1 Research Statement

Siddhant Khare

1.1 Research Direction

My research interest lies at the intersection of **AI systems, distributed systems, and security**, with a focus on building **efficient and secure infrastructure for agentic large language models (LLMs)**.

As LLMs transition from passive text generators to **tool-using, long-horizon agents operating in real software environments**, two fundamental bottlenecks emerge:

1. **Systems bottlenecks** — memory pressure from KV-cache and activations, inefficient inference serving, lack of observability, and poor cost–performance tradeoffs at scale.
2. **Security bottlenecks** — over-privileged tool access, prompt and tool injection, lack of principled authorization, and weak isolation between agents and environments.

My goal is to design systems that address both dimensions together, enabling **reliable, scalable, and least-privilege agentic AI**.

1.2 Core Research Questions

1. **Efficient memory and serving systems for long-context and agentic LLMs** How can we manage KV-cache and intermediate activations across heterogeneous memory tiers (VRAM, RAM, NVMe) with predictable latency, minimal overhead, and security-aware policies to support long-horizon inference and multi-agent workloads?

2. **Least-privilege authorization for tool-using agents** How can we enforce fine-grained, programmable authorization over agent tool calls—grounded in formal access-control models—while preserving usability and performance? Can Zanzibar-style authorization systems be adapted to dynamic agent–tool ecosystems?

3. **Robustness and security of agents in real environments** How do we systematically evaluate and harden agents against adversarial or noisy conditions such as prompt/tool injection, flaky tools, environment drift, and partial observability?

1.3 Prior Work and Preparation

I approach research through **systems-building and measurement**, grounding ideas in working prototypes and real-world constraints.

1.3.1 LLM Systems and Observability

- **TokenVM** explores treating KV-cache and activations as a **virtual memory system**, paging across VRAM, RAM, and NVMe to enable long-context inference without linear slowdown.
- **KV-Cache Profiler** and **LLMTraceFX** investigate observability at the GPU and runtime level, connecting kernel stalls, memory movement, and scheduling effects to model-level

inference behavior.

These projects aim to make LLM performance bottlenecks explicit and actionable, rather than opaque.

1.3.2 Agent Orchestration and Reliability

- **Agentflow** is a control-plane-style orchestration framework for LLM agents, inspired by distributed systems primitives (scheduling, retries, observability, cost controls).
- The system treats agents as managed workloads rather than ad-hoc scripts, enabling reproducible evaluation and failure analysis.

1.3.3 Security and Authorization

- **A2AS** explores agent-to-agent security and safe tool usage, focusing on threat models, privilege boundaries, and runtime enforcement.
- **actionsec** applies similar ideas to CI/CD, examining how LLMs introduce new attack surfaces in developer workflows.
- I am a **maintainer of OpenFGA**, a Zanzibar-inspired authorization system, where I have contributed to correctness, performance, and maintainability of production-grade access control infrastructure.

This experience gives me a practical foundation to research **authorization and security for agentic systems**, not just in theory but as deployable systems.

1.4 Research Agenda

In a PhD program, I aim to pursue the following directions:

- **Secure KV-cache and activation management** Designing memory-tiering policies that are performance-aware and leakage-aware, with explicit tradeoffs between latency, cost, and isolation.

- **Authorization frameworks for agent tool use** Extending distributed authorization models (e.g., Zanzibar/OpenFGA) to dynamic, multi-agent tool ecosystems, with formal guarantees and empirical evaluation.

- **Evaluation of agent robustness** Building benchmarks and harnesses that stress agents under realistic failure and adversarial conditions, enabling principled comparison of mitigation strategies.

I am particularly interested in research environments that value **systems rigor, empirical evaluation, and real-world relevance**, and I am motivated to publish work that bridges AI systems, security, and distributed systems communities.

1.5 Why Me

My background is as a systems engineer and open-source maintainer. I currently work at **Ona** (formerly Gitpod), building production systems used by developers at scale. Alongside this, I have independently built and published multiple open-source AI systems and security prototypes.

I am seeking a **direct PhD path** and am comfortable with intensive coursework and qualifying requirements. I want to focus fully on research, contribute to collaborative lab environments, and develop systems that make agentic AI practical, efficient, and secure.