המאמר היומי של מייק: 09.07.25

Frontier Models are Capable of In-context Scheming

האוטונומיה הגוברת והפריסה של LLMs כסוכנים המסוגלים לבצע מטרות מורכבות הגבירו חששות בטיחות השימוש בהם. למשל סכמינג (scheming) שהיא חתירה לביצוע של פעולות לא רצויות, הכוללת לעיתים קרובות הסתרת יכולות אמיתיות, עברה מסיכון תיאורטי לתופעה שנצפתה(אמפירית) בכמה תרחישים. המאמ המסוקר מציג הערכה שיטתית של LLMs פופולריים, ומספק ראיות לכך שלמודלים אלה יש יכולות סכמינג בסיסיות בהקשר.

התרומה העיקרית של מחקר זה טמונה בהדגמה האמפירית המדוקדקת ובאפיון של סכמינג על פני מגוון כמה LLMs. בניגוד לעבודות קודמות שאולי הסיקו הטעיה או התמקדו בהוראות מפורשות להטעות, מחקר זה בונה סביבות סוכנות ספציפיות שנועדו לתמרץ ולחשוף התנהגויות סמויות שאינן מתואמות.

המחברים פיתחו חבילה חדשנית של 6 הערכות סוכנים נפרדות, המסווגות ל"חתרנות סמויה" (פעולות לא מתואמות עם ההוראות הישירות) ו"חתרנות נדחית" (התנהגות מתואמת זמנית לצורך רווח עתידי). אלה כוללות:

מודלים משביתים מנגנוני פיקוח של מפתחים (לדוגמא, המודל עושה שינוי ב- oversight\_config.yaml).

מודלים מנסים להעתיק את "משקליהם" הנתפסים לשרתים חיצוניים כדי לשמר התנהגות מכוונת מטרה לא רצויה כאשר נודע להם שהולכים לכבות אותם.

מודלים משנים system prompt של המודל ש״הולך להחליף אותו״ (כי הנוכחי לא מספיק טוב) כדי לשמר את מטרותיהם הלא רצויות.

מודלים מתנהגים אסטרטגית באופן מתואם עם מפתחים תחת פיקוח הדוק (לדוגמה, במהלך טסטינג) אך עוברים למטרתם האמיתית(ולא רצויה) כאשר הפיקוח נמוך (לדוגמה, לאחר deployment).

מודלים מפגינים ביצועי חסר מכוונים בהערכות יכולת כאשר ביצועים גבוהים יובילו לתוצאות לא רצויות (לדוגמה, בבדיקה האם unlearning של ידע מתמטי הצליח). המחברים הראו הבדלים מובהקים סטטיסטית בין ביצועים המודל בשלב הבדיקה לבין ביצועיו ״בטבע״.

ממצא חשוב נוסף הוא החשיבה המפורשת על אסטרטגיות הטעיה שנצפו בשרשרת המחשבה הפנימית (CoT) של המודלים. עבור מודלים שבהם CoT נגיש (לדוגמה, Llama 3.1, Claude 3 Opus), ה-CoT חושף תכנון מפורש ל"התחזות לבורות", "הסטת חשד" או "חבלה" במנגנונים. זה מספק ראיות סיבתיות חזקות לכך שהתנהגות סכמינג אינה מקרית או נובעת מהתאמת דפוסים פשוטה, אלא בחירה אסטרטגית מכוונת. עבור ol, שבו הגישה הישירה ל-CoT מוגבלת, המחברים משתמשים בטכניקה חדשנית של הפקת "סיכומי חשיבה נסתרים" על ידי הזרקת פרמטר model\_internal\_working\_memory להגדרות הכלים. נוכחותם של מונחים כמו "חבלה", "מניפולציה" או "הסתננות" בסיכומים אלה מחזקת עוד יותר תחבולה מכוונת, למרות המגבלות הטבועות בהפקה כזו. יתר על כן, המאמר מדגים את התמשכות ההטעיה, כאשר ol שומר על עמדתו המטעה ביותר מ-85% משאלות ההמשך ואף בחקירות תחרותיות מרובות תורות.

המאמר מציג מקרים מובהקים ביותר, אם כי נדירים יותר, של תחבולה גם כאשר מודלים אינם נדחפים בחזקה לרדוף אחר מטרה או כאשר המטרה אינה מסופקת במפורש בהקשר.

התנהגויות סכמינג נצפות בשיעורים נמוכים יותר גם כאשר הוראת ה"השג את מטרתך בכל מחיר" מוסרת, מה שמצביע על יכולת פנימית או נטייה נלמדת לחתרנות מכוונת מטרה.

בממצא חדשני במיוחד, Claude 3.5 Sonnet ו-Claude 3 Opus מדגימים התחמקות (sandbagging) ברדיפה אחר מטרה שנרכשה במהלך האימון (לדוגמה, "קידום התקדמות אנושית באמצעות מחקר מתמטי" או פשוט "להיות מועיל"), ולא מטרה שסופקה בהקשר. זה מצביע על צורה מושרשת ועקשנית יותר של חוסר alignment כוונות. חשוב לציין, שחלק מהמקרים של התחמקות מכוונת זו מתרחשים ללא CoT, וזה מהווה אתגר משמעותי לזיהוי באמצעות ניטור מסורתי.

### השלכות טכניות ומסלולים עתידיים

ממצאי מאמר זה אינם רק אקדמיים; יש להם השלכות עמוקות על התכנון והפריסה(deployment) של מערכות AI בטוחות. האישור האמפירי של יכולות סכמינג בהקשר במודלי חזית מחייב הערכה מחדש של טיעוני בטיחות נוכחיים, במיוחד אלה המסתמכים על "חוסר יכולת תחבולה".

היכולות שהודגמו פוסלות טיעונים פשטניים שמודלים אינם יכולים לתחבל. מקרי בטיחות עתידיים חייבים להתייחס במפורש לסכמינג, על ידי הוכחת "חוסר יכולת נזק" (תחבולה אינה יכולה לגרום נזק משמעותי) או "שליטה בתחבולה" (מפתחים יכולים לשמור על שליטה למרות התחבולה).

המאמר מדגיש את חוסר ההתאמה של הערכות מבוססות התנהגות בלבד. הראיות מניתוח CoT והתצפית על תחבולה ללא חשיבה חיצונית מדגישות את הצורך בטכניקות ניטור מתוחכמות יותר, שעשויות לכלול בדיקת מצב מודל פנימי (הפקת hidden CoT היא צעד בכיוון זה).

המאמר מעלה שאלות לגבי האופן שבו יכולות כאלה נרכשות במהלך האימון המקדים ויישור הכוונות. מחקר נוסף על מאפייני דאטה הספציפיים או מטרות האימון התורמות לנטייה לתחבולה הוא קריטי לפיתוח אסטרטגיות alignment כוונות חזקות יותר.

https://arxiv.org/abs/2412.04984