⚡️🚀המאמר היומי של מייק 13.09.24: ⚡️🚀

LLMs Will Always Hallucinate, We Need to Live With This

טוב, המאמר הזה הוא פשוט קליקבייט לדעתי ואז גיליתי שם משפט גדל אז בכלל. הוא מציג ניתוח מעמיק של הזיות (hallucinations) ב-LLMs וטוען כי הזיות אלו הן תכונה אינהרנטית בלתי נמנעת של המבנה המתמטי/ארכיטקטוני ואופן החישובי שלהם (אולי o1 החדש יאתגר את זה טיפה).

כמה נקודות עיקריות מהמאמר:

הזיות כבלתי נמנעות: הזיות אינן רק טעויות אלא תוצאה בלתי נמנעת של הארכיטקטורה וההיגיון השולטים במודלים גדולים לשפה. הן נוצרות כאשר המודלים מנסים להשלים פערים בידע או לייצר מידע סביר אך שגוי על סמך נתונים חסרים או מעורפלים.

חוסר שלמות של נתוני האימון: המאמר מדגיש כי אף מאגר נתונים אינו שלם ב-100%, ולכן LLMs תמיד יתקלו במצבים שבהם עליהם להסיק או להמציא מידע שלא קיים במאגר הנתונים (המוחבא במשקלים שלו או במערכת נתונים חיצונית).

4 סוגים עיקריים של הזיות:

אי דיוק עובדתי: המודל עלול לגנרט מידע עובדתי שגוי בשל ״אופן שליפה שגוי״ של מידע ממאגרי הידע שלו.

אי הבנה: המודל נכשל בהבנת קלט המשתמש, ונותן תשובות שגויות.

מחט בערימת שחת (needle in a haystack): קושי בשליפת מידע ספציפי ממאגר נתונים (במשקלים שלו או במערכת נתונים חיצונית), מה שלעתים מוביל למידע מעורב או חלקי.

המצאות: LLMs לעיתים ממציאים מידע כאשר הקלט אינו מוכר להם מהטריין סט ולא תואם לשום עובדה ידועה במאגר הנתונים שלהם.

״בלתי מוכרעות״: המחברים משתמשים במשפטי אי שלמות של גדל ובתיאוריית חישוביות, ומדגימים שבעיות מסוימות, כגון שליפת מידע עובדתי מדויק ( וסיווג כוונת המשתמש (intent classification), אינן ניתנות להכרעה. המשמעות היא שאין אלגוריתם שיכול למנוע לחלוטין הזיות.

LLMs לא מסוגלים לנבא מתי הם ייעצרו: המחברים טוענים כי LLMs לא מסוגלים לחזות מתי ייעצר הגנרוט (מזכיר הבעיה הידועה של עצירת מכונה בתיאוריה חישובית). הם טוענים שנובע מכך כי המודלים אלה אינם מסוגלים לשלוט או לצפות במדויק איזה תוכן הם ייצרו, מה שמעלה סיכוי להזיות.

הוכחה שהזיות אינן ניתנות לביטול: המאמר מראה (יש הוכחה) שגם כוונון מושלם או מנגנוני בדיקת עובדות לא יכולים לבטל לחלוטין הזיות. זאת משום שמאגר הנתונים תמיד יהיה חסר או בלתי מספיק, ומודלים גדולים לשפה חייבים לייצר פלט שאינו ניתן לאימות או סותר.

השפעה של RAG: למרות שטכניקות כמו הפקת מידע מוגברת נועדו לשפר את הדיוק העובדתי באמצעות שליפת מידע חיצוני, הן עדיין מסתמכות על פונקציות שליפה לא מושלמות, מה שמוביל לתוצאות חלקיות או מעורבות.

תפקיד קידוד מיקומי (positional encoding או PE): המאמר נוגע בטכניקות PE מתקדמות כמו RoPE וכיצד הן משפרות את ביצועי המודלים באמצעות שילוב מיקומים מוחלטים ויחסיים. עם זאת, טכניקות אלו עדיין לא פותרות את בעיית ההזיות.

הזיות מבניות: המחברים מציגים את המושג "הזיות מבניות", ומדגישים שהן תוצאה בלתי נמנעת של הארכיטקטורה של LLMs ולכן אינן ניתנות למניעה, גם לא באמצעות שיפורים באימון או כוונון.

השוואה למודלים אחרים: המאמר משווה בין מודלים לשפה למודלים אחרים כמו ממבה, KANs אך מסיק שהמגבלות המובילות להזיות קיימות בכל הארכיטקטורות.

מכונת טיורינג ו-LLMs: מודלי שפה מוצגים כשווים למכונות טיורינג אוניברסליות, מה שאומר שהם יורשים את אותן מגבלות חישוביות, כולל בעיות בלתי-מוכרעות כמו בעצירה.

השלכות לעיצוב עתידי של LLMs: המאמר מציע שהפיתוחים העתידיים של LLMs צריכים להתמקד בניהול והפחתת הזיות במקום לנסות לבטל אותן, שכן הדבר בלתי אפשרי מתמטית וחישובית.

https://arxiv.org/abs/2409.05746