and clinicians struggling to provide consistent advice. At the same time, large language models (LLMs) and conversational AI tools are becoming increasingly common in digital health, yet current systems largely provide generic responses and fail to address the complexities of conflicting information.

Existing automated approaches to guideline conflict detection are limited, often relying on model-level "black box" decisions. Such approaches suffer from several issues: they may favour whichever recommendation appears most frequently, or give undue weight to options mentioned at the beginning or end of a document. Beyond accuracy, this kind of opaque decision-making undermines patient trust, as it provides no clear rationale for why one recommendation is preferred over another.

Our project is therefore novel in advancing fundamental methods for: formally representing heterogeneous guideline knowledge; algorithmically detecting inconsistencies; evaluating and weighting conflicting evidence; and designing explainable reconciliation strategies suitable for patient use. By shifting the focus from simple question answering towards conflict-aware, explainable reasoning, this project addresses a key technical and clinical gap.

The overarching aim is to develop a transparent and trustworthy QA framework that can provide reliable answers even when knowledge sources disagree. Specifically, the project will (1) create formal representation models for pregnancy-related guidance, (2) design algorithms to detect and categorise conflicts, (3) develop evidence weighting and reconciliation strategies, and (4) integrate explainable mechanisms into a patient-facing QA system. Pregnancy care is chosen as the initial testbed given its direct impact on maternal and fetal outcomes, but the proposed methods are domain-agnostic and can be applied across healthcare more broadly.

References

Braun, T., Rothermel, M., Rohrbach, M., and Rohrbach, A. (2025). Defame: Dynamic evidence-based fact-checking with multimodal experts. In: International Conference on Machine Learning (ICML).
Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. (2024). Medagents: Large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics (ACL).
Venktesh, V. and Setty, V. (2025). Factir: A real-world zero-shot open-domain retrieval benchmark for fact-checking. In: Proceedings of the ACM Web Conference 2025 (WWW). Resource Track.
Wu, H., Zeng, Q., and Ding, K. (2024). Fact or facsimile? evaluating the factual robustness of modern retrievers. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM).

Bayesian Meta-Reasoning in Large Language Models

Supervisor: Yulan He / Hanqi Yan

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

State-of-the-art large language models (LLMs) are typically pre-trained through statistical next-token prediction and often post-trained with reinforcement learning (RL) using reward models. While these methods enable strong performance on reasoning tasks such as mathematics and coding, they face key limitations such as: (1) sensitivity to task frequency and input perturbations; (2) dependence on task-specific annotations and preference data, which limits scalability; (3) unreliable self-generated feedback that fails to capture shared reasoning patterns. These constraints arise because LLMs are optimised to solve tasks directly rather than to learn how to reason. For better generalisation, we argue that models need to instead actively engage in learning-to-reason or meta-reasoning processes, and develop the ability to monitor, regulate, and adapt their reasoning strategies when needed.

This project aims to implement the Bayesian meta-reasoning framework we described in our position paper in ICML 2025. The new framework will enable models to move beyond task-specific optimisation towards adaptive reasoning strategies that generalise across domains. The research will first develop unified benchmarks and metrics to evaluate meta-reasoning across domains, moving beyond accuracy to measures of calibration error, logical consistency, and cross-domain generalisation. It will then design adaptive architectures that dynamically combine reasoning skills, using approaches such as Mixture-of-Experts and Bayesian inverse planning. Self-play will be explored to discover scalable and multifaceted rewards without costly human annotations. Finally, latent-space reasoning methods will be explored to reduce cascading errors and improve efficiency. Techniques such as diffusion-based reasoning and looped transformers will be explored. If time permits, the project may extend to interpretable meta-knowledge consolidation, using mechanistic interpretability to link reasoning skills to model components. This would allow selective fine-tuning and more efficient training.

References

Yan, H., Zhang, L., Li, J., Shen, Z. and He, Y., 2025. Position: LLMs Need a Bayesian Meta-Reasoning Framework for More Robust and Generalizable Reasoning. In 2025 International Conference on Machine Learning (ICML).

Towards controllable and interpretable large language model alignment

Supervisor: Hanqi Yan/Yulan He

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Large language models (LLMs) are rapidly transforming how we interact with technology, but making them aligned with human values and needs remains a major open challenge. Current post-training methods (e.g., reinforcement learning from human feedback) can make models more helpful, accurate, or supportive. However, they face important limitations: (i) Models often need to balance multiple objectives at once (e.g., being safe, empathetic, and logically sound). (ii) Alignment needs can change depending on the person and context. (iii) Models risk forgetting prior abilities when retrained for new goals.

This PhD project will explore how to make LLM alignment both controllable (so we can flexibly adjust a model's behavior) and interpretable (so we can understand why the model behaves as it does). The research will have two main strands: (i) Understanding Alignment Mechanisms: use mechanistic interpretability tools (e.g., sparse autoencoders, probing, causal interventions) to study how alignment signals are represented or entangled inside LLMs (ii) Building Controllable Alignment Frameworks. It is to develop adaptive methods that integrate multiple alignment goals. Explore inference-time control, such as representation steering to adjust model behavior dynamically in response to user queries. The framework will be applied to real-world challenges such as mental health chatbots, where LLMs must combine safety, reasoning, empathy, and clear communication. The broader impact will be methods for more transparent, reliable, and adaptable LLMs. This project is ideal for students interested in LLMs, interpretability, and alignment, and offers excellent opportunities for impactful research and publication.

References

[1] Thinking Hard, Going Misaligned: Emergent Misalignment in LLMs
[2] Sparse Activation Editing for Reliable Instruction Following in Narratives
[3] Soft Reasoning: Navigating Solution Spaces in Large Language Models through Controlled Embedding Exploration
[4] Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference
[5] NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning

Adaptive and Inclusive Cybersecurity Education

Supervisor: Tasmina Islam

Areas: Artificial Intelligence, Machine learning / Deep learning, Cybersecurity, Human-centred computing (human-computer interaction), Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Cybersecurity awareness is essential in a digital society, yet current initiatives are often generic and fail to address the needs of underrepresented groups such as children, minority communities, and people with disabilities. This project proposes to design adaptive awareness environments that apply artificial intelligence to deliver personalised, culturally relevant, and accessible cybersecurity training.

Participants will engage with interactive simulations of cyber threats (e.g., phishing, ransomware, data leakage) enhanced by generative AI, which can create customised attack narratives and provide context-aware, real-time feedback. Alongside the development of adaptive systems, the project will employ interviews and focus groups, combined with behavioural analytics and surveys, to understand user needs, evaluate effectiveness, and identify different types of cybersecurity users (e.g., risk-takers, cautious users, or overconfident participants). It will also explore new ways of measuring awareness outcomes, introducing resilience-based metrics such as detection speed, recovery performance, and transfer of skills across scenarios.

By combining technical innovation with human-centred research, this work seeks to establish a scalable framework for evidence-driven cybersecurity awareness, empowering diverse populations to navigate digital risks safely. Prospective students can discuss options with the supervisor.

References

1. Segupta, S., Varma, U., & Islam, T. (2025). Empowering Cybersecurity Education: A Review of Adaptive Learning Paradigms and Practical Implications. In Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, 16-18 June 2025, Munich, Germany.: 3rd International Workshop on Cyber Security Education for Industry and Academia.
2. Hedges, M., & Islam, T. (2024). VirSec - Immersive Security Training within Virtual Reality. In 17th International Conference on Advanced Visual Interfaces: 2nd International Workshop on CyberSecurity Education for Industry and Academia (CSE4IA 2024)
3. Islam, T & Zou, Y 2023, ChildSecurity: A Web-based Game to Raise Awareness of Cybersecurity and Privacy in Children. in Cybersecurity Challenges in the Age of AI,Space Communications and Cyborgs.

Lifelong and Culturally-Aware Model Editing: Towards Scalable, Context-Sensitive Knowledge Updates in Multilingual Language Models

Supervisor: Helen Yannakoudakis

Areas: Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Large language models are increasingly deployed in dynamic, multilingual settings where facts, APIs, and public knowledge shift frequently. Existing model editing techniques such as ROME, MEMIT, and MEND have shown success in making localised factual corrections, but they typically assume that edits are isolated, language-neutral, and structurally simple. These assumptions break down in real-world applications, where edits must accumulate over time and reflect differing cultural or linguistic realities. In practice, a factual update that is accurate and appropriate in English may require significant adaptation, or even a different framing entirely, when expressed in Arabic, Hindi, or Swahili. Such differences are not mere translation artifacts but stem from deeper issues of cultural context and knowledge representation.

This PhD project proposes a new framework for lifelong and culturally-aware model editing that supports both the scalability of long-horizon updates and the sensitivity required for cross-lingual propagation. At its core is a dual-phase editing system: rapid, lightweight interventions are used for immediate local updates, while slower consolidation phases ensure long-term stability through replay-based regularisation and continual learning techniques. The project is grounded in two central challenges: first, how to ensure that models remain stable and composable under long sequences of factual edits, without suffering interference, drift, or catastrophic forgetting; and second, how to propagate a single factual update across multiple languages in a way that preserves semantic intent while allowing for appropriate cultural and contextual variation, particularly in low-resource or typologically distant languages.

To address these challenges, the framework introduces editing in a shared causal latent space designed to align internal representations across languages, enabling coherent generalisation without assuming uniformity. At the same time, it permits language-specific or culture-specific adaptation when edits carry different implications across linguistic contexts. By integrating insights from model editing, multilingual NLP, and lifelong learning, this work aims to advance the reliability, adaptability, and cultural robustness of foundation models operating in a global context.

References

[1] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? EMNLP, 2023.
[2] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 1(1):12, 2021.
[3] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In International conference on machine learning. PMLR, 2020.
[4] Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark. NeurIPS, 2024.
[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521—3526, 2017.
[6] Jiaang Li, Quan Wang, Zhongnan Wang, Yongdong Zhang, and Zhendong Mao. Enhance lifelong model editing with continuous data-adapter association. arXiv e-prints, pages arXiv—2408, 2024.
[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935—2947, 2017.
[8] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. NeurIPS, 2017.
[9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. Advances in neural information processing systems, 35:17359—17372, 2022.
[10] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. ICLR, 2023.
[11] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. ICLR.
[12] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. ICLR. PMLR, 2022.
[13] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. ICLR.
[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. NeurIPS, 2022.
[15] Jonas Pfeiffer, Aishwarya Kamath, Andreas Ruckle, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. EACL, 2021.
[16] Jonas Pfeiffer, Ivan Vuli´c, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. EMNLP, 2020.
[17] Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. Cross-lingual knowledge editing in large language models. ACL, 2024.
[18] Mengqi Zhang, Xiaotian Ye, Qiang Liu, Shu Wu, Pengjie Ren, and Zhumin Chen. Uncovering overfitting in large language model editing. ICLR, 2025.
[19] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. NeurIPS.

Language-aware alignment and cross-lingual safeguards in LLMs

Supervisor: Oana Cocarascu

Areas: Natural Language Processing, Machine learning / Deep learning

(Back to Scholarship Not Allocated)

Project Description

Large language models (LLMs) have shown disparities in their responses within and across languages. Adversaries can exploit inconsistent behaviour in low-resource languages to bypass safety measures, spread misinformation or generate harmful content. With current evaluations and benchmarks primarily focusing on English, there is a need for language-aware alignment and cross-lingual safeguards in LLMs.

This PhD project aims to conduct a systematic evaluation of the behaviour of LLMs in multilingual settings and use interpretability techniques to examine the mechanisms behind unsafe generation to improve safety across languages. Concretely, it will focus on low-resource languages, dialectal variations within high-resource languages, and code-mixed languages. Through extensive evaluations across multiple languages and variations thereof, the project will provide a better understanding of model behaviour on safety-sensitive inputs across languages. To uncover the mechanisms leading to unsafe generation in LLMs across diverse languages, we will identify latent regions tied to safety as these can vary between languages or dialects, and perturb components to determine latent risks across diverse linguistic inputs. By analysing LLM responses to prompts in multiple languages, this work will uncover language-specific risks and will provide novel insights into the interaction between language and unsafe behaviour.