⚡️🚀המאמר היומי של מייק -14.10.24: ⚡️🚀  
LLMS KNOW MORE THAN THEY SHOW: ON THE IN-TRINSIC REPRESENTATION OF LLM HALLUCINATIONS

מאמר כחול-לבן זה מציג חקירה מקיפה של דפוסי השגיאות של LLMs והקשר שלהם עם הייצוגים הפנימיים של המודל. המחברים מבצעים סדרת ניסויים כדי לנתח כיצד LLMs מקודדים מידע על התשובה הנכונה וחוקרים את טבע השגיאות שהם מייצרים.

המחברים חקרו את הנושאים הבאים:  
  
שיפור זיהוי שגיאות:

המחברים גילו כי ניתן להגיד האם המודל ייתן תשובה נכונה או לא מהסתכלות בטוקנים ספציפיים המכילים "תשובה מדויקת" בתוך פלט המודל. כלומר עבור השאלה ״מה עיר הבירה של צרפת״ האינדיקציה האם המודל נותן התשובה הנכונה ניתן לגזור מייצוגי הטוקנים המופק על ידי שכבות מסוימות של המודל. על ידי התמקדות בטוקנים אלה, המחברים הצליחו לשפר משמעותית את דיוק זיהוי השגיאות במגוון משימות ומודלים.  
  
הכללה בין משימות:

המחקר בוחן האם יכולות זיהוי השגיאות ניתן להכללה בין משימות וסוגי דאטה שונים. התוצאות מראות הכללה מוגבלת, עם הצלחה מסוימת רק בין משימות הדורשות מיומנויות דומות (למשל, אחזור עובדתי או היסק שכל ישר). זה מרמז על כך של-LLMs יש מספר מנגנוני אמיתות "ספציפיים למיומנות" ולא מנגנון אוניברסלי אחד.

טקסונומיה של שגיאות:

המחברים מציעים טקסונומיה של שגיאות LLM המבוססת על התפלגות התשובות במספר דגימות. הם מזהים מספר סוגי שגיאות, כולל תשובות נכונות/שגויות באופן עקבי, תשובות נכונות/לא נכונות לסירוגין ומקרים עם תשובות מגוונות רבות. המחברים מדגימים שניתן לחזות סוגי שגיאות אלה מהייצוגים הפנימיים של המודל.

פער בין ייצוג פנימי להתנהגות חיצונית:

המחברים מראים פער זה באמצעות מערך ניסויי בו הם מייצרים מספר תשובות לכל שאלה ומשתמשים במודל מאומן(באמצעות probing) לבחירת התשובה הטובה ביותר על סמך ייצוגים פנימיים. הם הבחינו בשיפורים משמעותיים בדיוק עבור סוגי שגיאות מסוימים, במיוחד אלה בהם המודל אינו מראה העדפה ברורה לתשובה הנכונה בפלטים הרגילים שלו. לדוגמה, בקטגורית שגיאות "שגוי באופן עקבי אך מייצר את התשובה הנכונה לפחות פעם אחת", שיטת הבחירה מבוססת מודל הסיווג השיגה שיפורים של עד 40% בדיוק בהשוואה למצב הרגיל.

ממצא זה מרמז על כך שה-LLMs לעתים קרובות "יודעים" את התשובה הנכונה ברמה מסוימת, אך ידע זה לא תמיד משתקף בתהליך יצירת הפלט שלהם. פער זה מעלה שאלות חשובות לגבי טבע ייצוג הידע ב-LLMs והמנגנונים השולטים בתהליך יצירת הפלט שלהם. המחברים מציעים כי ממצא זה עשוי לשמש לפיתוח אסטרטגיות חדשות לשיפור דיוק ה-LLM, אולי על ידי שינוי תהליך יצירת הפלט כך שלוקח בחשבון גם את הייצוגים הפנימיים.

https://arxiv.org/abs/2410.02707