המאמר היומי של מייק: 25.06.25  
The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs

מאמר 🇮🇱

תפנית מעניינת מתרחשת בתקופה האחרונה בעולם של הערכת ביצועי מודלים. אנחנו כבר לא שואלים רק עד כמה המודל מצליח במבחן כלשהו, אלא שאלה מהותית יותר: האם ניתן לסמוך על מודל שפה שיחליף מתייג אנושי? זו לא שאלה שמדדים מסורתיים כמו דיוק, F1 או הסכמה בין מתייגים יכולים לענות עליה כראוי. תחת זאת, המאמר שנסקור היום מציג שיטה מבוססת סטטיסטיקה לפתרון ביעה זו. בלב המאמר עומדת קריאה להתרחק ממדדי התאמה שטחיים, ולעבור לנימוקים מבוססי השערות סטטיסטיות וניתוח עלות-תועלת.

התרומה המרכזית של המאמר היא שיטה חדשה בשם מבחן המתייג האלטרנטיבי (Alt-Test). השיטה הזו לא בודקת האם המודל מסכים עם רוב המתייגים או עומד ברף דיוק כלשהו. במקום זאת, היא שואלת שאלה עמוקה יותר: האם המודל עקבי יותר עם קבוצת המתייגים מאשר מתייג אנושי ממוצע? כך זה עובד: בכל פעם מוציאים מתייג אחד מן הקבוצה, ומשתמשים בשאר כקבוצת ייחוס. המודל והמתייג שהוצא נבדקים לפי מידת ההתאמה שלהם לקבוצה הנותרת. אם המודל עקבי יותר מהמתייג שהושמט הוא "מנצח" אותו. התהליך הזה חוזר על עצמו עבור כל מתייג.

החידוש כאן הוא שמדובר בשיטה שאינה דורשת כלל תוויות אמת (gold labels). היא גם יעילה בדגימות קטנות; אפשר להשתמש בה עם שלושה מתייגים וכמה עשרות דוגמאות בלבד. אבל אולי החשוב ביותר: היא מספקת הכרעה בינארית כלומר האם המודל יכול, סטטיסטית, להחליף את האדם? לא "כמה טוב הוא היה", אלא האם יש הצדקה להשתמש בו במקום מתייג אנושי.

כדי לאחד את תוצאות ההשוואות הללו בין המודל לכל מתייג, המאמר מציג מדד שנקרא שיעור הניצחון ω. זהו פשוט אחוז המתייגים שהמודל ניצח לפי מבחן האלטרנטיבה. למשל, אם המודל טוב יותר מ-4 מתוך 6 מתייגים, שיעור הניצחון שלו הוא שני שלישים ואם הוא טוב רק מ-2 מתוך 6, השיעור הוא שליש. המאמר קובע כלל הכרעה ברור: אם שיעור הניצחון גבוה מ-50% אז המודל טוב יותר ממתייג טיפוסי, ולכן ניתן להחליפו במודל באותו הקשר. שיטה זו לוקחת השוואות סטטיסטיות ומתרגמת אותן למדיניות תפעולית. היא גם מכירה בשונות בין מתייגים, ולא מניחה שכולם שקולים. זהו מנגנון ברור, מבוסס דאטה ואמין, שמאפשר קבלת החלטות על בסיס תצפיות ולא תחושת בטן.

אבל יש שאלה נוספת: אם אני מחליף בני אדם במודל איזה מודל עליי לבחור? כאן מציג המאמר מדד נוסף, רציף יותר, בשם הסתברות היתרון הממוצעת (ρ). הרעיון פשוט: עבור כל מתייג אנושי, בודקים מה הסיכוי שהמודל עקבי יותר עם שאר המתייגים ממנו. לאחר מכן מחשבים את ממוצע ההסתברויות האלה. הערך הסופי מתקבל כמדד: מה הסיכוי שהמודל טוב לפחות כמו מתייג אנושי אקראי. המדד הזה משמעותי מכמה סיבות:

- הוא רציף ודחוס, ולא ״קופצני״ כמו שיעור הניצחון (שיכול לקבל רק כמה ערכים בדידים).  
- הוא אינו תלוי ברף שרירותי, ולכן מתאים במיוחד להשוואה בין מודלים  
- הוא כללי ומתאים לכל סוגי המשימות: סיווג, דירוג, ואפילו הפקת טקסט חופשי.  
- הוא אינטואיטיבי וברור להסבר, גם לקהלים שאינם סטטיסטיקאים.

אבל מעבר לכך הוא מתמודד ישירות עם מה שמדדים קלאסיים מתעלמים ממנו: השונות בין בני אדם. בעוד דיוק או F1 מניחים שיש "אמת אחת", ρ בוחן עד כמה המודל מצליח ללכוד את התפלגות הדעות האנושיות. זהו מדד ראשון מסוגו שמכבד את המורכבות של שיפוט אנושי — ולא סתם מתקרב לממוצע.

אחת התוספות החדשניות והאלגנטיות ביותר במאמר היא הפרמטר אפסילון (ε) שמכיל בתוכו את העובדה שמודלים זולים, מהירים ומדרגיים יותר מבני אדם, ולכן לא חייבים להיות טובים כמוהם כדי להיות משתלמים. אפסילון מייצג את הפער המותר בביצועים כמה פחות טוב המודל יכול להיות, ועדיין להשתלם כלכלית. לדוגמה, אם ההשוואה היא למומחים יקרים נוכל להצדיק שימוש במודל גם אם הוא מעט פחות טוב. אבל אם מדובר מתייגים זולים המודל צריך להיות טוב יותר מהם בבירור.

באופן מעשי, המאמר ממליץ:

- להשוות מול מומחים עם ערך ε של 0.2 (כי הם יקרים)  
- להשוות מול מתייגים רגילים עם ערך ε של 0.1 (כי הם זולים יותר).

בכך, אפסילון הופך את המבחן האלטרנטיבי מכלי השוואתי טהור למסגרת פרגמטית לקבלת החלטות עסקיות. הוא מביא את כלכלת האנוטציה לתוך עולם ההסקה הסטטיסטית ןלא רק האם המודל טוב, אלא האם הוא מספיק טוב לאור מה שהוא חוסך לנו.

בקיצור מאמר לא רגיל בנוסף המודרני ובקטע טוב…

https://arxiv.org/abs/2501.10970