Jiankun (David) Wei

Researcher, Undergraduate Student

+1 (306) 201-9129 jiankun.wei@mail.utoronto.ca Toronto, ON, Canada
in/jiankun github.com/david-wei-01001 Personal Website

Profile

Passionate researcher with a robust background in Computer Science, specializing in Generative Modeling, Large Language Models, and Robotics. Possesses three years of research experience, including speculative decoding, integrating Large Language Models (LLMs) for robotics control, and object detection. Demonstrate academic excellence through high university grades and a diverse portfolio of projects, distinguished by exceptional communication skills and self-motivation.

Research Interest: Generative Modeling, Large Language Models, Embodied Agents/Robotics

Education

Honours Bachelor of Science in Computer Science *University of Toronto* **Toronto, Canada** *2020-PRESENT*

- **GPA**: 3.99/4.00
- **Audited**: CSC2221 - Theory of Distributed Computing, MAT344 - Introduction to Combinatorics
- **Poster Presentiation** in UGSRP (Undergraduate Summer Research Program) Research Showcase
- Attended TCURC (Trinity College Undergraduate Research Conference)
- Attended Toronto Ethics in AI Symposium

Awards

- **Finalist** in 2024 MIT Bitcoin Hackathon: Scaling Up
- **6T5 Scholarship for High Academic Achievement** for the 2021 Academic Year.
- **Drew Thompson Scholarship for High Academic Achievement** for the 2021 and 2022 Academic Year.
- **Dean's List Scholar in the Faculty of Art & Science** form 2021 to 2024

Research Experiences

Researcher, *(Department of Computer Science, University of Toronto)* **Toronto, ON, Canada** *05/2024 - 10/2024*

- Conducted extensive research on various speculative decoding methods of LLM inference, including self-drafting vs. independent drafting, lossless vs. approximate decoding, and greedy decoding vs. speculative sampling.
- Explored and visualized speculative behavior, measuring metrics such as total generation time, latency, and correct speculation rate, and traced time taken per token for each iteration as a potential side channel.
- Fingerprinted the speculative patterns to execute attacks using advanced machine learning techniques such as Random Forest, Bottleneck-CNN, and pretrained Transformers, achieving 90% accuracy in close world attacks and $15\% \sim 30\%$ in slightly open world attacks.
- Design attacks to reveal 2 of the 3 parameters of Lookahead Decoding speculation mechanism, and leak contents in the datastore used for the REST speculation mechanism.
- **Outcome:**
  - Poster presentation at the Undergraduate Summer Research Program of the University of Toronto
  - first-author manuscript currently under review at the MLSYS conference

Researcher, *(MEDCVR Lab, University of Toronto)* **Mississauga, ON, Canada** *01/2024 - 05/2024*

- Spearheaded the integration of Large Language Models as the lower-level controller for robotics while granting fault tolerance by implementing multi-agent discussion and active re-planning to detect and correct failures.
- Designed few-shots prompt and deployed Google Gemini within Unity using C#, significantly enhancing real-time application performance.
- Contributed to the LLMUnity package and participated in the macOS testing, enhancing adoption and compatibility.
- Established a seamless remote connection framework to Google cloud within Unity, enabling smooth communication with the Gemini model.
- Regularly prepared and presented comprehensive weekly PowerPoint updates, highlighting ongoing progress and key developments.

Volunteer Researcher, *(AIT Lab, ETH Zürich)* **Remote** *07/2023 - 11/2023*

- Proficient with Detectron2 and Torchvision's Faster R-CNN libraries for advanced object detection and short-term object interaction anticipation.
- Adopt X101-FPN model in Detectron 2 library as the backbone to analyze ego-centric video frames.
- Skilled in analyzing and interpreting large source code bases, enhancing problem-solving and development efficiency.
- Responsible and efficient in managing GPU-accelerated server resources, ensuring fair use without compromising others' processes and maintaining optimal application performance.
- Highly self-motivated team contributor, driven to achieve collective goals and advance project progresses.

Researcher, *(Munk School of Global Affairs & Public Policy)* **Toronto, ON, Canada** *09/2021 - 04/2022*

- Identified underlying threats in the Chinese real estate sector such as capital chain stresses and landlord complaints.
- Expertly processed and tokenized free-form Chinese text paragraphs, performing sentiment analysis using Python libraries NLTK and Jieba to identify emerging trends in the real estate sector, highlighting potential risks.
- Demonstrated comprehensive expertise in utilizing Excel for advanced data manipulation and analysis, as well as employing Python to enhance and streamline the integration and computational processing of Excel datasets.
- Skilled in scrutinizing annual reports to assess assets and liabilities, utilizing financial websites such as Qichacha, Tianyancha, and East Money to gather critical financial data and insights.
- Highly dedicated, taking full responsibility for my work, and effectively collaborating with team members to ensure project success and timely delivery of ideas.

Projects

LLM-Enhanced Robotics Manipulation *01/2024 - 04/2024*

- Implemented adaptive control strategies in robotics by utilizing Gemini model for image understanding and decision making, significantly enhancing real-time feedback and task execution efficiency.
- Implemented multi-agent discussion algorithm to stabilize Gemini outputs and reduce hallucination.
- Integrated LLMs as robot's low-level controller, enhancing decision-making and autonomous error detection and adjustments.

Crypto Care *GitHub Repository* *04/19/2023 - 04/20/2023*

- Created an online cryptocurrency donation platform for non-profit organizations.
- Implemented the database for the CryptoCare project including user registration, secure login authentication, real-time wallet balance updates, and comprehensive documentation of donation activities.
- Significantly contributed to the project's UI design and coded the interfaces, creating a beautiful, comfortable appearance with smooth animations for transitions and clicks.
- Rapidly mastered JavaScript and React within 2 days with no prior experience, demonstrating a strong capacity for learning and applying new techniques to deliver a sophisticated and functional application.
- Collaborate with teammates to implement MetaMask Ethereum transactions and balance checks on the Sepolia testnet.

Multi-Style Transfer *GitHub Repository* *01/2023 - 04/2023*

- Implemented multi-style transfer algorithms to blend multiple artistic style to photographs utilizing both CycleGAN and Neural Style Transfer (NST)
- Conducted a thorough comparison of the two methodologies, showcasing differences in style application through detailed visuals and performance benchmarks (SSIM, FID, Style Consistency).
- Demonstrated proficiency in PyTorch and TensorFlow frameworks, alongside a deep understanding of the CycleGAN architecture.

Collaborative Community Software *GitHub Repository* *09/2021 - 12/2021*

- Engineered a learning community platform to foster sharing of educational materials, employing rigorous software development methodologies.
- Demonstrated good software developing habits by writing clear and easily comprehensible specification documents, applying various design patterns to solve complex problems effectively, and implementing comprehensive testing to guarantee code quality.
- Managed the project's evolution with professional Git usage and detailed code documentation, facilitating smooth collaboration within the team and ensuring efficient progress tracking and code integration.

Other Experiences

QA Engineer, *(Uken Games Inc.)* **Remote** *05/2023 - 04/2024*

- Quickly and accurately develop comprehensive test cases while providing valuable reviews for peers' tests.
- Proficient in interpreting large specification documents, ensuring effective communication with product teams and developers.
- Well-versed in software testing life-cycles, including A/B tests, rinse requests, smoke tests, and product sanity checks.
- Familiar with essential testing and monitoring tools such as TestRail, Bridge, and Kibana, facilitating thorough testing processes.

Volunteer Executive Member *(University of Toronto Buddha's Light Club)* **Toronto, ON, Canada** *09/2022 - 04/2023*

- Strategically planed and coordinated club events, enhancing community engagement and member participation.
- Designed and managed engaging content for social media platforms to promote club activities and increase online presence.
- Participated in weekly Buddhist worship ceremonies, deepening cultural understanding and personal practice.

Skills

- **Soft Skills:** Self-motivation, Creative Problem-Solving, Communication, Accountability, Presentation, Teamwork, Adaptability

- **Programming Languages:** Python, Java, C, C#, JavaScript, HTML5, CSS3, Matlab, R
- **Libraries & Tools:** Unity, PyTorch, OpenCV, TensorFlow, React, Detectron 2, NumPy, SciPy, Panda, Matplotlib
- **Software:** PowerPoint, FireBase, TestRail, Bridge, Kibana, Jira
- **Language:** Chinese (Native), English (Fluent), French (Beginner)
- **Hobbies:** Cooking, Meditation, Reading, Philosophy