JINJIE NI

Email: jinjieni@nus.edu.sg | Homepage: https://jinjieni.github.io/

Research Interests

Large (multi-modal) language models pretraining; Reinforcement learning; Diffusion models

Experiences

Academia

National University of Singapore 2023 - present
Research Fellow
- Foundation Models.

Nanyang Technological University 2020 - 2023
Ph.D. in Computer Science
- Efficient Language Models and Dialogue Systems.

Harvard University, Institute for Applied Computational Science
Research Assistant (remote)
- VAE-GAN variants.

Northwestern Polytechnical University 2016 - 2020
B.Eng. in Electrical Engineering
- Multimodal Models.

Industry

Research Associate at Sea AI Lab, Singapore Oct 2024 – Present
Sea AI Lab
- In charge of LLM pretraining and architectures, (multi-modal) reinforcement learning for reasoning, and diffusion language models.

Research Intern at Alibaba Group, Singapore April 2022 - Oct 2022
DAMO Academy
- In charge of modality alignment for pre-trained models. Worked with Dr. Yukun Ma.

Research Intern at Continental Sept 2020 - March 2022
Continental-NTU Corp Lab
- In charge of fusing task-oriented and open-domain dialogue agents. Worked with Dr. Rui Mao.

Research Intern at Chinese Academy of Sciences
Institute of Automation Feb 2020 - June 2020
- In charge of anchor-free position estimation and object detection. Worked with Dr. Sen Xin.

Institute of Computing Technology Oct 2018 - Nov 2018
- Training abstractive summarization models. Worked with Dr. Shuai Jiao.

Featured Research

For full publication list, see Google Scholar.

- **NoisyRollout**

  - NoisyRollout: Reinforcing Visual Reasoning with Data Augmentation. Neurips 2025 in submission. [Twitter1][Twitter2]

  - Xiangyan Liu*, **Jinjie Ni***, Zijian Wu*, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, Michael Qizhe Shieh.

  - NoisyRollout is a **simple, zero-cost** reinforcement learning improvement method that achieves **state-of-the-art visual reasoning and perception results** across five out-of-domain benchmarks,

demonstrating exceptional sample efficiency (2.1K training samples) and scalability without requiring additional training costs or complex modifications to the RL objective.

- **SynthRL**
  - SynthRL: Scaling Visual Reasoning with Verifiable Data Synthesis. Neurips 2025 in submission. [Twitter1][Twitter2]
  - Zijian Wu*, **Jinjie Ni***, Xiangyan Liu*, Zichen Liu, Hang Yan, Michael Qizhe Shieh.
  - SynthRL is a scalable and guaranteed method that automatically synthesizes verifiably correct and more challenging training questions at scale for visual reasoning models from an initial 8K seed dataset, **achieving consistent and significant performance gains** across five out-of-domain visual math reasoning benchmarks, with improvements most pronounced on the hardest evaluation samples where deeper, more complex reasoning is required.
- **RAPID**
  - Long-Context Inference with Retrieval-Augmented Speculative Decoding. **ICML 2025 (spotlight).**
  - Guanzheng Chen, Qilong Feng, **Jinjie Ni**, Xin Li, Michael Qizhe Shieh
  - Developed RAPID, a novel retrieval-augmented speculative decoding framework that accelerates long-context inference by **over 2x (up to 2.69x)** while simultaneously enhancing generation quality, boosting performance on InfiniteBench from **39.33 to 49.98** for LLaMA-3.1-8B and improving dialogue quality scores from **2.82 to 4.21** by synergistically integrating the benefits of RAG and long-context models.
- **MixEval-X**
  - MixEval-X: Any-to-Any Evaluations from Real-World Data Mixtures. **ICLR 2025 (Spotlight, top 5.1% Papers)**. [Twitter]
  - **Jinjie Ni**, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, Zian Zheng, Kaichen Zhang, Mahir Shah, Kabir Jain, Yang You, Michael Qizhe Shieh.
  - MixEval-X is the **first** any-to-any, real-world benchmark featuring **diverse input-output modalities**, **real-world task distributions**, **consistent high standards across modalities**, and **dynamism**. It achieves up to **0.98** correlation with arena-like multi-modal evaluations while being way more efficient.
- **MixEval**
  - MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. **NeurIPS 2024** main track (poster). [Twitter]
  - **Jinjie Ni**, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, Yang You.
  - Building golden-standard LLM evaluation from off-the-shelf benchmark mixtures. The **best** LLM evaluation at the time of release for its **SOTA** model ranking accuracy (0.96 correlation with Chatbot Arena) and efficiency (6% the time and cost of running MMLU). Moreover, it's dynamic.
- **OpenMoE**
  - OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models. **ICML 2024** (poster). [Twitter]
  - Fuzhao Xue, Zian Zheng, Yao Fu**, Jinjie Ni**, Zangwei Zheng, Wangchunshu Zhou, Yang You.
  - The **first fully open** MoE-based Decoder-only LLM trained over chinchilla scaling law.
- **InstructWild**
  - Instruction in the Wild: A User-Based Instruction Dataset. Github.
  - **Jinjie Ni**, Fuzhao Xue, Yuntian Deng, Jason Phang, Kabir Jain, Mahir Hitesh Shah, Zangwei Zheng, Yang You.

- The **first** large-scale instruction tuning dataset harvested from the web.

- **GHA**
  - Finding the Pillars of Strength for Multi-Head Attention. **ACL 2023** main track (poster).
  - **Jinjie Ni**, Rui Mao, Zonglin Yang, Han Lei, Erik Cambria.
  - Cutting off redundancy for Transformer layers. **SOTA** efficiency and performance among efficient transformers. Concurrent work of GQA, cited and discussed in the GQA paper.

- **PAD**
  - Adaptive Knowledge Distillation between Text and Speech Pre-trained Models. **ICASSP 2023** (oral).
  - **Jinjie Ni**, Yukun Ma, Wen Wang, Qian Chen, Dianwen Ng, Han Lei, Trung Hieu Nguyen, Chong Zhang, Bin Ma, Erik Cambria.
  - Knowledge distillation between text and speech pre-trained models. The **SOTA** text-speech distillation method at the time of release.

- **HiTKG**
  - HiTKG: Towards Goal-Oriented Conversations via Multi-Hierarchy Learning. **AAAI 2022** (oral).
  - **Jinjie Ni**, Vlad Pandelea, Tom Young, Haicang Zhou, Erik Cambria.
  - The **first** work that trains agents to actively guide the conversations. It ushers in **a new era** of intelligence for dialogue agents. The **SOTA** approach for turn-level dialogue reasoning tasks.

- **FusedChat**
  - FusedChat: Towards Fusing Task-Oriented Dialogues and Chitchat in Multi-turn Conversational Agents. **AAAI 2022** (oral).
  - Tom Young, Frank Xing, Vlad Pandelea, **Jinjie Ni**, Erik Cambria.
  - The **first** attempt of fusing task-oriented and open-domain dialogue systems.

- **Recent Advances in Deep Learning Based Dialogue Systems**
  - Recent Advances in Deep Learning Based Dialogue Systems. **AIRE**.
  - **Jinjie Ni**, Tom Young, Vlad Pandelea, Fuzhao Xue, Erik Cambria.
  - An 80-page systematic review for dialogue systems. One of the **most** cited dialogue system reviews.

Services

Conference PC Member / Reviewer

- Neurips 2025, ICML 2025, ICLR 2025, Neurips 2024, ACL 2024, EMNLP 2024, ACL 2023, EMNLP 2023, AAAI 2023

Journal Reviewer

- Knowledge-Based Systems, Information Fusion, Artificial Intelligence Review, Cognitive Computation

Co-organizer

- MLNLP community