המאמר היומי של מייק: 18.04.25  
Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM’s Reasoning Capability

מאמר די מעניין העוסק בשיפור יכולות הנמקה של מודלי שפה בשאלות שיש להם תשובות חד משמעיות (כגון שאלות מתמטיות ושאלות קוד שניתן לבחון את נכונות הפתרון באמצעות סט מקיף של טסטים). המאמר מגדיר מושג טוקן קריטי (critical token) שהוא למעשה מהווה סוג של סימן האם המודל הולך לתת תשובה נכונה או לא נכונה לשאלה.

המחברים שמו לב כי שבתוך מסלולי הנמקה(reasoning) שגויים, ישנם טוקנים שהם כמעט בוודאות מובילים לתוצאות שגויות. טוקנים אלה משבשים את הרצף הלוגי, מעוותים קשרים או מכניסים שגיאות חישוביות, וכך משפיעים באופן משמעותי על התוצאה הסופית. בשונה מטוקנים אחרים שעשויים להשפיע בצורה לא משמעותית על תהליך האינפרנס, "הטוקנים הקריטיים" האלו מהוות סוג של נקודות כשל. זיהוי הטוקנים הללו הוא חיוני, משום שלעיתים קרובות הימנעות מהם או תיקונם יכולה להוביל לתוצאה נכונה – גם בתוך מסלול הסקה שגוי.

המאמר מציעה שיטה לזיהוי של טוקנים קריטיים. טוקן מזוהה כקריטי עם כל מסלולי ההנמקה המתחילים ממנו מסתיימים בתשובה שגויה ועבור כל הטוקנים שבאים 95% מהמסלולים המתחילים מהם מסתיימים בתשובה לא נכונה. שימו לב שיש טוקנים המופיעים בטקסט במיקומים שהם אחרי הטוקן הקריטי שלא כל מסלולי הנמקה שלהם מכילים את הטוקן הקריטי, כך לא מן הנמנע שיש בינם מסלולים המסתיימים בתשובה נכונה. המחברים ביצועו כמה בדיקות כדי לוודא שהטוקנים שזוהו בצורה באמת טוקנים קריטים.

לאחר מכן המאמר מפתח שיטת RLHF ליישור מודל שפה שבמרכזה מזעור של הנראות של הטוקנים הקריטיים (כי הם מובילים לשגיאות). בשביל כך המאמר מציע לאמן שני מודלים (עם פיינטיון) - אחד שמגנרט תשבות נכונות והשני מגנרט תשובות לא נכונות (שמעתם נכון).

לאחר מכם המחברים מנסחים דרך לשערוך הנראות של האם הטוקן הוא קריטי בהינתן הפרומפט וטוקני התשובה לפניו. הנוסחה היא הפרש ממשוקל של נראויות הטוקנים (מותנים) מהמודל של התשובות הנכונות לבין זה של המודל של התשובות השגויות. שערוך זה מקבל ערך נמוך עבור התשובה הנכונה וערך גבוה עבור התשובה הלא נכונה. בשלב האחרון המודל עובר פיינטיון עם DPO שזה קיצור של Direct Preference Optimization כאשר על הזוגות של שאלות עם התשובות הנכונות והשגויות. כדי למזער את הסיכוי להופעת טוקן קריטי המאמר משנה את האיבר המכיל נראות של תשובה שגויה בנוסחה העיקרית של DPO על ידי הכפלה על הנראות השלילית של טוקן להיות קריטי. שימו לב שמכיוון שההתחשבות בנראות מתרחשת ברמה של טוקן ה-DPO במאמר עובר להיות token-level ולא sample-level כמו במאמר המקורי.

https://arxiv.org/abs/2411.19943