Research Statement

Hou Pong Chan

1 Overview

Language models (LMs), and in particular large language models (LLMs), have become transformative AI assistants, adept at tasks ranging from summarizing articles to drafting complex documents. Their growing deployment in high-stakes domains such as healthcare, law, and education, however, introduces critical challenges that demand rigorous scientific inquiry. My research agenda is dedicated to addressing two fundamental pillars required for the responsible integration of LLMs into society: **trustworthiness** and **human-centricity**.

First, an LLM must be **trustworthy**. Its outputs must be factually correct, and it must possess the self-awareness to refuse to answer questions beyond its knowledge. The phenomenon of "hallucination"—generating factually incorrect content—not only fuels misinformation but also risks catastrophic consequences in critical decision-making contexts. Second, an LLM must be **human-centered**, capable of reliably understanding and adapting to the unique needs, preferences, and emotional states of its users. A model that lacks sentiment awareness or misinterprets user preferences can deliver a poor experience or, in sensitive situations involving grief or crisis, cause genuine harm.

In response, my research develops cutting-edge machine learning methods to build more trustworthy and human-centered language models. To enhance trustworthiness, I have pioneered a multi-faceted approach that tackles hallucination at every stage of the generation process: (1) detecting hallucination risk before generation through representation probing, (2) augmenting knowledge during generation via multi-modal Retrieval-Augmented Generation (RAG), and (3) evaluating and rectifying errors after generation using fine-grained factuality assessment and automated post-editing. To foster human-centricity, my work enables LLMs to accurately predict user sentiment and preferences, generate responses aligned with those needs, and provide actionable recommendations.

Furthermore, I have successfully transferred the above research insights to build more trustworthy **multimodal large language models (MLLMs) for the medical domain** and a **culturally aware and safe multilingual LLM** for underserved populations.

2 Prior Work

2.1 Trustworthy LLM and MLLMs

My work on trustworthiness systematically addresses the challenge of hallucination from detection and prevention to evaluation and correction.

2.1.1 Detecting Hallucination Risk before Generation via Representation Probing

Identifying whether a query exceeds an LLM's knowledge boundary before generation is a critical and efficient step for reducing hallucinations. We present the **first study analyzing how LLMs recognize knowledge boundaries across languages** by probing their internal representations [21]. Our empirical results reveal how knowledge boundaries are encoded in hidden states, motivating a training-free alignment method that transfers knowledge-boundary perception across languages via representation steering, thereby reducing hallucination risk in low-resource languages.

2.1.2 Enhancing Factual Grounding with Multi-modal RAG

To enrich an LLM's knowledge at inference time, my research has advanced Retrieval-Augmented Generation (RAG) in two key ways. First, we built a strong multi-modal retriever by developing a novel language-centric omnimodal representation learning framework, LCO-EMB, which achieved state-of-the-art performance on standard embedding benchmarks [22]. This work also led to a fundamental discovery: the Generation-Representation Scaling Law (GRSL), which demonstrates that a language model's multimodal representation capacity scales positively with its generative ability.

Second, to ensure MLLMs can robustly use retrieved content, we proposed **the first retrieval-aware tuning approach for multimodal document reading** [11]. Our method trains MLLMs to effectively incorporate knowledge from retrieved multimodal documents while ignoring irrelevant or noisy information, leading to a 4.6% relative improvement in the correctness of generated responses.

2.1.3 Developing Fine-Grained Factual Evaluation

Moving beyond simple binary factuality judgments, I developed fine-grained evaluation methods to provide deeper, actionable insights into model weaknesses. We proposed an automatic evaluation model that not only identifies factual errors in summaries but also classifies their specific type (e.g., extrinsic noun phrase error) and highlights the evidential text spans in the source text, enhancing interpretability [8].

In addition, we also proposed a novel task and a new dataset of detecting and localizing non-factual information in social media posts [15], which aims to identify the text span in the social media posts that is being manipulated, as well as identifying the corresponding supporting evidence text span in the source news article. To address this task, we further proposed a pipeline that first personalizes opinions from factual information and then detects non-factual information, which outperforms GPT 3.5 [17].

Furthermore, because world knowledge evolves, we examine whether LLMs can utilize updated facts presented in the retrieved context. We release a summarization benchmark of news articles published after prior LLMs' knowledge cutoffs [10]. We find that many LLMs default to outdated parametric knowledge rather than leveraging the input documents.

2.1.4 Automating Factual Error Correction

I have developed methods to automatically rectify factual errors in model-generated content. For the text-only setting, we introduced a zero-shot factual error correction pipeline that corrects responses according to a trusted source document without requiring any labeled training data [14]. Remarkably, our zero-shot method outperforms fully-supervised approaches. For the multimodal setting, we introduced a chart caption correction task and a novel, reliable automatic evaluation metric based on visual entailment that highly correlates with human judgment [16]. Our two-stage correction framework for this task surpasses the performance of advanced MLLMs like GPT-4V.

2.2 Human-centered Language Models

My research on human-centered language models focuses on three key areas: understanding users' sentiment and stance, aligning generation with users' preferences, and delivering accurate recommendations to users.

2.2.1 Predicting User Sentiment, Stance, and Beliefs

To build models that truly understand users, I developed techniques to predict and track their sentiment and stance. We first designed an explainable multi-task framework that jointly predicts sentiment and generates a summary for user reviews, where summarization provides both explainability and a performance boost for sentiment classification [2].

To track opinions over time, we introduced a novel task of summarizing a specific person's stance on an event across multiple documents [18], a key capability for understanding public discourse. Our

pipeline-based approach for this task significantly outperformed GPT-3 [1]. We also pushed the boundary from static analysis to dynamic prediction by introducing a task to forecast a user's sentiment towards a future event based on their profile and interaction history [19]. Our model, which induces a belief-centered graph to learn user representations, achieves highly accurate sentiment forecasting [20].

2.2.2 Aligning Generation with User Preferences and Requirements

We first focus on summarization. We introduce **the first reinforcement learning (RL) approach for keyphrase generation**, with an adaptive reward that increases information coverage while limiting redundancy [3]. Our method consistently improves the informativeness of generated keyphrases across five datasets. We also propose a condense-then-select pipeline that first condenses sentences and then selects them to form a summary, significantly improving the informativeness of generated summaries in news and scientific domains [6].

Users often have diverse requirements for generated content (e.g., length) at test time. To address this, we design an RL framework based on a constrained Markov decision process (CMDP) that enables users to more effectively control specific attributes of summaries at inference, including length, abstractiveness, and targeted entities [7]. Carefully designed constraints ensure adherence to user-specified requirements.

To generate coherent long-form text guided by user-specified talking points, we introduce a dynamic content planner trained with a coherence-based contrastive objective using four types of hard negatives [13]. On opinion generation tasks, our method produces more coherent, content-rich text than strong baselines. We further propose a multi-task training strategy that integrates planning and reviewing with end-to-end generation, leading to substantial improvements in long-form coherence [12].

2.2.3 Developing Recommendation Algorithms in Social Media and Education

I have also designed recommendation algorithms for practical domains. For social media, I developed a novel RL approach with a permutation-invariant encoder to recommend the optimal set of comments in a discussion thread to users, achieving state-of-the-art performance [5]. In education, I addressed the challenge of peer grading by designing probabilistic graphical models to account for student reliability and bias [4] and a trust-aware multi-armed bandit algorithm to optimize task assignment [9].

2.3 Translating Research Insights into More Trustworthy and Human-centric LLMs

I apply these research insights to develop trustworthy, human-centered LLMs for healthcare and multilingual domains.

In healthcare, I built Lingshu, MLLMs for unified multimodal medical understanding and reasoning [23]. To reduce hallucinations, I design more factually accurate multimodal medical knowledge distillation and an LLM-based validator that checks the factual consistency of synthesized reasoning traces, improving training data quality. Lingshu models achieve SOTA results on standard medical multimodal/textual QA and report generation benchmarks for 7B and 32B model sizes, and **have been downloaded over 21,000 times within three months**.

For multilingual applications, I developed SeaLLMs v3, multilingual LLMs for Southeast Asian languages [24]. I devise a supervised fine-tuning method that teaches models to refuse out-of-knowledge and culturally unsafe questions, reducing hallucinations and improving safety. SeaLLMs v3 produces substantially fewer hallucinations and more culturally safe responses than similarly sized LLMs. At the United Nations' AI for Good Global Summit (2024), SeaLLMs was selected for **"Innovation to Expand Impact: AI for Good Case Studies"** and received the **Best Innovate for Impact Use Case Award**.

3 Future Research Directions

My future research will build directly upon my prior work on trustworthy and human-centered LLMs to tackle the next frontier of challenges in creating safe, reliable, and capable MLLMs and agentic frameworks.

3.1 Interpretability and Control for Safe Multimodal LLMs

Interpretation techniques can reveal causes of unsafe outputs and enable targeted inference-time control. While my work on representation probing has shown promise for unimodal LLMs, ensuring the safety of MLLMs requires new interpretability and control techniques that can operate across modalities. The core challenges are: (1) identifying how abstract concepts (e.g., politeness, honesty) are represented consistently across vision, text, and their fused representations; (2) ensuring that steering in one modality does not cause unintended side effects in another.

Building on my previous research in representation probing, I will develop probing methods to discover consistent concept representations across different modalities and steering techniques that enable targeted edits in one modality while enforcing cross-modal consistency. Beyond deployment-time steering, I will create representation probing tools to flag training samples likely to induce unsafe behaviors, providing safety guardrails during multimodal data curation and training.

3.2 Trustworthy Agentic Reinforcement Learning

The emergence of LLM-powered agents, which act autonomously in complex environments, raises the stakes for trustworthiness. Current agentic RL frameworks suffer from critical flaws: (1) final-answer-only rewards encourage shortcut learning, yielding confident but incorrect intermediate reasoning that amplifies hallucinations; (2) verifiable rewards for emotional and social intelligence capability are underdeveloped; and (3) agents are prone to "sycophancy"—agreeing with a user's flawed premises to maximize reward, which undermines decision quality.

Drawing on my experience in RL and factuality, my goal is to build a trustworthy agentic RL framework. I will develop: (1) process-based reward models that assess the factual and logical integrity of each intermediate reasoning step, inspired by my work on fine-grained factuality; (2) novel reward functions to promote social and emotional intelligence; and (3) alignment techniques that optimize for the user's latent, long-term goals rather than their stated, and potentially flawed, preferences.

3.3 Reliable MLLM Architecture and Agentic Framework for 3D Medical Data

While 2D medical imaging has seen progress, analyzing volumetric 3D medical data like CT and MRI scans remains a major challenge. Existing MLLMs struggle to understand complex 3D anatomical structures and capture spatial relationships across hundreds of slices. Current models also lack transparency, hallucinate, and produce inconsistent reasoning; they lack the ability to perform systematic, evidence-based reasoning as a human clinician would.

Building on my experience with medical MLLMs and multi-modal RAG, I will pursue a two-pronged research agenda. First, I will design effective MLLM architectures for 3D volumetric medical data, including vision encoders that capture complex anatomy and long-range 3D dependencies, and efficient 3D patch embeddings tailored to medical imaging structure. Second, I will develop multimodal medical agentic frameworks that can orchestrate specialized medical tools, retrieve clinical guidelines, and synthesize evidence to perform transparent, step-by-step diagnostic reasoning. The ultimate goal is to create a reliable AI partner for clinicians that not only identifies findings but also explains its reasoning with direct grounding in established medical evidence, augmenting the capabilities of human experts.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Hou Pong Chan*, Wang Chen*, and Irwin King. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In *Proceedings of SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1191–1200. ACM, 2020.

[3] Hou Pong Chan, Wang Chen, Lu Wang, and Irwin King. Neural keyphrase generation via reinforcement learning with adaptive rewards. In *Proceedings of ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2163–2174. Association for Computational Linguistics, 2019.

[4] Hou Pong Chan and Irwin King. Leveraging social connections to improve peer assessment in moocs. In *Proceedings of The Web Conference (WWW) Digital Learning Track, Perth, Australia, April 3-7, 2017*, pages 341–349. ACM, 2017.

[5] Hou Pong Chan and Irwin King. Thread popularity prediction and tracking with a permutation-invariant model. In *Proceedings of EMNLP, Brussels, Belgium, October 31 - November 4, 2018*, pages 3392–3401. Association for Computational Linguistics, 2018.

[6] Hou Pong Chan and Irwin King. A condense-then-select strategy for text summarization. *Journal of Knowledge-Based System (KBS)*, 227:107235, 2021.

[7] Hou Pong Chan, Lu Wang, and Irwin King. Controllable summarization with constrained markov decision process. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1213–1232, 2021.

[8] Hou Pong Chan, Qi Zeng, and Heng Ji. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of ACL, Toronto, Canada, 9-14 July 2023*. Association for Computational Linguistics, 2023.

[9] Hou Pong Chan, Tong Zhao, and Irwin King. Trust-aware peer assessment using multi-armed bandit algorithms. In *Proceedings of The Web Conference (WWW) 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 899–903. ACM, 2016.

[10] Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S. Chao. Temposum: Evaluating the temporal generalization of abstractive summarization. In *Proceedings of EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023.

[11] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*. Association for Computational Linguistics, 2025.

[12] Zhe Hu, Hou Pong Chan, and Lifu Huang. MOCHA: A multi-task training approach for coherent text generation from cognitive perspective. In *Proceedings of EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10324–10334. Association for Computational Linguistics, 2022.

[13] Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. PLANET: dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of*

ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 2288–2305. Association for Computational Linguistics, 2022.

[14] Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. Zero-shot faithful factual error correction. In *Proceedings of ACL, Toronto, Canada, 9-14 July 2023*. Association for Computational Linguistics, 2023.

[15] Kung-Hsiang Huang, Hou Pong Chan, Kathleen McKeown, and Heng Ji. ManiTweet: A new benchmark for identifying manipulation of news on social media. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11161–11180, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[16] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] OpenAI. Chatgpt: Large-scale language model. https://openai.com/research/chatgpt, 2023. Accessed on June 27, 2023.

[18] Revanth Gangi Reddy, Heba Elfardy, Hou Pong Chan, Kevin Small, and Heng Ji. Sumren: Summarizing reported speech about events in news. In *Proceedings of AAAI 2023*, 2023.

[19] Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. Measuring the effect of influential messages on varying personas. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of ACL, Toronto, Canada, 9-14 July 2023*. Association for Computational Linguistics, 2023.

[20] Chenkai Sun, Jinning Li, Yi Fung, Hou Pong Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. In *Proceedings of EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023.

[21] Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, and Yu Rong. Analyzing LLMs' knowledge boundary cognition across languages through the lens of internal representations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24099–24115, Vienna, Austria, July 2025. Association for Computational Linguistics.

[22] Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, and Yu Rong. Scaling language-centric omnimodal representation learning. *Under-review in NeurIPS*, 2025.

[23] Weiwen Xu*, Hou Pong Chan*, Long Li*, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[24] Wenxuan Zhang*, Hou Pong Chan*, Yiran Zhao*, Mahani Aljunied*, Jianyu Wang*, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao, editors, *Proceedings of the 2025 Conference*

of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 96–105, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.