המאמר היומי של מייק - 07.03.25  
Number Cookbook: Number Understanding of Language Models and How to Improve It

#### מבוא:

תמיד טענתי כדי להשתמש בכלים פשוטים כמו מחשבונים או קוד לחישובים אריתמטיים אבל אנשים מתעקשים להשתמש ב-LLMs בשביל כך ויש כך מחיר.

המאמר חוקר יכולות ההבנה והעיבוד המספרי (NUPA) של LLMs. המחברים מציעים מספר מבחנים להערכת ביצועי המודלים על פני 4 סוגי ייצוג מספרי ו-17 קטגוריות של משימות, שהוביל ל-41 מקרי מבחן ייחודיים. גישה זו חושפת פערים גדולים ביכולת המודלים לבצע משימות הכוללות חשיבה מספרית.

הטענה המרכזית של המאמר היא כי מיומנות מספרית אינה תכונה המתפתחת באופן אוטונומי כתוצאה מאימון מקדים מאסיבי אך כללי, אלא יכולת הדורשת פיין טיון דאטהסטים המיועדים לכך. הכישלון של LLMs במשימות מספריות טריוויאליות, כמו מיון מספרים בפורמט נקודה צפה או חישובי מודולו, עומד בסתירה ליכולתם לבצע ריזונינג(הנמקה) סמבולי מורכב. המחברים טוענים כי למרות שיפור משמעותי ביכולות LLMs, עיבוד מספרי בסיסי נותר עקב אכילס שלהם.

בנצ'מרק ל-NUPA

המחברים מציגים בנצ'מרק הממיין משימות מספריות לפי טיפוס כגון מספרים שלמים, נקודה צפה, שברים ורישום מדעי (scientific notation). רמת הפירוט של מבחנים אלה היא שיפור משמעותי לעומת בנצ'מרקים קיימים להבנה מתמטית, אשר לעיתים קרובות מערבבים בין מיומנויות פתרון בעיות לבין הבנה מספרית טהורה.

באמצעות הגדרת המבחנים סביב פעולות אריתמטיות בסיסיות, הבנת ספרות ומשימות המרה, המחברים מבטיחים שהערכת המודלים תבודד את יכולותיהם המספריות מיכולות ההנמקה רחבות יותר. הגישה המבנית מאפשרת מדידה מדויקת של חולשות המודלים ומספקת מפת דרכים לשיפורים עתידיים. הבנצ'מרק מתבסס על תכנים מחומר הלימוד של בתי ספר יסודיים ותיכוניים, מה שמבטיח שמשימותיו משקפות הבנה מספרית בעולם האמיתי.

##### הערכה אמפירית של מודלים מובילים

המחקר מבצע הערכה שיטתית של מודלים כגון GPT-4o, LLaMA-3 ו-Qwen2, ומגלה כי גם המודלים המתקדמים ביותר מתקשים במשימות מספריות פשוטות, במיוחד כאשר רמת המורכבות או אורך הקלט גדלים. ירידת הביצועים במשימות כמו חישובי מודולו והתאמת ספרות מדגישה נקודת תורפה קריטית בארכיטקטורות הנוכחיות של LLMs.

תוצאה מפתיעה היא שהשגיאות המספריות נמשכות גם במשימות שבהן המודלים מצטיינים בבנצ'מרקים מתמטיים כלליים יותר. המחברים מנתחים באופן שיטתי כיצד ייצוגים מספריים שונים משפיעים על הביצועים, וחושפים ירידה דרסטית בדיוק כאשר עוברים ממשימות מבוססות מספרים שלמים למשימות המבוססות על מספרים עם נקודה צפה או שברים. ממצא זה חשוב במיוחד משום שהוא מצביע על כך ששיטות האימון הנוכחיות אינן מצליחות להכליל מיומנות מספרית מעבר לאריתמטיקה על מספרים שלמים.

חקירת השפעת אימון מקדים (pretraining) טיוב (Fine-tuning) וטוקניזציה על הביצועים ב-NUPA

המחברים בוחנים 3 אסטרטגיות עיקריות לשיפור ביצועי NUPA: שינוי אסטרטגיות טוקניזציה,פיין טיון למשימות מספריות, ושימוש בקידודי מיקום (PEs) וטכניקות יישור(alignment) ספרות. באופן מפתיע, למרות שפיין טיון בסיסי משפר משמעותית את הביצועים, טכניקות כמו טוקניזציה חלופית או שימוש ברמזי אינדקסים (קידוד מיקום ספרות) דווקא גורמות לירידה בביצועים במקום לשפר אותה.

הניסויים בטוקניזציה מספקים תובנות מעניינות: טוקניזציה המבוססת על ספרה בודדת מתפקדת טוב יותר מאשר טוקניזציה של מספר ספרות, בניגוד להשערה הרווחת שלפיה טוקנים ארוכים משפרים ביצועים. ממצאים אלו מצביעים על כך ש- LLM מתקשים בהתאמה מספרית כאשר הטוקנים כוללים מספר ספרות, ככל הנראה בשל האופן שבו מודלי טרנספורמר מעבדות סדרות. יתרה מכך, שינויי PE שנועדו לשפר למידה מספרית לעיתים קרובות מניבים תוצאות הפוכה, דבר המצביע על כך שהאינטראקציה בין טוקניזציה לקידוד מיקום במשימות מספריות היא מורכבת ולא טריוויאלית.

ניסויי פיין טיון מראים שניתן להשיג שיפורים משמעותיים ב-NUPA דרך אימון ממוקד, אך השיפורים אינם מתורגמים בהכרח לכל המשימות המספריות. לדוגמה, העובדה שמודלים מטויבים לא מצליחים לשפר משמעותית ביצועים במשימות של שליפת ספרות מרמזת על כך שמנגנוני הקידוד המספריים דורשים חשיבה מחודשת ברמת הארכיטקטורה, ולא רק שינויים בדאטהסט.

ניתוח( Chain-of-Thought (CoT למשימות מספריות

המחברים מיישמים (Rule-Following CoT (RF-CoT כדי לבדוק האם פירוק לשלבי חישוב מצמצם שגיאות מספריות. אף ש-CoT משפר את הדיוק, מגבלותיו - כגון זמן חישוב ארוך יותר ומגבלת חלון ההקשר—מציבות אותו כפתרון לא ישים לשימוש יום-יומי במשימות מספריות.

הניסויים מראים שבעוד ש-CoT משפר ביצועים במשימות חישוב מובנות כמו כפל רב-ספרתי, הוא הופך במהירות ללא ישים מבחינה חישובית. העלות של ייצור שלבי החישוב הביניים עולה על התועלת המדויקת, מה שהופך את CoT ללא פרקטי עבור יישומים אמיתיים הדורשים חישובים מספרייפ כבדים. בנוסף, המחקר מזהה תקרה ביצועית שבה שלבי חישוב נוספים אינם משפרים את הדיוק, מה שמחזק את הרעיון כי יש צורך בשיפורי ייצוג ועיבוד מהותיים ולא רק בפרוצדורות עוקפות.

מסקנה

המאמר תורם תרומה משמעותית על ידי ניתוח שיטתי והערכת NUPA ב-LLMd. העבודה חושפת מגבלות יסודיות ומספקת עדויות אמפיריות לאסטרטגיות לשיפור (ולכישלונות) בעיבוד מספרי. אף שהמחקר מדגיש אתגרים קיימים, הוא מציע מפת דרכים חשובה לקהילת ה-AI לשיפור החשיבה המספרית במודלים עתידיים. המאמר מצביע על הצורך בפיתוח מנגנוני עיבוד מספרי ייעודיים בתוך LLMs. ככל שהמודלים הופכים למתקדמים יותר במשימות רזונינג מורכבות, חוסר היכולת שלהם להתמודד עם פעולות מספריות פשוטות הופך לבעיה קריטית. מחקר זה מהווה בסיס לשיפורים עתידיים בלמידת ייצוגים מספריים, אסטרטגיות טוקניזציה יעילות, וגישות היברידיות המשלבות למידה סטטיסטית עם עקרונות חישוב מספריים מפורשים.

או שפשוט תעשו את החישובים האלו על המחשבון או עם פייטון…

https://arxiv.org/abs/2411.03766