Research Statement
Ching Nam Hang

My research interests lie at the intersection of data science and artificial intelligence (AI), with an emphasis on their implementation in healthcare and education. Specific research projects include the theoretical exploration and practical applications of machine learning for pandemic response, the management of misinformation during crises, and the augmentation of self-learning in large-scale educational systems. Recently, I have been focusing on the applications and innovations of large language models (LLMs) to enhance health informatics and educational learning. In particular, I am the Principal Investigator of a RGC Faculty Development Scheme (FDS) project titled "*Harnessing Large Language Models for Social Trust: Automated Fact-checking for Public Health Informatics*", funded under the Competitive Research Funding Schemes for the Local Self-financing Degree Sector. My work is devoted to turning data-driven insights into practical, real-world applications, notably improving our response to global health crises and nurturing autonomous learning. My professional journey and continued engagement in data science and AI deepens my expertise and problem-solving ability, providing a strong foundation for developing innovative solutions with broad and meaningful impact.

Ph.D. Research

(1) Machine Learning for Pandemic Response

In the face of the COVID-19 pandemic, an array of digital contact tracing (DCT) strategies has been proposed to curb virus transmission. Existing literature, however, largely focuses on DCT application design and deployment, with less emphasis on machine learning optimization. Our work in [1], which received the Best Paper Award, presents a comprehensive review of machine learning-enhanced DCT strategies, formulating relevant research questions and outlining potential solutions using existing machine learning methodologies. We introduce a new taxonomy categorizing DCT strategies into forward, backward, and proactive contact tracing. We examine the integration of privacy-preserving strategies, such as federated learning, with big data and machine learning to enhance the privacy and security of DCT systems. Additionally, we demonstrate how convolutional neural networks, graph neural networks (GNNs), and transformer-based learning optimize DCT strategies. As a solution to limited data availability in the early phase of pandemics, we show how generative adversarial networks (GANs) can be used to tackle such challenges in DCT.

(2) Machine Learning for Infodemic Risk Management

The COVID-19 pandemic has led to an infodemic of unprecedented scale with the rapid spread of misleading or false information via online social networks. Network analysis plays an indispensable role in fact-checking efforts, modeling and learning infodemic risks through large-scale graph computations and statistical processes. Drawing insights from our work in [2], which showcases the efficacy of the proposed graph algorithm in handling efficient computations on large-scale graphs, we identify a unique intersection where these methodologies can be applied to decode the complexities of infodemic propagation. Notably, our contributions in [2] received the Honorable Mention Award at the MIT/Amazon/IEEE Graph Challenge 2018. Building upon the foundational principles established in [2], in [3], we present MEGA (*M*achine Learning-*E*nhanced *G*raph *A*nalytics), a novel framework enabling automated machine learning for mega-scale graph datasets. Within MEGA, we employ a blend of feature engineering and graph neural networks to enhance learning performance and enable parallel computation on extensive graphs. Infodemic risk analysis serves as a unique application of MEGA, encompassing spambot detection through triangle motif counting and influential spreader identification via distance center computation. Performance evaluations using the COVID-19 Infodemic Twitter dataset underscore MEGA's superior computational efficiency and classification accuracy. Further examining this issue, in [4], we introduce TrumorGPT, a maximum truth-seeking AI. TrumorGPT synergizes large language models (LLMs) like GPT and knowledge graphs to facilitate automated fact-checking, demonstrating efficacy on large-scale datasets.

(3) Artificial Intelligence for Large-Scale Learning

The COVID-19 pandemic has significantly reshaped learning environments and modalities, elevating the importance of self-learning in STEM education. In [5], we offer a comprehensive review of transformer-based LLMs for AI-assisted programming, showcasing the potential of LLMs in facilitating programming learning. In [6], we introduce an AI-assisted programming tool that integrates LLMs with Xcode, enhancing developer productivity in the Apple software ecosystem. Through advanced natural language processing techniques and a chat interface, our software facilitates code generation, autocompletion, and interactive decision-making, as

evidenced by case studies demonstrating its efficacy with LLMs like GPT. Continuing this line of research, in [7-8], we design AI-driven software tools aimed at promoting interactive self-learning in STEM education.

Junior Faculty Research

(1) Generative AI for Personalized Education

Building on my research in AI-driven learning tools, my recent work explores how generative systems can advance personalized education. In [9], we introduce MCQGen, an LLM-powered framework that automatically generates multiple-choice questions for blended and flipped classrooms. By integrating retrieval-augmented generation with advanced prompt engineering, MCQGen produces contextually relevant questions aligned with curriculum objectives and common student misconceptions. Selected as a Featured Article, this work demonstrates how automated question generation can reduce educators' workload while providing adaptive learning materials for students. In [10], we extend this direction through a multimodal LLM-based approach for personalized vocabulary learning with AI-generated images. Guided by prompt engineering and dual coding theory, the system creates customized flashcards that pair words with visuals tailored to different developmental stages and cultural contexts to enhance engagement and support age-appropriate learning experiences.

(2) AI for Information Integrity and Discovery

A central focus of my current research is leveraging AI to strengthen information integrity and enhance knowledge discovery. In [11], we enhance TrumorGPT, a graph-based retrieval-augmented LLM designed for fact-checking public health claims. The system constructs and queries semantic health knowledge graphs updated with recent medical information to verify "trumors" (true rumors) and debunk misinformation. Through few-shot learning and real-time retrieval, TrumorGPT mitigates the limitations of static training data and hallucination in standard LLMs. Evaluations on COVID-19 and other health datasets demonstrate its effectiveness in providing accurate and timely verifications, contributing to greater public trust in health communication. In [12], we extend this direction to scientific knowledge discovery by assessing how LLMs such as GPT-4 can function as intelligent research assistants. Our study evaluates their ability to retrieve known references from paper abstracts, identify related literature, and suggest new, relevant works. We introduce metrics such as reference recall and hallucination rate to quantify performance, showing that with augmented features, modern LLMs can effectively retrieve key citations while reducing fabricated outputs.

(3) AI for Networks and Systems Analysis

My research also extends to applying AI in the analysis and optimization of complex networked systems. In [13], we conduct a comprehensive review of how LLMs can transform network engineering and management. This work identifies limitations in traditional communication networks, such as static configurations and manual operations, and discusses how generative AI can automate network design, optimize traffic routing, and enhance security. It also outlines emerging opportunities for integrating LLMs into networking tasks ranging from configuration to analytics, offering a roadmap toward more adaptive and intelligent network systems. In [14], we apply network science to the academic domain through citation network analytics for measuring research novelty and influence. We introduce a quantitative metric that combines a publication's originality with its subsequent impact, revealing how innovative ideas spread within scholarly communities. By analyzing citation graphs of foundational AI papers, we find that early groundbreaking works demonstrate high novelty and influence, while later contributions tend to be more incremental. The analysis also uncovers an inverse relationship between content similarity and novelty, showing that papers diverging most from prior work are often the most original.

Future Research Directions

While data science and AI are now deeply embedded in modern society, pushing the boundaries of these technologies, such as developing systems like LLMs, is widely acknowledged to be a challenging field. I plan to undertake the following long-term directions in my future research to meet these challenges:

- LLMs have shown exceptional prowess in understanding and generating human-like text, but they still struggle with reasoning in specialized domains and grasping nuanced meanings in diverse datasets. To unlock their full potential across various applications, we must explore how to improve the intelligence and reasoning capabilities of LLMs in domain-specific contexts. This includes developing methods to enhance LLM performance even under limited data or computational resources.
- As LLMs and other AI tools generate an overwhelming volume of information, the reliability of this information becomes critically important. It raises the question: are we receiving authentic information or machine-fabricated misinformation? Addressing this requires research into making AI systems more

trustworthy and truthful. Improving the transparency, factual accuracy, and ethical alignment of LLMs will be essential to ensure users can trust AI-generated content in high-stakes settings.

- The explosion of information has also impacted academic publishing and research. Conferences and journals now receive thousands of submissions, overwhelming the peer review process. With limited human reviewing capacity, it is important to investigate how AI tools, such as LLMs, can help alleviate this burden by assisting in preliminary reviews or filtering of submissions. At the same time, researchers can leverage AI to efficiently search through vast literature and identify relevant, novel work. Using AI to streamline literature reviews and knowledge discovery will enhance research productivity and help the scientific community manage information overload.

I hope to contribute significantly to the rapidly evolving landscape of data science and AI, specifically in the areas of LLMs, trustworthy AI, and AI-driven knowledge discovery. Beyond the immediate scope of these areas, I am also keen on diligently exploring the integration of AI with other disciplines, such as healthcare and education, to maximize the broad societal impact of AI advancements. While having specific milestones is crucial to guiding research efforts, I remain open-minded and constantly alert to emerging research opportunities. As the fields of data science and AI continue to progress and evolve, I am committed to ongoing learning, adeptly adapting to new challenges, and leading innovative research at the frontiers of technology.

References

[1] **C. N. Hang**, Y. -Z. Tsai, P. -D. Yu, J. Chen and C. W. Tan, Privacy-Enhancing Digital Contact Tracing with Machine Learning for Pandemic Response: A Comprehensive Review, ***Big Data and Cognitive Computing, special issue on Digital Health and Data Analytics in Public Health***, Vol. 7, No. 2, 2023.

[2] C. -Y. Kuo, **C. N. Hang**, P. -D. Yu and C. W. Tan, Parallel Counting of Triangles in Large Graphs: Pruning and Hierarchical Clustering Algorithms, ***IEEE Conference on High Performance Extreme Computing (HPEC), MIT/Amazon/IEEE Graph Challenge (Honorable Mention)***, 2018.

[3] **C. N. Hang**, P. -D. Yu, S. Chen, C. W. Tan and G. Chen, MEGA: Machine Learning-Enhanced Graph Analytics for Infodemic Risk Management, ***IEEE Journal of Biomedical and Health Informatics***, Vol. 27, No. 12, pp. 6100-6111, 2023.

[4] **C. N. Hang**, P. -D. Yu and C. W. Tan, TrumorGPT: Query Optimization and Semantic Reasoning over Networks for Automated Fact-Checking, ***58th Annual Conference on Information Sciences and Systems (CISS)***, 2024.

[5] M. F. Wong, S. Guo, **C. N. Hang**, S. W. Ho and C. W. Tan, Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review, ***Entropy, special issue on Statistical Machine Learning with High-Dimensional Data and Image Analysis***, Vol. 25, No. 6, 2023.

[6] C. W. Tan, S. Guo, M. F. Wong and **C. N. Hang**, Copilot for Xcode: Exploring AI-Assisted Programming by Prompting Cloud-based Large Language Models, ***IJCAI Symposium on Large Language Models (LLM@IJCAI'23)***, 2023.

[7] J. Li, C. W. Tan, **C. N. Hang** and X. Qi, A Chatbot-Server Framework for Scalable Machine Learning Education through Crowdsourced Data, ***ACM Conference on Learning at Scale***, 2022.

[8] C. W. Tan, L. Ling, P. -D. Yu, **C. N. Hang** and M. F. Wong, Mathematics Gamification in Mobile App Software for Personalized Learning at Scale, ***IEEE Integrated STEM Education Conference***, 2020.

[9] **C. N. Hang**, C. W. Tan and P. -D. Yu, MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning, ***IEEE Access***, Vol. 12, pp. 102261-102273, 2024.

[10] **C. N. Hang** and S. M. Ho, Personalized Vocabulary Learning through Images: Harnessing Multimodal Large Language Models for Early Childhood Education, ***IEEE Integrated STEM Education Conference***, 2025.

[11] **C. N. Hang**, P. -D. Yu and C. W. Tan, TrumorGPT: Graph-Based Retrieval-Augmented Large Language Model for Fact-Checking, ***IEEE Transactions on Artificial Intelligence***, 2025.

[12] **C. N. Hang**, P. -D. Yu, C. W. Tan and D. M. Chiu, Beyond Search: Measuring LLM Performance for Scientific Literature Discovery, ***IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)***, 2025.

[13] **C. N. Hang**, P. -D. Yu, R. Morabito and C. W. Tan, Large Language Models Meet Next-Generation Networking Technologies: A Review, ***Future Internet, special issue on Featured Papers in the Section Internet of Things***, Vol. 16, No. 10, 2024.

[14] **C. N. Hang**, P. -D. Yu, C. W. Tan and D. M. Chiu, When Ideas Go Viral: Measuring Scholarly Novelty and Viral Influence via Citation Network Analysis, ***IEEE Global Communications Conference (GLOBECOM)***, 2025.