⚡️🚀המאמר היומי של מייק -12.10.24: ⚡️🚀  
GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

האם מודלי שפה גדולים מסוגלים לעשות ריזונינג? השאלה הזו מעסיקה חוקרים רבים לאור יכולות די מרשימות שמודלי שפה מפגינים בפתרון שאלות לא פשוטות (אבל רק בתנאים מסוימים 🙂). המאמר בוחן את יכולות החשיבה המתמטית(שזה תת-יכולת של ריזונינג כללי) של LLMs ומציג את GSM-Symbolic, בנצ'מרק חדש לבחינת יכולות אלו שהם פיתחו.

החוקרים מצאו שביצועי LLMs(נבחן מגוון רחב של מודלים: Gemma, Phi, Mistral, Llama3, GPT-4 ו-o1) משתנים באופן משמעותי כאשר משנים מעט את השאלות המתמטיות, מה שמעלה ספקות לגבי אמינות המדדים הקיימים. הביצועים של רוב המודלים יורדים כאשר עוברים מ-GSM8K המקורי ל-GSM-Symbolic, מה שמרמז על אפשרות של זיהום דאטה (contamination) במהלך האימון (כתבתי על זה לא מעט).

בנוסף המודלים מראים רגישות גבוהה יותר לשינויים במספרים מאשר לשינויים בשמות עצם, מה שמעיד על חוסר יציבות ביכולות החשיבה שלהם. ככל שמספר המשפטים בשאלה עולה, הביצועים יורדים והשונות בביצועים עולה, מה שמצביע על קושי בטיפול בשאלות מורכבות יותר.

החוקרים יצרו בנצ'מארק GSM-NoOp, שבו נוספו משפטים לא רלוונטיים לשאלות, וגילו ירידה דרמטית בביצועים של כל המודלים. אפילו כאשר ניתנו למודלים דוגמאות של אותה שאלה או שאלות דומות, הם התקשו להתגבר על האתגרים של GSM-NoOp.

המחקר מצא שאימון נוסף על משימות קלות יותר וגם הגדלת כמות דאטה לאימון לא שיפרו את הביצועים במשימות מורכבות יותר.

קצת מנחם שלפחות מודלים חדשים יותר, כמו o1-preview ו-o1-mini, הראו ביצועים חזקים יותר, אך עדיין סבלו מהמגבלות שזוהו במחקר

הממצאים מעלים ספקות לגבי היכולת האמיתית של LLMs לבצע חשיבה מתמטית פורמלית. נראה כי המודלים מסתמכים יותר על התאמת תבניות מאשר על חשיבה לוגית אמיתית. המחקר מדגיש את הצורך בשיטות הערכה אמינות יותר ובמחקר נוסף על יכולות החשיבה של מודלי שפה גדולים.

https://arxiv.org/abs/2410.05229