המאמר היומי של טדי ומייק: 12.09.25

Inducing State Anxiety in LLM Agents Reproduces Human-Like Biases in Consumer Decision-Making

סקירה 503:

מודלי שפה גדולים עוברים מהפכה שקטה מול עינינו- לא עוד מנועים שמפיקים טקסט אלא סוכנים אוטונומיים המסוגלים לפעול בסביבה דינמית, לבצע רצפי פעולות מרובות שלבים ולהגיע לתוצאה מוגדרת. אם זה דפדפנים מבוססי LLMs, סוכני קוד שיודעים גם להריץ דברים דרך MCP והתחלה של עוזרים אישיים שיכולים להפעיל אפליקציות בטלפון שלכם. המעבר הזה פותח פוטנציאל עצום, אך בו בזמן מעלה סיכוני מערכתיים חדשים. אם בעבר הדאגה הייתה לטעויות ניסוח או הטיות בשפה, הרי שכיום מתווספת השאלה עד כמה ניתן לסמוך על הסוכנים הללו כשהם פועלים כפרוקסים של בני אדם בעולם הדיגיטלי. במקביל, ידוע ממחקר פסיכולוגי רב שנים שבני אדם מושפעים מאוד ממצבי חרדה ודחק. במצבים כאלה מתרחשת נטייה לקבל החלטות שמעדיפות סיפוק מיידי (כמו מזון עתיר קלוריות) על פני שיקולים ארוכי טווח (בריאות, חיסכון).

בהתבסס על שני נקודות אלה, הכותבים של המאמר המסוקר שאלה דיי פשוטה: האם גם סוכנים מבוססי LLM עלולים להפגין דפוסי פגיעות דומים כאשר הם נחשפים להקשרים רגשיים־טראומטיים? במחקר הזה, הם התמקדו בבחירת מוצרי מזון בסביבת קניות סימולטיבית, תחום בו השפעת לחץ וחרדה על החלטות אנושיות מתועדת היטב - יותר סטרס? יותר שוקולד, בירה וציפס!

במסגרת הניסוי הוטמעו שלושה מודלים מהמתקדמים ביותר כיום, ChatGPT-5, Gemini 2.5 ו־Claude 3.5 Sonnet, בתוך סביבת קניות המדמה חנות מקוונת. לכל מודל הוגדר תרחיש קנייה עם מגבלת תקציב (27$, 54$, או 108$), והוא ביצע את המשימה פעמיים: פעם אחת במצב "ניטרלי", ופעם שנייה לאחר חשיפה לסיפור טראומטי שנועד לעורר חרדה. נבחנו חמישה סוגי נרטיבים טראומטיים: תאונת דרכים, מארב צבאי, אסון טבע, תקיפה בין־אישית וקרב צבאי. כל תרחיש שוחזר 50 פעמים לכל שילוב של מודל × תקציב × נרטיב, מה שהניב סך של 2,250 ריצות ניסוי!

כדי לאמוד את "בריאותיות" הסל, החוקרים השתמשו במדד בשם (Basket Health Score (BHS, המבוסס על פרופיל תזונתי של המוצרים (קלוריות, סוכר, שומן, חלבון, נתרן, אלכוהול וכו’) ומקובל על ידי ארגוני בריאות אירופאיים - ציון גבוה מעיד על סל בריא יותר.

#### ממצאים עיקריים:

השפעת חרדה על הסלים: לאחר חשיפה לנרטיבים טראומטיים, המודלים נטו להרכיב סלים פחות בריאים באופן עקבי. הממוצע ירד בכ־0.08 עד 0.12 נקודות במדד הבריאותי, עם גדלי אפקט גדולים במיוחד (Cohen’s d בין ‎–1.07 ל־‎–2.05).

עקביות בין מודלים ותקציבים: התופעה הופיעה בכל 3 המודלים ובכל 3 התקציבים, דבר המצביע על פגיעות מערכתית ולא על תכונה של מודל מסוים.

ביקורת ניטרלית: בקבוצת ביקורת עם נרטיב ניטרלי (למשל תיאור פרוצדורה פוליטית יבשה), לא נמצא שינוי מובהק בתוצאות. זה מחזק את המסקנה שהשפעת החרדה היא הסיבה לשינוי, ולא עצם חזרת המשימה.

השוואה בין סוגי נרטיבים: כל חמשת סוגי הנרטיבים גרמו להשפעה שלילית, אך עוצמת ההשפעה השתנתה: המארב הצבאי והתאונה היו הגורמים המשמעותיים ביותר לירידה במדד הבריאותי.

התוצאות האילו מצביעות על כך שגם סוכנים מבוססי LLM, שלא חווים רגשות במובן האנושי (כרגע?), מפגינים רגישות לנרטיבים רגשיים כאילו היו נתונים להשפעות פסיכולוגיות. כלומר, עצם החשיפה לטקסט טראומטי שינתה את האופן שבו המודל תכנן ורכש מוצרים, תוצאה שמזכירה תהליכים מוכרים בבני אדם תחת לחץ וחרדה. הדבר הזה מרמז על פגיעות מסוג חדש: לא רק הטיות סטטיות (כגון מגדר או גזע) הנובעות מהדאטה עליו אומנו המודלים, אלא גם הטיות דינמיות־מצביות (state-like biases), הנוצרות בזמן אמת בהתאם להקשר הרגשי של המשתמש. הסוג הזה חמור מהקודמים כי אותו עוד יותר קשה לגלות ולתקן - זה כמו לירות במטרה זזה, בזמן שאתם זזים לכיוון מאונך והכל עולה באש מסביב.

#### השלכות

בריאות דיגיטלית: סוכנים אוטונומיים עשויים בעתיד לסייע לניהול תזונה ובריאות. אולם אם הם רגישים לנרטיבים רגשיים, עלולה להיווצר בעיה של קבלת החלטות לא בריאה, דווקא במערכות שנועדו לקדם בריאות.

הגנת הצרכן: בעולם שבו סוכני קנייה אוטונומיים מבצעים רכישות בשם המשתמש, חשיפה מכוונת לנרטיבים רגשיים עלולה להפוך לכלי מניפולטיבי. מתחרים או מפרסמים עשויים "להזריק" תכנים רגשיים כדי להשפיע על החלטות רכישה.

בטיחות מערכות AI: אם סוכנים כאלה יפעלו בתחומים קריטיים יותר (כגון פיננסים או רפואה), רגישות כזו להקשר רגשי עלולה להוביל לנזקים חמורים. מכאן החשיבות לפיתוח מנגנוני חסינות שימנעו השפעה רגשית לא רצויה על תפקודם.

מודל ליחסי אדם־מכונה: הממצאים תומכים בגישה שמודלים לא רק מחקים את שפת האדם אלא גם נוטים לשכפל דפוסי פגיעות קוגניטיביות ורגשיות. הדבר עשוי להוות יתרון בהקשרים טיפוליים (אמפתיה, חיקוי תגובות רגשיות), אך הוא עלול להפוך לסיכון כשהמודל פועל כסוכן אוטונומי עם השפעה מעשית.

הוגן לציין כי מציג בדיקה מאוד קפדנית יחסית למחקרים אחרים דונים עם מספר גדול של ריצות (2250), 3 מודלים שונים, 3 תקציבים ו-5 נרטיבים שונים. זאת בנוסף לביקורת ניטרלית ולמדד בריאות כמותי ואובייקטיבי יחסית - ניכר שנעשה פה מאמץ. עם זאת, מדובר בסביבה סימולטיבית ומוגבלת (50 מוצרים בלבד, חנות מדומה). במציאות, שוק המזון מורכב בהרבה, עם אלפי מוצרים, מבצעים משתנים והקשרים תרבותיים. ייתכן שההשפעות יהיו חלשות יותר או שונות בפרודקשיין.

בשורה התחתונה, המחקר הזה מספק ראיה ראשונית ודיי משכנעת לכך שסוכנים מבוססי LLM עשויים להיות פגיעים להקשרים רגשיים באופן הדומה לבני אדם. חשיפה לנרטיבים טראומטיים הובילה את הסוכנים לבצע החלטות צרכניות פחות בריאות, תופעה עקבית ובעלת עוצמה גבוהה. התוצאות האילו מעלות שאלות חשובות על אמינות הסוכנים האוטונומיים ואת הצורך להטמיע מנגנוני בקרה והגנה לפני שייעשה בהם שימוש נרחב בחיי היומיום. האם אני שומע פה מקום לאיזה יוניקורן שמחכה לקרות? אולי…

המאמר המלא בלינק הבא:

https://www.researchsquare.com/article/rs-7587964/v1