PhD projects (Natural Language Processing)

- Semantic Visual Analytics for Exploring the Evolution of Folklore across Cultures and Media
- Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge
- Learning for Healthy Human-AI Co-Evolution
- Explainable AI for Biomedical Discovery
- Evolving Cybercrime: Improving Threat Recognition and Response
- Verbal reasoning with Bayesian networks
- Scalable Oversight of AI Agents
- Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement
- Cross-cultural argument evaluation and generation
- Language-aware alignment and cross-lingual safeguards in LLMs
- Towards controllable and interpretable large language model alignment

Semantic Visual Analytics for Exploring the Evolution of Folklore across Cultures and Media

Supervisor: Alfie Abdul-Rahman

Project Description

This PhD project explores how semantic modelling, artificial intelligence, and interactive visualization can support the human-centred exploration of complex textual and cultural information. Using folklore across cultures and media as its research context, the project will develop semantic visual analytics methods that enable users to explore, question, and critically assess computational representations of narrative knowledge.

Singapore and Malaysia provide the primary context, offering interconnected histories, multilingual environments, and diverse cultural traditions. Case studies may examine figures such as the pontianak, orang bunian, toyol, Sang Kancil, and Hang Tuah, tracing their evolution across historical texts, literature, film, television, comics, games, and digital culture.

The project will investigate how LLMs, semantic embeddings, information extraction, and knowledge graphs can be combined with interactive visualization to explore narrative relationships, timelines, semantic similarities, and adaptation pathways. Particular attention will be given to provenance, uncertainty, ambiguity, multilingual representation, and competing interpretations of AI-derived cultural knowledge.

Using a human-centred design-study methodology, the research will work with domain experts to design and evaluate new visualization approaches. Expected contributions include novel semantic visualizations, methods for communicating provenance and uncertainty, approaches to human-AI visual sensemaking, and an open-source visual analytics platform.

References

El-Assady, M., Kehlbeck, R., Collins, C., Keim, D. A., & Deussen, O. (2020). Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1001—1011.
Janicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis STARs.
Kahng, M., et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. CHI EA '24.
Liem, J., et al. (2023). A Workflow Approach to Visualization-Based Storytelling with Cultural Heritage Data. IEEE VIS Workshop on Visualization for the Digital Humanities.
Liu, S., et al. (2019). Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE Transactions on Visualization and Computer Graphics, 25(7), 2482—2504.
Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., & Elmqvist, N. (2018). ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. IEEE Transactions on Visualization and Computer Graphics, 24(1), 361—370.

Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge

Supervisor: Alfie Abdul-Rahman

Project Description

This PhD project investigates how interactive visualization and visual analytics can enable people to explore, interrogate, validate, and refine knowledge generated by Large Language Models (LLMs). Rather than treating AI-generated information as definitive, the project will develop methods that help users understand how computational interpretations are constructed, examine supporting evidence, and identify uncertainty or disagreement.

The research will investigate how LLMs, semantic embeddings, information extraction, retrieval-based methods, and knowledge graphs can be integrated with interactive visualization. Potential approaches include knowledge graphs, semantic landscapes, evidence views, document visualizations, and model-comparison interfaces that enable users to reason with, question, and refine machine-generated knowledge.

A central focus will be semantic provenance: understanding not only where information originates, but how AI-generated interpretations are derived from underlying evidence. The project will explore visual approaches for communicating provenance, uncertainty, contradictory evidence, and alternative interpretations, while enabling users to validate, modify, or reject AI-generated representations.

Using a human-centred visual analytics methodology, the research will combine iterative design with qualitative and quantitative evaluation across domains such as scientific literature, cultural collections, and policy documents. Expected contributions include novel visualization techniques, methods for communicating semantic provenance and uncertainty, design guidance for human-AI visual analytics, and an open-source platform.

References

El-Assady, M., Kehlbeck, R., Collins, C., Keim, D. A., & Deussen, O. (2020). Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1001—1011.
Kahng, M., et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. CHI EA '24.
Liu, S., et al. (2019). Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE Transactions on Visualization and Computer Graphics, 25(7), 2482—2504.
Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., & Elmqvist, N. (2018). ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. IEEE Transactions on Visualization and Computer Graphics, 24(1), 361—370.
Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2016). The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics, 22(1), 240—249.

Learning for Healthy Human-AI Co-Evolution

Supervisor: Yulan He

Project Description

Personalised AI systems are typically designed to adapt to users' preferences and provide responses they are likely to find helpful or satisfying. But the relationship between human and AI is two-way: while human feedback can improve AI systems, interacting with AI can also change what people know, how they reason, and what they are able to do. Most existing alignment methods, such as reinforcement learning from human feedback, mainly optimise AI responses for immediate user preferences. This can encourage undesirable behaviours such as sycophancy, where an AI agrees with users or reinforces their existing views rather than helping them learn or make better decisions. We therefore view human-AI interaction as a process of co-evolution. Healthy co-evolution should benefit both sides. The AI should become more capable, and the human should also gain better cognitive capabilities. This project will develop a new learning approach that rewards AI interaction strategies according to whether they lead to measurable improvements in human capability over time, rather than simply maximising immediate satisfaction. It will: (i) develop a reward signal that captures capability growth, (ii) design an algorithm that can identify which earlier AI actions contributed to that growth, and (iii) evaluate the approach in controlled A/B studies against systems optimised mainly for user satisfaction or stated preferences.

References

1. Liu, Wei, Siya Qi, Yali Du, and Yulan He. "Self-play only evolves when self-synthetic pipeline ensures learnable information gain." ICML 2026.
2. Sharma, Mrinank, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus et al. "Towards understanding sycophancy in language models." ICLR 2024.
3. Cheng, Myra, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. "Sycophantic AI decreases prosocial intentions and promotes dependence." Science 391, no. 6792 (2026): eaec8352.
4. Ding, Qianggang, Xingyao Wang, Rui Feng, Zhibin Wang, Feixiang Wang, Kelong Mao, Hao Sun et al. "ComBodied Agents: a New Paradigm of Human-Centric Agentic AI." arXiv preprint arXiv:2608.10915 (2026).

Explainable AI for Biomedical Discovery

Supervisor: Dr Sophia Tsoka and Prof Yulan He

Project Description

AI systems can increasingly combine structured information and scientific literature to support scientific reasoning and discovery [1], but it remains difficult to understand how reliable their reasoning is and what evidence supports their conclusions [2]. This PhD project will develop new AI methods for explainable scientific discovery, with biomedical data providing a rich and challenging application domain.

The project will explore how foundation models, including large language models, can be combined with multimodal learning, structured reasoning, knowledge representation and optimisation-based methods in biomedicine [1—6]. Research questions will include how to represent information across molecular data and scientific literature, how to generate hypotheses that are traceable to supporting evidence and how to design AI systems that can express alternatives and uncertainty rather than producing a single opaque answer. A key part of the PhD will be developing methods to test whether AI-generated hypotheses remain consistent across independent datasets and data types and are supported by existing scientific knowledge. Applications will include single-cell and spatial multi-omics, disease-mechanism discovery, medical prognosis and therapeutic hypothesis generation. The project is particularly suited to students interested in both developing novel AI methodology and applying it to important real-world scientific problems.

References

1. da Silva, I. L., Yan, H., Gui, L., & He, Y. (2026). GraphMind: Unveiling Scientific Reasoning through Contextual Graphs for Novelty Assessment. Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, 732—742. https://doi.org/10.1145/3770855.3818195
2. Li, J., Yan, H., & He, Y. (2025). Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6850—6866. https://doi.org/10.18653/v1/2025.acl-long.340
3. Liapis, G. I., Tsoka, S., & Papageorgiou, L. G. (2025). Interpretable optimisation-based approach for hyper-box classification. Machine Learning, 114(3). https://doi.org/10.1007/s10994-024-06643-7
4. Li, Y., Cardoso-Silva, J., Kelly, J. M., Delves, M. J., Furnham, N., Papageorgiou, L. G., & Tsoka, S. (2024). Optimisation-based modelling for explainable lead discovery in malaria. Artificial Intelligence in Medicine, 147, 102700. https://doi.org/10.1016/j.artmed.2023.102700
5. Li, Y., da Costa Avelar, P. H., Chen, X., Zhang, L., Wu, M., & Tsoka, S. (2024). CLigOpt: controllable ligand design through target-specific optimization. Bioinformatics, 40(Supplement_2), ii62—ii69. https://doi.org/10.1093/bioinformatics/btae396
6. da Costa Avelar, P. H., Cardoso-Silva, J., Wu, M., & Tsoka, S. (2026). SCONE: a subset-contrastive method for multi-omics network embedding. Briefings in Bioinformatics, 27(4). https://doi.org/10.1093/bib/bbag381

Evolving Cybercrime: Improving Threat Recognition and Response

Supervisor: Tasmina Islam / Kovila Coopamootoo

Project Description

Cybercrime continually evolves as offenders adapt their language, tactics and modes of interaction to new technologies and user behaviours. The growing use of generative AI may accelerate this process, making emerging threats harder both to detect computationally and to recognise by people.

This project will investigate how changes in malicious communications, such as phishing, scams and social engineering, influence threat recognition and response. Computational methods including natural language processing, machine learning, metadata analysis and temporal modelling will be used to identify how malicious narratives and behavioural patterns evolve over time. Human-centred studies will investigate how people recognise and interpret changing threats, the cues they rely on, common misconceptions, and factors that influence their security decisions.

The project will also investigate how people can be better prepared to recognise unfamiliar and changing threats. It will explore how computational findings about emerging threat patterns can inform cybersecurity learning and develop and evaluate approaches that support transferable threat-recognition and decision-making skills rather than reliance on known examples or fixed warning signs.

References

1. Islam, T., Becker, I., Posner, R., Ekblom, P., McGuire, M., Borrion, H., & Li, S. (2019). A socio-technical and coevolutionary framework for reducing human-related risks in cyber security and cybercrime ecosystems. In International Conference on Dependability in Sensor, Cloud, and Big Data Systems and
Applications (pp. 277-293).
2. Ng M, Bada M, Coopamootoo KPL (2023) . What We Do in the Shadows: How does Experiencing Cybercrime Affect Response Actions & Protective Practices? . Conference: 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) vol. 00 , 659 - 672 .
3. J. Li and T. Islam, "Learncyber - Addressing Security and Privacy Concerns: Personalised Learning Tools," 2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Yogyakarta, Indonesia, 2025, pp. 467-472,

Verbal reasoning with Bayesian networks

Supervisor: Jeroen Keppens

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.

Scalable Oversight of AI Agents

Supervisor: Yali Du

Project Description

This PhD project will investigate how less capable AI systems can reliably supervise more capable language model agents. As agents undertake increasingly complex tasks, their outputs and explanations may become difficult to verify, creating a need for oversight methods that remain effective across capability gaps.

The research will develop an empirical and methodological framework for understanding when oversight succeeds, why it fails, and how limited supervision resources should be allocated. It will examine three connected questions: how capability differences affect error detection; whether evidence, verification tools, and structured debate improve supervisory judgements; and how selective escalation to stronger overseers can balance reliability and cost. Particular attention will be paid to plausible but misleading explanations and failures that emerge when agents adapt to oversight.

Using tasks with independently verifiable outcomes, the project will combine controlled experiments, modelling of oversight performance, and the development of adaptive supervision methods. Evaluation will measure accuracy, robustness, computational cost, and generalisation across tasks and models.

The expected contribution is a principled understanding of the limits of scalable oversight, alongside practical methods for improving the reliability and efficiency of supervising increasingly capable AI agents.

References

Engels et al. Scaling Laws For Scalable Oversight. https://arxiv.org/abs/2504.18530
Bowman et al. Measuring Progress on Scalable Oversight for Large Language Models. Paper
Bengio. Why are AI agents lying, cheating and coordinating? https://arxiv.org/abs/2211.03540
https://yoshuabengio.org/en/blog/why-are-ai-agents-lying-cheating-and-coordinating

Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement

Supervisor: Yali Du

Project Description

This PhD project will investigate how systems of interacting large language model (LLM) agents can improve their capabilities through collaborative training and iterative self-improvement. Agents may assume complementary roles as task generators, solvers, critics, and evaluators, producing learning signals that support improvements across successive training cycles.

The research will examine when multi-agent interaction produces reliable capability gains beyond individual self-improvement. Key questions include how agent diversity, communication, and training objectives influence learning; how agents can generate increasingly useful tasks and feedback; and whether improvements to task generation and evaluation enable sustained recursive self-improvement (RSI). The project will distinguish improvements arising from additional inference computation, memory, or workflow adaptation from those achieved through model training.

Methods will combine multi-agent reinforcement learning, synthetic data generation, and adaptive curricula. Controlled experiments will compare collaborative and individual training under matched resource budgets, using held-out tasks and independent evaluation to assess generalisation, efficiency, and stability. Particular attention will be paid to error amplification, reward exploitation, and loss of diversity.

The expected contribution is a framework for understanding and developing self-evolving agent systems, identifying both the mechanisms that enable sustained improvement and the conditions that limit it.

References

1. *MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning.* (2025). [arXiv:2502.18439](https://arxiv.org/abs/2502.18439).
2. *MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models.* (2024). [arXiv:2402.01620](https://arxiv.org/abs/2402.01620).
3. Chen, Z., et al. (2024). *Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models.* [arXiv:2401.01335] (https://arxiv.org/abs/2401.01335).
4. Zhao, A., et al. (2025). *Absolute Zero: Reinforced Self-play Reasoning with Zero Data.* [arXiv:2505.03335] (https://arxiv.org/abs/2505.03335).
5. Zhang, J., et al. (2025). *Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents.* [arXiv:2505.22954] (https://arxiv.org/abs/2505.22954).

Cross-cultural argument evaluation and generation

Supervisor: Oana Cocarascu

Project Description

Argumentation (i.e. the way individuals formulate arguments and evaluate them) varies substantially across cultures. Argument quality, how evidence is presented, how politeness and emotions are expressed, how hierarchy is signalled, and how persuasive strategies are used in natural language vary significantly across cultures. However, large language models (LLMs) have been found to fail to generate and evaluate arguments in ways that align with cultural norms, and multi-cultural value awareness of LLMs remains an active research topic.

The project will investigate how LLMs interpret and generate arguments across cultural contexts while preserving argument quality. In particular, the project will conduct a comprehensive analysis of how argument quality and persuasiveness differ across cultures, and how LLMs can be evaluated, aligned, and controlled to respect culturally diverse argumentative norms. The project aims to conduct a systematic evaluation of how cultural value systems shape quality and persuasiveness of arguments generated by LLMs, and develop models that can generate culturally aligned arguments while maintaining their validity. The research will evaluate the behaviour of LLMs in moral dilemmas across languages and cultures and develop models that evaluate argument quality, coherence, factual grounding, emotional appeal, and persuasiveness, while accounting for cultural context.

Language-aware alignment and cross-lingual safeguards in LLMs

Supervisor: Oana Cocarascu

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. Previous works on safety alignment have been predominantly done in monolingual settings. More recent works have proposed multilingual safety models for safeguarding LLM generations, albeit on high-resource languages, prioritising breadth over language depth. There is a need for language-aware alignment and cross-lingual safeguards in LLMs that goes beyond human and machine translated samples to incorporate linguistic patterns such as dialectical variations within language.

This project will evaluate the behaviour of LLMs in a variety of (low-resource) languages and dialects, as well as mixed-language data to ensure AI systems behave as intended. The project aims to complete three key objectives: O1) conduct a systematic evaluation of the behaviour of LLMs in multilingual settings, O2) use interpretability techniques to examine the mechanisms behind unsafe generation, and O3) develop a framework to enhance safety alignment for low-resource languages. Concretely the research will focus on i) low-resource languages, ii) dialectal variations within high-resource languages, and iii) code-mixed languages. As models can bypass safety guardrails when prompted in languages underrepresented in pretraining, the project will evaluate how susceptible LLMs are to jailbreak attacks across low-resource languages, dialects of high-resource languages and how code-mixed prompts affect safety mechanisms. 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, the project 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. In particular, the project will investigate the model's internal activations to localise latent components that correlate with jailbreaks through parameter scaling and perturbation, as well as activation patching. The project will provide novel insights into the interaction between language and unsafe behaviour and will develop a framework for enhancing safety alignment across languages and dialects.

Towards controllable and interpretable large language model alignment

Supervisor: Hanqi Yan/Yulan He

Project Description

Large language models (LLMs) are increasingly deployed in high-stakes settings, yet current alignment methods, most notably reinforcement learning from human feedback (RLHF), introduce a critical vulnerability: reward hacking. Rather than genuinely learning to be helpful, models may exploit the reward signal to produce outputs that satisfy evaluators without reflecting true knowledge or intent. This manifests in several failure modes including sycophancy, deceptive alignment, and faithfulness failures in chain-of-thought reasoning.

This PhD project pursues two research strands. First, using mechanistic interpretability tools including probing, causal interventions, and gradient-based analysis to identify where and how deceptive tendencies arise inside LLMs. Second, developing practical detection and mitigation methods including lightweight probing-based detectors and inference-time interventions that correct deceptive behaviour without retraining. Methods will be evaluated across high-stakes domains including education, healthcare, and legal assistance, contributing to safer and more trustworthy LLM deployment. This project is ideal for students interested in LLMs, interpretability, and alignment, and offers excellent opportunities for impactful research and publication.

References

[1] When thinking backfires: mechanistic insights into reasoning-induced misalignment
[2] Large Language Models Hack Rewards, and Society
[3] Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference
[4] GRADE: Probing Knowledge Gaps in LLMs through Gradient Subspace Dynamics
[5] PreAct-Bench: Benchmarking Predictive Monitoring in LLMs
[6] Automonitor-bench: Evaluating the reliability of llm-based misbehavior monitor