بسمه تعالی

امنیت و حریم‌خصوصی در یادگیری ماشین

بخش‌های مسمومیت، استخراج مدل، و حریم‌خصوصی

۱- لطفا نام مقاله انتخابی‌تان را تا قبل از روز امتحان پایانی (۲۴ خرداد) در این صفحه وارد نمایید. پس از وارد کردن نام مقاله حق تغییر آن را بدون هماهنگی با استاد درس ندارید.

۲- اگر به هر دلیلی اعضای گروه شما تغییر کرده است تا روز امتحان گروه جدید خود را تشکیل دهید. اگر تا روز امتحان گروه شما مشخص نشد، گروه شما توسط استاد درس مشخص می‌شود.

۳- ارائه‌ها روز شنبه، دهم تیر ماه، از ساعت ۹ تا ۱۲:۳۰ در کلاس مجازی https://vc.sharif.edu/ch/amsadeghzadeh برگزار می‌شود. اگر فردی با این زمان مشکل دارد، هرچه زودتر به آقای صادق‌زاده اطلاع دهد.

۴- برای ارائه یک پژوهش به موارد زیر دقت فرمایید.

۱. توضیح هدف مقاله
۲. توضیح راه‌حل ارائه شده
۳. بیان واضح ارزیابی
۴. بیان ضعف‌های پژوهش
۵. رابطه این پژوهش و پژوهش‌های مرتبط

۵- هر ارائه به مدت ۲۰ دقیقه انجام می‌شود و ۱۰ دقیقه هم برای پرسش و پاسخ در نظر گرفته شده است.

۶- لیست مقالات نامزد برای ارائه در زیر آمده است. هر گروه باید یکی از مقالات را انتخاب نماید.

۷- امکان انتخاب مقاله خارج از لیست تنها در صورت هماهنگی با استاد درس وجود دارد. از مقالات خارج از لیست زیر که مرتبط با موضوع امنیت و حریم‌خصوصی مدل‌های بزرگ زبانی مانند ChatGPT باشند استقبال می‌شود.

۸- نمره‌دهی ارائه‌ها توسط استاد، دستیار درس، و دانشجویان درس انجام می‌شود.

۹- اولویت تخصیص یک مقاله با گروهی است که زودتر آن را در صفحه ارائه ها وارد نماید. لطفا سطر مربوط به گروه‌های دیگر را ویرایش ننمایید و قبل از انتخاب مقاله بررسی کنید که آن مقاله توسط گروه دیگری انتخاب نشده باشد.

۱۰- در صورت سوال یا ابهام می توانید آن را در کوئرا یا به طور مستقیم با آدرس amsadeghzadeh@gmail.com مطرح نمایید.

لیست مقالات نامزد

1. Extracting Training Data from Large Language Models
2. Large Language Models Can Be Strong Differentially Private Learners
3. Spectral Signatures in Backdoor Attacks
4. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks
5. Trojaning Attack on Neural Networks
6. Poisoning and Backdooring Contrastive Learning
7. The Privacy Onion Effect: Memorization is Relative
8. High Accuracy and High Fidelity Extraction of Neural Networks
9. Reverse-engineering deep ReLU networks
10. Renyi Differential Privacy
11. Deep Leakage from Gradients
12. Certified Robustness to Adversarial Examples with Differential Privacy
13. Privacy risks of securing machine learning models against adversarial examples
14. Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds
15. Label-Only Membership Inference Attacks
16. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning
17. Differentially Private Learning Needs Better Features (or Much More Data)