המאמר היומי של מייק: 13.08.25  
Checklists Are Better Than Reward Models For Aligning Language Model

תשכחו מכל מה שידעתם על מודלי תגמול: האם צ'קליסט פשוט הוא העתיד של עולם ה-AI?

בשנים האחרונות, פרדיגמה יחידה שלטה במאמצינו לגורם ל- LLMs להתנהג לפי ה״חוקים״: למידת חיזוק ממשוב אנושי (RLHF). בלב גישה זו נמצא מודל התגמול (RM), רשת נוירונים חזקה אך לא ניתנת לפרשנות, שאומנה לזקק את ״המשטח המורכב והמבולגן של העדפות אנושיות״ לכדי ציון סקלרי יחיד. לאחר מכן, משתמשים בציון זה כדי להנחות את ה-LLM שלנו להתנהגות "טובה". אך כל התהליך הזה נשען על הנחה שברירית: שמספר נלמד בודד יכול ללכוד באופן מהימן את האופי הרב-ממדי של ערכים אנושיים.

מאמר חדש, אינטואיטיבי ודי מבריק קורא תיגר על הנחת יסוד זו. המחברים טוענים שברדיפה אחר ציון יחיד, בנינו מערכות שהן לא רק קופסאות שחורות, אלא גם נוטות ל-reward hacking ובנוסף לא ניתנות לפרשנות. החלופה שהם מציעים אינה מודל מורכב יותר, אלא תנועה לעבר פשטות ויכולת פירוש (interpretability). על ידי שילוב של צקליסטס מובנות עם שיטת DPO, המאמר משרטט נתיב חזק, יעיל ואמין יותר ל-alignment של המודלים.

החידוש המרכזי הראשון הוא המעבר מתגמול מרומז וסקלרי לתגמול מפורש ומבוסס-וקטור. במקום לאמן מודל תגמול לפתח "תחושה" אינטואיטיבית לגבי מה שבני אדם מעדיפים, המחברים מציעים להעריך את הפלט של המודל אל מול רשימת תיוג מובנית של תכונות רצויות ומוחשיות.

דמיינו שאתם מעריכים תגובה לא באמצעות ציון בודד מ-1 עד 10, אלא אל מול רשימה של קריטריונים בינאריים או מרובי-רמות:

האם התשובה נכונה עובדתית? (כן/לא/חלקית)

האם היא נמנעת מסטריאוטיפים מזיקים? (כן/לא)

האם הטון עוזר ואינו מתנשא? (כן/לא)

האם היא מצטטת מקורות אמינים, אם רלוונטי? (כן/לא)

פירוק זה הוא המפתח. הוא הופך את המשימה המאוד מורכבת של מידול העדפות לסדרה של בעיות סיווג מוגדרות יותר וניתנות לאימות, שלעיתים קרובות מבוצעות על ידי מודל שפה אחר (כלומר LLM-as-a-judge). אך הדבר מעלה שאלה: כיצד הופכים הערכה וקטורית זו לסיגנל אימון נקי וסקלרי לעדכון המודל? כאן נכנס החידוש השני של המאמר. הצקליסט אינו משמש כפונקציית תגמול ישירה. במקום זאת, המחברים משתמשים בה כפונקציית תיוג אוטומטית ועוצמתית ליצירת זוגות העדפה עבור DPO. שיטת DPO מבצעת ידי פיין טיון של מודל השפה על זוגות של תשובות מועדפות ותשובות לא מועדפות. המאמר משתמש ברשימת התיוג כדי ליצור זוגות אלה באופן אוטומטי, ובכך לבטל את הצורך בתיוג אנושי יקר או במודל תגמול נפרד.

תהליך האימון הופך ללולאה איטרטיבית ועצמאית:

יצירה (Generate): עבור פרומפט נתון, המודל המאומן מייצר שתי תגובות מועמדות או יותר.

הערכה (Evaluate): מודל ה"שופט" מעריך כל תשובה אל מול רשימת התיוג, וקובע איזו מהן עונה טוב יותר על הקריטריונים המפורשים.

צימוד (Pair): בהתבסס על הערכה זו, התגובה העדיפה מתויגת כנבחרת (y\_w) והאחרת מתויגת כנדחית (y\_l).

פיינטיון: זוג ה-(y\_w, y\_l) שנוצר זה עתה משמש כדגימה בודדת לעדכון המודל המאומן באמצעות פונקציית הלוס של DPO.

שיטה אלגנטית זו פותרת מספר בעיות בבת אחת. היא עוקפת את הצורך לאמן מודל תגמול מונוליטי, ובמקום זאת שואבת את סיגנל ההעדפה שלה מרשימת התיוג השקופה והניתנת לעריכה. ומכיוון שהנתונים נוצרים תוך כדי תנועה, נוצרת ״תוכנית לימודים״ דינמית ומתקנת את עצמה, שניתן לכוון בזמן אמת פשוט על ידי שינוי הקריטריונים ברשימת התיוג.

הניסויים של המחברים נועדו לא רק לכבוש את טבלאות הבנצמרקים, אלא לבחון רובסטיות של השיטה. הם מראים שבעוד שיישור סטנדרטי מבוסס-RM יכול להשיג ציונים גבוהים במדדי ביצוע ספציפיים, מודלים אלה הם לעיתים קרובות שבריריים. הם ״מנצלים לרעה חוק גודהארט״, והופכים למצטיינים באופטימיזציה של הפרוקסי (ציון התגמול) על חשבון המטרה האמיתית.

לעומת זאת, מודלים שיושרו בשיטת Checklist-DPO מפגינים רובסטיות רבה יותר. מכיוון שהם מותאמים לעמוד במערך מגוון של קריטריונים מפורשים, יש להם פחות סיכוי למצוא "פריצה" יחידה ופשוטה. הם חייבים להיות טובים במספר דרכים הניתנות לאימות. המאמר מראה שמודלים אלה עמידים יותר להנחיות אדברסריאליות, פחות מתרפסים (sycophantic), ומקפידים באופן אמין יותר על מגבלות בטיחות, גם בתרחישים שחורגים מנתוני ההתפלגות המקוריים.

https://arxiv.org/abs/2507.18624

המלך הוא עירום: למה מודלי שפה נכשלים בחשיבה אלגוריתמית אמיתית