⚡️🚀המאמר היומי של מייק -20.10.24: ⚡️🚀  
RL, BUT DON’T DO ANYTHING I WOULDN’T DO

אוקיי, אחרי כמה סקירות יחסית קלילות הגיע הזמן לסקור מאמר קצת כבד לפחות מהמבט הראשון. המאמר בנושא של אימון מודלי שפה עם השיטות מעולם למידה באמצעות חיזוקים או בקצרה RL. דרך אגב הטענות המתמטיות הרבות שהמחברים הוכיחו (לא אמפירית אלא הוכחות מתמטיות רציניות) לא מוגבלות רק לאימון LLMs עם RLHF.

בד״כ כאשר אנו מאמנים LLM על RLHF פונקציית הלוס שאנו מעוניינים לאפטם מורכבת מסכום של שני איברים (לפעמים מוסיפים עוד אבל אני מדבר כאן על כאלו המופיעים ברוב המאמרים על יישור (alignment) של LLMs עם שיטות RL.

האיבר הראשון אחראי על מקסום של פונקציית reward שזה מודל שמאומן לפני על דאטהסט המכיל זוגות של תשובות מועדפות יותר ומועדפות פחות(המתויג על ידי בני אדם) לסט של שאלות. מודל reward מאומן לתת ערך גבוה לתשובה טובה וערך נמוך לתשובה לא טובה לשאלה. אז האיבר הראשון מאפטם את משקלי ה-LLM המאומן כך שימקסמו את פונקציית ה-reward ובכך יגרמו ל-LLM להיות יותר מיושר (aligned)עם הציפיות שלנו (לפחות היינו רוצים להאמין בכך).

האיבר השני הינו איבר רגולריזציה השומר את המשקלים של המודל המאומן קרובים יחסית (במונחי מרחק KL בין התפלגויות הטוקנים) למשקלי המודל ההתחלתי (שלו אנו עושים פיין טיון). איבר זה נדרש כי בלעדיו המודל יעשה את מה שנקרא ״reward hacking" ובמקום להתיישר עם ציפיותנו ימקסם את reward אבל כתוצאה נקבל עוד יותר גרוע ממה שהיה (או לפחות פחות טוב ממה שניתן לקבל עם איבר רגולריזציה זה).

אולם מחברים המאמר טוענים שאיבר זה לא מספיק ולא תמיד ימנע ממשקלי המודל להתכנס למצבים לא רצוים. הסיבה לכך היא שמודל בסיס שאנו רוצים לשמור את המודל המאומן קרוב אליו מהווה בעצמו קירוב של מודל ״בטוח ומיושר עם ציפיותנו״ (בדרך כלל אומן על התשובות הרצויות). ומתברר שגם אם מודל הבסיס שלנו קרוב מספיק ל״מודל הבטוח״ והמודל שאנו מאמנים קרוב למודל הבסיס במונחי KL, עדיין לא ניתן להבטיח שהמודל המאומן יהיה קרוב מספיק ל״מודל הבטוח״ (גם במונחי KL) - כלומר אי שוויון המשולש לא מתקיים כאן. גם נאמן את מודל בסיס על יותר דאטה ויותר משאבים, עדיין נתקשה להבטיח את קרבתו של המודל המאומן ל״מודל הבטוח״

הסיבה לכך היא קצת (מבחינה קונצפטואלית) דומה לכך למה דגימת Langevin בצורתה הקלאסית (ללא רעש) לא עובדות לדאטה בעלת מימד גבוה מאוד כמו (תמונות). בגלל המימד המאוד גבוה של המרחב הסמנטי של הדאטה מודל הבסיס יגיע למקומות ש״המודל הבטוח״ לא היה מגיע בכלל ואז הוא יתקשה לתת שערוך אמין להסתברויות הטוקנים. וזה יגרום למודל המאומן להיות לא אמין באותה המידה.

המחברים קוראים למאורעות אלו (הגעה למצב שהמודל הבטוח לא היה מגיע אליו) מצבי חסר תקדים (unprecedented events) וטוען שכאשר הם קורים מודל הבסיס ייטה לתת תשובה "פשוטה מדי" ולרוב לא נכונה וכך יעשה המודל המאומן. הפשטות הזו נובעת כנראה (איליה סלוצקבר מדבר על זה רבות) בגלל ה algorithmic-information-theoretic inductive bias שעוזר לרשתות נוירונים להגיע ליכולת הכללה טובה עקב נטייתם להתכנס לפתרונות פשוטים בעלות סיבוכיות תכנותית נמוכה (שזה בעצם סיבוכיות kolmogorov) כלומר הן פעולות לפי עקרון התער של אוקם (זו ההנחה כמובן). זה גורם למודלים להפגין התנהגות פשוטה(מדי) ולא טובים במקרה שהם נתקלים במאורעות חסרי תקדים האלו. והמודל המאומן על RLHF "יורש מהם" את הפגם זה.

מאמר מאוד עמוק, דורש זמן בשביל להפנים אבל שווה קריאה בהחלט…

https://arxiv.org/abs/2410.06213