Multilingual LLMs

This research focus is on the evaluation and advancement of multilingual large language models (LLMs), with a strong emphasis on linguistic inclusivity and real-world relevance. Efforts go beyond English-centric benchmarks, incorporating low-resource languages and non Western cultures to better represent the world's full linguistic and cultural diversity. Key areas include the development of culturally grounded synthetic datasets and instruction tuning strategies that enhance model performance across a wide range of languages and contexts. The research aims to uncover hidden model failures and design solutions that ensure LLMs are more equitable and robust for speakers of all languages.

PhD students with interest in multilingual natural language processing, fairness and equity in AI, and the representation of under-resourced languages and cultures are encouraged to explore opportunities within this area.

AI Infrastructure

This research focus is centered on developing robust and efficient systems that support cutting edge AI and machine learning workloads, with a particular emphasis on the training and serving of large language models (LLMs). The work spans multiple layers of the technology stack, from ML-based algorithmic innovations to low-level kernel and systems implementations. A key aspect of this research is designing systems that leverage an in-depth understanding of workload characteristics, enabling the discovery of novel optimizations in the rapidly evolving hardware and software landscape of AI. Recent efforts in this area include scheduling optimizations to achieve high throughput and low latency in LLM serving, advanced memory management techniques for large models, and near-zero overhead checkpointing for distributed machine learning training.

PhD students interested in systems for AI, large-scale machine learning infrastructure, algorithmic optimization, and end-to-end performance in LLM deployments are encouraged to consider opportunities in this research area.

Grounded and Verifiable Reasoners

This research focus addresses the critical challenge of ensuring that advanced reasoning models produce outputs firmly grounded in real-world constraints, particularly when reasoning over private or domain-specific data. The aim is to develop efficient reasoning models whose outputs consistently align with domain knowledge, formal logic, and established systems of reasoning, as well as to design verifiers that can rigorously assess this grounding. Key research questions in this area include:

1. How do we train reasoning models such that their outputs are grounded and verifiable by design?

2. How do we verify a model's reasoning output in structured domains like math and code, where only final answers may be easily checkable?

3. Beyond math and code, how do we define and verify grounding in less structured tasks such as document generation?

4. How do we build efficient reasoning models, including efficiency at both training and inference time?

PhD students with a strong interest in reasoning, verification, formal methods, and the intersection of machine learning with structured and unstructured domains are encouraged to explore opportunities within this research area.

GenAI for Education

This research focus explores the transformative potential of GenAI to improve education globally, with particular emphasis on low-resource settings and serving the needs of the global majority. The aim is to create innovative, equitable, and accessible AI-driven educational solutions that empower both students and educators. Key objectives and areas of inquiry include:

1. **Understand the Impact of GenAI in Education:** Investigate how GenAI can support educational practices, improve learning outcomes, and broaden access to quality education—particularly in resource-constrained environments.

2. **Enhancing Mathematical Reasoning:** Develop GenAI that effectively assist students in grasping complex mathematical concepts through step-by-step explanations, clear reasoning, and accurate solutions.

3. **Advancing Visual Reasoning in Education:** Improve the capabilities of GenAI to understand and explain complex visual content, such as geometric figures, diagrams, and charts—especially relevant to STEM subjects.

4. **Multimodal Content Generation:** Create rich interactive educational materials that integrate text, images, videos, and other media as effective visual aids for educators.

5. **Designing Accessible Learning Experiences:** Co-design, iteratively improve, and evaluate educational experiences for children and teachers in schools for the blind. This involves generating accessible content—including audio-first, tactile-first, and Braille materials—and ensuring effective delivery in both physical and hybrid educational settings.

Research initiatives such as Shiksha Copilot and Ludic Design for Accessibility reflect these goals, advancing the mission to enhance learning, promote equity, and empower educators and learners worldwide.

PhD students seeking to harness GenAI to transform education, especially in underserved and low-resource settings, should explore this research area to develop equitable and accessible AI-driven learning solutions.

LLMs for Healthcare

This research focus investigates the transformative applications of Large Language Models (LLMs) in healthcare, with a vision of enabling faster, more accessible, and better-informed decision-making across clinical and community settings. LLMs have the potential to revolutionize care delivery—from empowering chatbots that offer reliable, easy-to-understand health information to patients, to reducing clinical burden by supporting healthcare professionals with timely assistance. Additionally, with advancements in multimodal foundation models, LLMs are now integrating diverse data types, including biomedical imaging and ECG lead data, to assist doctors in generating and refining radiology reports, saving time and improving consistency.

Key research directions include:

1. **Patient Engagement and Education:** Using LLM-powered conversational agents to answer patient queries, provide health education, and support patient self management, especially in resource-constrained environments.

2. **Clinical Documentation and Support:** Assisting medical professionals in generating and refining radiology reports, saving time, and improving the consistency and quality of service.

3. **Decision Support for Community Health Workers:** Delivering timely and accurate guidance grounded in medical guidelines to frontline workers, strengthening the continuum of care beyond hospital walls.

4. **Medical Coding and Billing:** Leveraging the reasoning capabilities of LLMs to power medical coding and billing workflows for revenue cycle management (RCM) providers, further streamlining administrative processes.

These research themes underscore the potential for LLMs to bridge gaps in healthcare access, quality, and equity across diverse care settings.

PhD students interested in AI for healthcare, clinical decision support, health informatics, or equitable technology implementation are encouraged to explore opportunities within this area.

Next-Generation Retrieval Models for Chat, Search, and Recommendation

This research aims to tackle complex, large-scale challenges in information retrieval. The primary goal is to achieve state-of-the-art retrieval accuracy and efficiency across various applications, including search, recommendation systems, and retrieval-augmented generation (RAG) within Copilot experiences.

Following are the specific focus areas:

1. Generative retrieval, an alternative to dense retrieval that seeks to directly generate the identifiers of the most relevant documents for a given input (e.g., Scaling the Vocabulary of Non-autoregressive Models for Efficient Generative Retrieval [KDD '25]).
2. New transformer architectures for improved retrieval accuracy and reduced latency across multiple applications.
3. Methods for injecting parametric knowledge into language models (e.g., MOGIC: a Metadata-infused Oracle Guidance framework for Improved Extreme Classification [ICML '25]).
4. Efficient similarity search over high-dimensional datasets, incorporating techniques in high-dimensional geometry, graph-based algorithms, vector quantization, and data compression (e.g., DiskANN: fast accurate billion-point nearest neighbor search on a single node [NeurIPS '19]).

PhD applicants with interests in information retrieval, large language models, system optimization, or novel algorithms for search and recommendation are encouraged to consider joining this research area.

Reliable Agentic Systems

This research theme focuses on advancing the reliability of agentic systems—AI agents powered by large language models (LLMs) capable of autonomous decision-making and tool use. A full-stack approach is taken, innovating across infrastructure, model design, reasoning strategies, and real-world applications. Current research focus is on two core areas:

1. **Agentic Reasoning:** enhancing the ability of LLM agents to plan, adapt, and coordinate actions across multi-step tasks. By integrating reasoning and tool use through reinforcement learning, significant improvements in both performance and interpretability can be realized. This results in more accurate and robust systems that set new benchmarks in complex problem solving (Agentic Reasoning and Tool Integration for LLMs via Reinforcement Learning - Microsoft Research).

2. **Agentic Safety:** understanding and mitigating the risks posed by autonomous agents in everyday tasks. By building a benchmark suite to rigorously evaluate safety and potential harms across domains such as web interaction, code generation, and textual reasoning, a comprehensive risk assessment can be achieved. In parallel, developing mitigation techniques to reduce unsafe behaviors and improve overall agent alignment ensures more reliable and trustworthy systems.

PhD students interested in AI safety, reinforcement learning, human-AI interaction, or scalable agentic reasoning will find this research area especially valuable.

Practical Cryptography and AI Security

This research focuses on various problems in practical cryptography (e.g., secure multi-party computation, differential privacy) and security problems in AI systems (e.g., cryptographic solutions to preventing information leakage in AI systems). Some recent works include Project EzPC, and Private Benchmarking for AI. A variety of problems are pursued, requiring either the development of new cryptographic protocols or the design of secure systems based on sound cryptographic principles.

PhD students with a strong interest in cryptography, privacy-preserving technologies, and AI security are encouraged to engage with this topic.