המאמר היומי של מייק: 08.09.25

Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation

מתחילים את המאה השישית :  סקירה 501

בעולם האימונים של LLMs, שבו ריצת אימון בודדת עולה יותר מבית מפואר אפילו בישראל, כל החלטה(הייפרפרמטרים) היא הימור על מיליוני דולרים. אנו מסתמכים על ניסויים קטנים, סימולציות זעירות וחסכוניות של הדבר האמיתי כדי להנחות את החלטות הללו. אנו מאמנים צי של מודלים עם מיליארדים בודדים פרמטרים כדי לחזות את התנהגותו של מודל ענק עם מאות מיליארד פרמטרים, בתקווה שהמגמות שאנו מודדים במעבדה יתקיימו גם במפעל. האמת המטרידה, עם זאת, היא שלעתים קרובות זה לא קורה. הדירוגים מתהפכים, תחזיות הסקייל (scaling) נכשלות, ואנו נותרים לתהות מדוע המבחנים (benchmarks) המהימנים שלנו הוליכו אותנו שולל.

המאמר שנסקור היום חוקר לעומק את הסוגיה הזו. חדשנותו אינה בזיהוי הבעיה, אלא באספקת מסגרת אבחון מדויקת, ניתנת לחישוב ואינטואיטיבית להפליא, המאפשרת להבין מדוע מבחנים מסוימים(בנצ'מארקים) הם מדריכים אמינים ואחרים הם אשליות סטטיסטיות. המאמר מציג טרמינולוגיה חדשה להערכת ההערכות שלנו.

החדשנות המרכזית של המאמר טמונה בכך שהמחברים מתייחסים למבחנים כאל מכשירי מדידה, כמו טלסקופ או גלאי חלקיקים. כל מכשיר טוב חייב לעשות שני דברים: להבחין בין תופעות שונות (אות) ולהפיק קריאות עקביות של אותה תופעה (רעש).

האות של מבחן, בניסוח של המאמר, הוא יכולת הבידול הטבועה בו בין מודלים באיכות משתנה. דמיינו שאתם מעריכים תריסר מודלים על משימה. אם כולם מקבלים ציון בין 90% ל-91%, למבחן יש אות נמוך. הבדלי הביצועים הולכים לאיבוד באבק הנקודות העשרוניות. לעומת זאת, מבחן בעל אות גבוה פורס את הציונים על פני טווח רחב וברור, כך שניתן לראות בבירור אילו מודלים עדיפים. המאמר מכמת במדויק את ה"פיזור" הזה כהפרש המרבי המנורמל בין ציוני שני מודלים כלשהם, מדד שהם מכנים פיזור יחסי (relative dispersion).

הרעש הוא האקראיות המתסכלת והמובנית בביצועי מודל על מבחן נתון. התובנה החדשנית והחשובה ביותר של המחברים כאן היא זיהוי של קירוב (proxy) זול ועוצמתי לאי-יציבות זו: התנודתיות בין צ'ק פוינטים סמוכים (checkpoint-to-checkpoint variability). גם בשלבי האימון האחרונים, הדיוק של מודל במבחן כמו ARC-Challenge יכול לקפוץ בפראות בצעד אימון אחד(training step). המחברים מראים(אמפירית) כי לתנודתיות זו, שקל למדוד, יש קורלציה גבוהה עם מקורות ״רעש״ יקרים יותר, כמו שינויים בסדר דאטה באימון או באתחול המשקולות. זהו "משנה משחק"; פירושו שניתן לאבחן מקור מרכזי לחוסר אמינות מבלי לאמן מספר מודלים יקרים מאפס.

המחברים מציינים כי אף אחד מהם, אות או רעש, אינו קובע לבדו את עוצמתו של מבחן. מה שקובע הוא היחס ביניהם. מבחן יכול להיות בעל אות פנטסטי (הפרדה מצוינת בין מודלים) אך להיות כה רועש עד שהדירוגים הם אקראיים למעשה מנקודת שמירה אחת לאחרת (כמו ARC-Challenge). לעומת זאת, מבחן יכול להיות יציב במיוחד ובעל רעש נמוך, אך אם אין לו אות, הוא חסר תועלת להשוואת מודלים.

יחס אות לרעש (SNR) הוא המדד שלוכד באלגנטיות את הפשרה הזו. הממצא האמפירי המרכזי של המאמר הוא קורלציה חזקה בין יחס האות לרעש של מבחן לבין "דיוק ההחלטה" (decision accuracy) שלו, הסבירות שהדירוג היחסי של מודלים בקנה מידה קטן יתקיים גם בקנה מידה גדול. זהו הגביע הקדוש: תכונה זולה וניתנת לחישוב של מבחן, החוזה את ערכו הכלכלי בתהליך הפיתוח.

מסגרת זו מאפשרת שלוש התערבויות חדשניות ועוצמתיות:

בחירת חלקים רלוונטיים במבחנים לפי SNR: מבחנים רבים הם אוספים של תת-משימות (למשל, 57 הנושאים של MMLU). המחברים מראים שלעתים קרובות ניתן ליצור הערכה אמינה יותר על ידי בחירה של תת-המשימות בעלות ה-SNR הגבוה ביותר, גם אם המבחן שנוצר מכיל פחות שאלות. עבור MMLU, יחס האות לרעש המרבי מושג עם 16 תת-המשימות המובילות בלבד. זוהי תובנה מצוינת ונוגדת-אינטואיציה: הערכה טובה יותר באמצעות הפחתה אסטרטגית.

הפחתת רעש באמצעות מיצוע: מכיוון שהתנודתיות בין נקודות שמירה היא מקור רעש עיקרי, תיקון פשוט הוא למצע את הציונים של כמה צ'ק פוינטים האחרונים במקום להסתמך על האחרונה בלבד. פעולת החלקה פשוטה זו משפרת באופן עקבי את דיוק ההחלטה ואת אמינות תחזיות חוקי הסקייל.

שינוי סוג המדידה: המאמר מספק סיבה לתעדוף סוגי מדידה מסוימים. המחברים מראים כי מעבר ממדדים בדידים כמו דיוק למדדים רציפים כמו ביטים-לכל-בייט (BPB) מגדיל באופן משמעותי את ה-SNR עבור משימות רבות, במיוחד במשימות גנרוט קשות שבהן מודלים קטנים מציגים ביצועים הקרובים לניחוש אקראי. אות הלוס הרציף רועש פחות ומופיע מוקדם יותר באימון, מה שהופך אותו למכשיר מדידה טוב יותר לחיזוי. BPB כאן הוא לוג של נראות מירבית של התשובה מנורמל באורך התשובה.

המחברים סיפקו יותר מאוסף ממצאים אלא פרדיגמה. הם הסיטו את השיח מתוכן המבחן לתכונותיו הסטטיסטיות ככלי מדידה. בכך שהעניקו לנו את שפת האות, הרעש וה-SNR, הם נתנו לקהילה כולה ארגז כלים זול, עוצמתי ומבוסס תיאורטית, לא רק כדי לבחור מבחנים טובים יותר, אלא כדי לשפר באופן פעיל את אלו שכבר יש לנו.

https://arxiv.org/abs/2508.13144

לאופטימייזר ללא בגדים: ניפוץ מיתוס ההאצה של פי 2