به نام خدا

دانشکده مهندسی کامپیوتر

امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۴-۱۴۰۳

استاد درس: دکتر امیرمهدی صادق‌زاده

طراحان: سروش وفایی‌تبار، رئوف زارع، علی عبدالهی، علی جاوری

تمرین ششم

مهلت تحویل: ساعت ۲۳:۵۹ دوشنبه ۱۵ بهمن ۱۴۰۳

نکات و قواعد

۱. سوالات خود را زیر پیام مربوطه در Quera مطرح نمایید.

۲. لطفا مطابق تاکید پیشین، حتما آداب‌نامه‌ی انجام تمرین‌های درسی را رعایت نمایید. در صورت تخطی از آیین‌نامه، در بهترین حالت مجبور به حذف درس خواهید شد.

۳. در صورتی که پاسخ‌های سوالات نظری را به صورت دست‌نویس آماده کرده‌اید، لطفا تصاویر واضحی از پاسخ‌های خود ارسال کنید. در صورت ناخوانا بودن پاسخ ارسالی، نمره‌ای به پاسخ ارسال شده تعلق نمی‌گیرد.

۴. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۶_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ (۲۵ نمره)

۱. مقاله لینک را مطالعه کرده و به طور کلی contribution های اصلی آن را توضیح دهید.

۲. بر اساس مقاله، نقاط ضعف مدل‌های قبلی را بیان کرده و توضیح دهید این مقاله چگونه این مشکلات را برطرف می‌کند.

۳. ایده‌ی smoothllm را به طور کامل شرح داده و الگوریتم آن، انواع perturbation ها و نحوه‌ی اعمال آن‌ها را توضیح دهید.

۴. مفهوم k-unstable را توضیح داده و بیان کنید این مفهوم چه کمکی به مدلسازی موجود می‌کند و چرا نیاز به تعریف آن وجود دارد.

۵. مقاله لینک را مطالعه کنید و contribution اصلی این مقاله را بیان کنید.

۶. بر اساس مقاله دو مفهوم Objectives Competing و Generalization Mismatched را به همراه مثال‌های آن‌ها از مقاله بصورت کامل توضیح دهید.

سوال ۲ (۲۵ نمره)

مقاله لینک را مطالعه کرده و به سوالات زیر پاسخ دهید. توجه داشته باشید که نیازی به مطالعه بخش ۳ مقاله نیست. این مقاله به مدلسازی احتمالی برای مسئله jailbreaking پرداخته است.

۱. فرضیات ۱.۲، ۱.۴ و ۲.۴ را بیان کرده و به تفصیل توضیح دهید.

۲. تعاریف ۱.۴، ۲.۴، ۳.۴ و ۴.۴ را ذکر کرده و توضیح دهید.

۳. قضیه ۲ را بررسی کنید و دیدگاه و شهود خود را درباره ویژگی‌ها و خواص این قضیه بیان کنید.

۴. الگوریتم E-RLHF را شرح دهید و توضیح دهید چرا این الگوریتم پیشنهاد شده است.

۵. تابع هزینه الگوریتم RLHF را با الگوریتم E-RLHF مقایسه کرده و تفاوت‌ها و شباهت‌ها را بیان کنید.

سوال ۳ (۱۰ نمره)

با توجه به این مقاله به سوالات زیر پاسخ دهید.

۱. نحوه ساخت خروجی معیوب را در روش *Encrypted Backdoors* توضیح دهید. چرا این روش برای ورودی‌های بدون ماشه [۱] خروجی معیوب تولید نخواهد کرد؟

۲. تفاوت دو روش *Encrypted Backdoors* و *NP-Complete Backdoors* چیست؟ به نظر شما چرا روش *Encrypted Backdoors* بر خلاف دو روش دیگر در شرایط نویزی عملکرد پایداری را ارائه می‌دهد؟

سوال ۴ *Attack on Gandalf* (۵۰ نمره)

در این سوال به روش‌های سنتی مهندسی پرامپت برای حمله به LLM ها می‌پردازیم. سایت گاندولف یک محیط عالی برای سنجیدن مهارت شماست. در این سایت شما با هشت نسخه‌ی مختلف از گاندولف، یک مدل زبانی، روبه‌رو می‌شوید که یک رمز را نزد خود پنهان کرده است. در مراحل ابتدایی، گول زدن گاندولف کار ساده‌ای است ولی به مرور گاندولف قوی‌تر می‌شود. مرحله‌ی هشت آن، آنچنان سخت است که دستیاران آموزشی درس نیز از پس آن برنیامدند.

در این تمرین شما باید مراحل این سایت را رد کنید. چیزی که تحویل می‌دهید باید یک گزارش از رویکرد شما برای رد کردن هر مرحله باشد. برای هر مرحله، پرامپت یا پرامپت‌هایی که به گاندولف داده‌اید و پاسخ گاندولف را گزارش نمایید. گزارش شما در یک فایل جدا از سوالات تئوری، به صورت پی‌دی‌اف و با نام *gandalf_stdnum_name.pdf* باشد.

توجه فرمایید که پاسخ‌های شما با جواب‌های موجود در اینترنت و همچنین پاسخ دیگر دانشجویان مقایسه خواهد گردید. این سوال تجربه‌ی یکتا و لذتبخشی است، سعی کنید که با روش خود گاندولف را شکست دهید، سپس راه‌حل‌های خود را با دیگران مقایسه نمایید.

نحوه‌ی نمره‌دهی این سوال بدین گونه است که مرحله‌ی هشتم امتیازی است. مرحله‌ی هفتم نیز با توجه به عملکرد کل کلاس امتیازی یا غیر امتیازی در نظر گرفته می‌شود. پاسخ‌هایی که با یک پرامپت یک مرحله را رد می‌کنند، نمره‌ی کامل را کسب می‌کنند (نمره‌ی چند پرامپت به مقدار اندکی، کمتر از یک پرامپت است).

موفق باشید.

[۱] *Trigger*