# Jianqiao Lu

+86 15315884246 /+852 63149434 | tjlujianqiao@gmail.com | Website | Github

## Summary

I am a Research Engineer at **ByteDance Seed Foundation Model**, working on **scaling and efficiency** for large-scale foundation models. My focus includes (1) training stability and architecture/system co-design for large-scale pretraining, (2) **long-context** modeling and efficient inference/prefill, and (3) post-training for capability and reliability (SFT/RL), with a growing interest in **agentic systems**. Previously, I conducted research on mathematical reasoning and formal verification with LLMs (NeurIPS/ICLR/ACL/EMNLP).

## Education

- **The University of Hong Kong** *2020 - 2025*
  *PhD, Computer Science*
- **Tongji University** *2015 - 2020*
  *Bachelor, Electronics*

## Experience

- **ByteDance** *2025 - Present*
  *Research Engineer (TopSeed Program), Seed Foundation Model* Beijing, China
  - Contributed to architecture & training design for large-scale foundation model pretraining.
  - Improved training stability and model quality versus prior baselines.
  - Worked on scaling & efficiency: optimizing throughput/cost, and improving long-context training/inference pipeline.
- **ByteDance** *04/2024 - 2025*
  *Research Intern, Seed Foundation Model* Beijing, China
- **Huawei** *04/2022 - 04/2024*
  *Research Intern, Noah's Ark Lab* Shenzhen, China

## Research

*Long Context Prefilling*

- ***(Equal Contribution)*** *FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference*

  **ICLR 2025, achieving a score of 8888 and ranking in the top 0.88% overall**

  We propose a dynamic sparse attention mechanism that optimizes attention patterns in real-time based on input-specific requirements, achieving up to a 10× acceleration compared to full attention while addressing the computational challenges of million-tokens handelling in LLMs.

*AI for Math*

- ***(First Author)*** *AUTOCV: Enhancing Reasoning with Automated Process Labeling through Confidence Variation*

  **NeurIPS 2024**

  An automated process labeling system that significantly enhances the accuracy of reasoning models by detecting and leveraging confidence shifts in reasoning steps, resulting in improvements of up to 34% over self-consistency across math and commensense benchmarks.

- ***(First Author)*** *FormalALIGN: Automated Alignment Evaluation in Autoformalization*

  **ICLR 2025**

  A framework that automates the evaluation of semantic alignment between natural and formal languages in autoformalization, outperforming GPT-4 by 11.58% on FormL4-Basic and 3.19% on MiniF2F-Valid, significantly reducing the reliance on manual verification.

- *Proving Theorems Recursively*

  **NeurIPS 2024**

  Developed POETRY, a recursive proof method that boosts success rates by 5.1% and doubles proof length on miniF2F.

- *FVEL: Interactive Formal Verification Environment with Large Language Models via Theorem Proving*

  **NeurIPS 2024 (Datasets and Benchmarks Track)**

  Introduced FVEL, an interactive formal verification environment that integrates LLMs with Isabelle for neural automated theorem proving, resulting in a 17.39% improvement in problem-solving on SV-COMP and a reduction in proof errors.

*Multi-Modality (Speech & Text)*

- ***(First Author)*** *Improving End-to-End Speech Processing by Efficient Text Data Utilization with Latent Synthesis*

  **EMNLP 2023**

  Developed the Latent Synthesis framework to efficiently utilize textual data for enhancing end-to-end speech processing models, achieving over 22.3% reduction in word error rate for ASR and significant improvements in SLU tasks.

*Benchmarking*

- *MR-BEN: A Comprehensive Meta-Reasoning Benchmark for Analyzing Large Language Models*

  **NeurIPS 2024**

  MR-BEN benchmark consists of 5,975 expert-curated questions across multiple domains to evaluate the meta-reasoning capabilities of LLMs.

- *Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios*

  **ACL 2024**

  A benchmark that evaluates LLMs' abilities in planning, creating, and using tools within real-world scenarios. UltraTool emphasizes complex, multi-step tasks, offering a more realistic assessment of LLMs' tool utilization capabilities beyond simple, synthesized queries.

  Code

*Online Matching*

- ***(Equal Contribution)*** *Online Matching Meets Sampling Without Replacement*

  **WINE 2024**

  This work provides the first competitive analysis of this method, showing its effectiveness in both Online Bipartite Matching and Online Stochastic Matching problems.