**Statement of Purpose**

A new era of the AI implementation is approaching, as we find ourselves at the intersection of machine learning and engineering practice. In the face of the emerging breakthroughs in AI algorithms, the computing infrastructure and engineering challenges like machine learning systems prove to be bottlenecks to the implementation of algorithms and the democratization of technology. As a promising researcher, I am committed to this tech revolution, transforming breakthrough AI algorithms into scalable and efficient machine learning systems through advanced engineering and infrastructure.

My professional engagement as a research assistant in Professor Lu’s team exposed me to machine learning algorithms and training. Nevertheless, a project focused on dataset preparation has revealed that scientific research often relies on tedious yet crucial engineering steps. To support my team's research on person and vessel re-identification, I used 3D models and Blender programming to generate a usable synthetic image dataset. For the first time, I found that engineering work is the prerequisite for the implementation of algorithms.

With this fresh understanding, I made a conscious effort to develop my engineering practice. In my senior year, an excellent opportunity appeared, enabling me to transform my research into a usable system. In the RTDNet project, I developed a video gaze estimation algorithm that achieved state-of-the-art performance on the EVE and EYEDIAP datasets. I not only designed a temporal difference (TD) module and recurrent network architecture to enhance prediction accuracy, but also deployed it as an online service, allowing users to upload videos and obtain inference results via a webpage. To make the idea become reality, I needed to address the preprocessing of diverse input videos, design a stable interface, and manage the stability of GPU-intensive inference during deployment. Experiencing the entire process inspired me: AI algorithms can only truly unleash their value when they are engineered and embedded into a comprehensive process.

To further strengthen my engineering skills, despite facing the most important graduation research and four curricula, I shifted my role from a student of the LI Team Project Module to a teaching assistant. Under heavy load, my scientific and technical thinking, particularly my engineering mindset, contributed significantly. By standardizing and automating repetitive questions raised by students, I developed a website to post FAQs and update it accordingly. Besides, a meeting reservation system is provided for students to schedule meeting with me automatically. It achieved my purpose of automation, and reduced inefficient email exchanges and ad hoc communication. In terms of the meeting slots, I arranged them for the after-lunch period, which fulfilled my TA duties and spared me a large chunk of time to focus on my final project. For the academic courses, I compiled summaries for each lesson based on the exam requirements.

The multiple approaches of managing workloads with standardized, batched, and traceable records supported being competent for my schedule, also allowed me to develop skills in engineering collaboration in this TA position. The rapid advancement of AI has become a double-edged sword, as I've noticed that team members often unconstrainedly use AI to generate code. From a professional angle, instead of improving efficiency when utilizing AI, it resulted in style conflicts, code redundancy, increased communication costs, and ultimately reduced overall output. To tackle this issue and avoid diminishing human subjectivity, I required each team to develop a unified style guideline and establishing a clear definition of task completion. Meanwhile, I reiterated the boundaries of AI's use among the 42 students clearly, making it serve as a "translator," translating natural language requirements into code snippets if necessary, rather than a “worker” that directly generated complete projects. This approach has proven to facilitate smoother student collaboration and minimize rework. The outcome also demonstrated that the essence of integrating team norms with a systems engineering framework is in making AI tools truly improve system development efficiency.

Experiencing “AI-in-the-loop” prompted me to explore the interplay between AI and systems engineering. The emphasis on the core issue of machine learning systems: how to effectively embed AI into software and computing processes while ensuring system reliability and maintainability. The challenge is no longer squeezing another 1% accuracy, but integrating engineering system capabilities (distributed training, model service-oriented, inference optimization, and resource scheduling) into the entire lifecycle of data processing, feature engineering, model iteration, deployment and monitoring, while constraining the system boundaries and ensuring scalability.

My in-depth thinking and my passion for engineering-oriented methods for AI stimulated me to pursue the MS in Electrical and Computer Engineering offered by Carnegie Mellon University. The program aligns with my academic goal of conducting engineering-oriented training rather than simply pursuing algorithmic benchmarks. Diving into the AI/ML Systems concentration aligns with my objective of conducting algorithm research and engineering implementation. Courses such as *18-763: Systems & Toolchains for AI Engineers*, *18-667: Algorithms for Large-scale Distributed Machine Learning and Optimization* and *18-847F: Foundations of Cloud and Machine Learning Infrastructure* will facilitate my understanding of the engineering process from model training to deployment. Given the opportunity to work with renowned scholars, I’m especially interested in working at the Parallel Data Lab under the guidance of Prof. Greg Ganger on large-scale systems for big data, and in Prof. Gauri Joshi’s group on designing scalable algorithms for distributed machine learning. Besides, internships, tech talks, and hackathons will pressure-test ideas and help distill methods into reusable practices.

The theoretical and practical enhancement at CMU will cultivate me into an ML Systems engineer. Upon graduation, I aspire to work at NVIDIA to tackle the scalability, cost, and reliability issues of training/inference. With three to five years’ preparation, I will move to the next stage, pursuing my long-term goal of developing a general-purpose inference platform, aiming at multimodal and real-time applications (such as video, robotics, and AR). Collaborated with like-minded peers, I can promote data-model-system co-design and contribute reusable components and standards to open source community. From building models to engineering systems, I have seen that true innovation lies at their intersection. CMU is where I intend to refine this vision and contribute to the next wave of AI engineering.