פורסם ב- SIGTRS, ינואר 2026

הצעה לשיפור התנהגות מודלים מול משתמשים ב chat gpt 5 -בטיפול בטקסט עברי / עפר דרורי

נושא :פער בין כוונת המשתמש האנושית לבין תוצאת הביצוע של מודלים גנרטיביים

רקע:
במהלך עבודה ממושכת עם מודל בינה מלאכותית ,(CHATgpt / DALL·E) נתקלתי בתופעה עקבית שבה בקשות מפורשות אינן מתממשות כנדרש, למרות ניסוח ברור, חזרות והדגשות. דוגמה בולטת לכך היא הוראה לייצר **מגן דוד אדום על סרט חובש ישראלי** ,אשר שוב ושוב מומרת ל־**צלב אדום או כוכב מערבי** —בניגוד מוחלט להנחיה.

הבעיה:
נראה שהמודל נשלט בפועל על-ידי הסתברויות של דוגמאות למידה ("מה סביר שייראה כמו סרט רפואי") ולא על-ידי **כוונת המשתמש** בזמן אמת.
כך נוצר מצב שבו המערכת "מעדיפה" תבנית נלמדת על פני הוראה ספציפית, גם כשזו חוזרת שוב ושוב—
כלומר, המודל מתנהג **כאילו יש לו רצון מובנה משלו** ,ומבטל את רצון המשתמש בפועל.

המשמעות:

- אובדן שליטה אנושית על תוצאה יצירתית.
- פגיעה באמינות הכלי לצרכים מקצועיים או תיעודיים.
- קושי באימוץ המערכת בתחומים רגישים (היסטוריה, רפואה, ביטחון, תרבות).

המלצה לשיפור:

1. להוסיף שכבת בקרה (post-processing) שמוודאת נאמנות להוראות המשתמש המפורשות, גם במחיר חריגה מהסתברויות למידה.
2. לאפשר למשתמש **"מצב נאמנות מוחלטת להנחיה** – "מצב שבו המודל מתעדף מילולית את הטקסט האנושי מעל לכל הטיה סטטיסטית.
3. להוסיף ממשק דיווח פשוט ("המודל חרג מההנחיה שלי") כדי לאפשר למשתמשים לתקן בזמן אמת.

הערה עקרונית:
כמי שעוסק מעל 40 שנה במחשוב, אני רואה כאן לא בעיית קוד, אלא בעיית תפיסה:
המודלים צריכים לזכור שהם **עוזרים לבני אדם להביע כוונה** ,לא ממציאים כוונה משלהם.