The infrastructural network in this information age generates a data deluge in digital space. Information, including confidential data of SSN, financial records, and even national security, is digitally stored, both in a direct or indirect way. However, news on severe computer security vulnerabilities could be seen nearly every day. Establishing systems to control and prevent safety attacks becomes a social consensus. According to Cybercrime Magazine, there will be a global talent shortfall of 3.5 million professionals in the field of Cybersecurity by the year 2025. Solving cybersecurity issues needs joint efforts, and I expect to be one of the contributors.

I built a broad skill set for my advanced study in Cybersecurity during college. Comprehensive training in computer science and software engineering was completed at Xi’an Jiaotong University. I was exposed to knowledge related to computer organization and operation, data structures, computer architecture, computer networks, and software engineering development. Meanwhile, diverse system development courses ingrained in me the importance of crafting secure system architectures, safeguarding sensitive data and resources within systems, and proactively preventing systems from security attacks. These courses and training equipped me to grasp the essence of a system effectively, assisting in identifying potential security vulnerabilities.

In recent years, the combination of artificial intelligence and information security has proven to hold significant research potential. In fact, I undertook initiatives that integrate AI with cyber security. I fulfilled a project to use deep learning models for URL-based phishing website detection. I also finished the research of employing NLP techniques to construct a knowledge graph for CTI report generation, which laid the groundwork for exploring novel technologies in the intelligent handling of malicious traffic based on knowledge-assisted decision-making. The AI-supported security exposure convinced that the integration of AI and security will become a trend in information security. With a strong foundation in software engineering and substantial knowledge and practical experience in machine learning and deep learning, I am well-positioned to be a contributor to this trend.

I joined Dr. Peng Zhang’s cybersecurity research group and researched Phishing Website Detection based on URL Sequences, which was also my first independent project. To learn the tests of phishing websites, I conducted plenty of literature reviews. I collected 50+ highly cited papers published after 2021 related to utilizing URL sequence-based phishing website detection methods. I thoroughly reviewed all the papers and closely studied the 12 most representative ones. I distilled relevant methods, seeking suitable ones to support my research. In the end, I independently established datasets, developed deep learning models, and achieved a significant reduction in detection time while maintaining the accuracy of the detection to be equivalent to the state-of-the-art models. The research was also concluded into a paper.

Innovative ideas were involved in this project. I created models that combined the advantages of CNN, RNN, and attention mechanisms. Using parallel CNN layers with different-sized convolutional kernels, I could extract local features from different dimensions and reduce the time for training and detection. Then, the output of the convolutional layers was put into GRU models. Through this type of RNN model, we could shorten the time with an efficient extraction of contextual information. Finally, a multi-head attention layer was applied for weighted enhancement, followed by a fully connected layer for result output. I also independently established a proprietary dataset. I utilized web scraping and data processing skills to create datasets containing 564,434 latest URLs, consisting of 276,239 phishing website URLs and 288,195 legitimate URLs. The datasets had been open-sourced. In this project, I achieved an accuracy rate of 98.3%, which was close to that of the SOTA models, while significantly improving operational efficiency by 34.93%.

I was greatly attracted by Professor Wang’s research group and eventually joined it to fulfill a knowledge graph for CTI reports based on NLP. The problems in this project were threefold. The writing styles of various reports naturally varied significantly, causing the challenge for models to adapt accurately to the new writing styles of reports, extract entities, and conduct relationship analysis. Meanwhile, using the present limited knowledge base was hard to extract the new entities in updated reports. Finally, there was no highly effective method for extracting relationships between entities at long distances.

Tackling these problems was a bumpy way. To deal with the different writing styles, I classified different types of entities. Specifically, for those with clear structures, such as Indicators Of Compromise (IOC), rule-based algorithms were employed for extraction. In terms of entities without clear structures but included in the knowledge base, the Aho-Corasick(AC) algorithms were used for multi-pattern string matching. To reduce false positives, I used NLP techniques incorporating Part-of-Speech Tagging. In this way, the entities were matched. Problems caused by the limited knowledge base were solved through rule-based algorithms. By analyzing reports from different vendors and sources, I could introduce limited patterns used in text to express those new entities and thus develop pattern rules recognizable by NLP processing. Combining the Part-of-Speech tagging results, dependency graphs for sentences in the report were constructed using the Spacy library to extract new entities. Meanwhile, I cooperate with team members to integrate rule-based methods with a Large Language Model (LLM) using ChatGLM models. The rule-based method was used to filter most of the new entities, followed by ChatGLM filling in the gaps. The method effectively leveraged the simplicity and efficiency of the rule-based algorithm while addressing the resource-intensive nature of LLM. It allowed for subsequent follow-up using LLM to handle omissions in the rule-based algorithms. The final problem regarding the relationship object was solved through Sentence BERT models. I computed the STIX Relationship Objects (SRO) in the knowledge base and embeddings of sentences extracted from reports. Then, I calculated the cosine similarity between these embeddings, normalized them, and utilized the results as confidence values. In this way, relationships between entities could be determined when the confidence exceeded a specified threshold. This lab research experience greatly enhanced my skills in the NLP field and honed my problem-solving capacities, benefiting my future research.

In the future, I hope to further my study in the master degree in Cybersecurity and eventually attain a Ph.D. degree. Your program can expose me to extensive research opportunities, preparing me for further academic trials. Moreover, your core security courses can provide me with systematic training in Cybersecurity. I am especially interested in Applied Cryptography, which is a novel domain I expect to explore. Meanwhile, the in-depth electives can help me learn Cybersecurity from different perspectives. I also want to learn Biometrics to explore security from the perspective of biometric technology, thus generating more diverse perspectives in my future research. I believe that learning from the prestigious faculties in your program will help me get closer to my goals.