My name is Bing Han, and I am a Ph.D. candidate in Computer Science at Stony Brook University. My research interests lie in GPU performance optimization and ML/AI systems, specifically in making GPU job sharing more efficient.

My research, conducted under the guidance of Dr. Anshul Gandhi and Dr. Zhenhua Liu, focuses on improving GPU job sharing efficiency. In my recent work, accepted at SoCC 2024, I developed predictive models and optimization techniques based on spatial sharing framework MPS, which reduced completion time by 36% and increased throughput by 1.5x through effective job colocation. In addition, I have modified the Kubernetes scheduler framework to further optimize GPU sharing for machine learning tasks, demonstrating my ability to work on large-scale system design.

Beyond ML systems research, I have practical experience in machine learning model development. I secured first place in the AICUP Chinese Medical Dialogue Analysis Competition by training BERT models that achieved 92% accuracy in analyzing complex medical dialogues.

I am eager to bring my skills in Python, C++, Kubernetes, and machine learning frameworks to collaborate with Microsoft's esteemed researchers. I am confident that this internship will not only contribute to my Ph.D. dissertation but also foster the kind of innovative work that has historically emerged from your summer internship programs.

I look forward to the opportunity to contribute to your exciting research initiatives. Please feel free to contact me at (631) 479-9014 or via email at bingshiunhan@gmail.com.

Sincerely,
Bing Han