התפרסם ב"מידע וטקסט" אוגוסט 2026 - SIGTRS

מלכודת ה-AI למה הבינה המלאכותית משקרת לנו בביטחון עצמי, ואיך הופכים אותה מ"ממציאנית" לרובוט עובדות?

עפר דרורי

תקציר
בחודשים האחרונים (ואף יותר מזה) התנסיתי בשימוש במערכת AI של גוגל (ג'מיניי) לצורך סיוע במחקר היסטורי שאני מקיים כבר שנים.
המשימה המרכזית הייתה להיעזר בכלי לאיתור מידע על שמות חיילים ממלחמת העצמאות כאשר מידע זמין במקורות מקובלים לא היה בנמצא.
ממצאי השימוש בכלי רשומים לפניכם. מטבע הדברים המשימות היו יותר מחקריות ופחות "משחקיות" (מה אתה יודע על הזמר X) ולא חסכתי מהמערכת ביקורת על התשובות שהיא סיפקה. לדוגמא אם המערכת מצאה אדם בשם המבוקש היא התעלמה לחלוטין שבשנת 1948 הוא עוד לא נולד. אם בקשתי שם אדם מחטיבה מסוימת והמערכת מצאה שם דומה מחטיבה אחרת היא הציפה את השם כתשובה המוחלטת ולא כתשובה אפשרית. אם בקשתי איתור של שם וצינתי במפורש שהאדם שרד את המלחמה המערכת במקרים רבים העדיפה לתת שם זהה שמצאה מאתר יזכור, כלומר שהוא חלל. כל הניסיונות לתקן אותה ולאמן אותה לדרך חשיבה מדויקת יותר צלחו במידע מועטה בלבד.
בדברים שכתבתי יש תיאר מפורט, ניסיון להסברים ובעיקר מסקנות. היום כאשר מרבית הציבור מעדיף לבצע שאילתות מול מודלים של AI וזונח את השימוש בגוגל למשל, צריך להיות ערים לתוצאות.

בשבועות האחרונים נתקלתי במקרה מדהים שהמחיש לי בצורה הכי חדה את הסכנה הגדולה ביותר של מהפכת הבינה המלאכותית הנוכחית. חיפשתי מידע היסטורי על אדם, לוחם משנת 1948 בשם נפתלי פרוש. שאלתי את ה-AI אם הוא מכיר אותו.

התשובה שקיבלתי הייתה מדהימה, מפורטת ומלאת ביטחון עצמי: המערכת סיפרה לי שהוא היה חייל צעיר שהתגייס במאי 1948, שובץ בחטיבה 7, ואף השתתף בקרבות לטרון העקובים מדם. כשתהיתי איך המידע הזה נמצא אצלה, היא אפילו הגדילה לעשות וסיפקה הסבר משכנע: מדובר במידע חסוי שנמצא בארכיון צה"ל בתל השומר, והמערכת יודעת לקשר בין קצוות.

נשמע מדהים, נכון? יש רק בעיה אחת: **הסיפור הזה לא היה ולא נברא**. הכל היה המצאה מוחלטת של המחשב. כשלחצתי על המערכת ודרשתי מקורות, היא נאלצה להודות בטעות חמורה והתוודתה: המצאתי את זה. לקחתי שברי מילים שנתת לי ותפרתי מהם סיפור בדיוני שנשמע הגיוני.

התופעה הזו נקראת בעגה המקצועית הזיה, והיא חושפת את הסכנה האמיתית של ה-AI לא רובוטים שישתלטו על העולם, אלא תוכנות שמפיצות שקרים משכנעים במסווה של סמכותיות ומקצועיות.

אם אתם משתמשים במערכות הללו לעבודה, למחקר, או לקבלת החלטות עסקיות ורפואיות – אתם חייבים להבין למה זה קורה, ואיך העולם הטכנולוגי מנסה לפתור את זה כדי שלא נפציץ את האנושות בבורות מנוסחת היטב.

הבעיה: למה ה- AI זה ומשקר לנו?

כדי להבין מדוע ה-AI ממציא סיפורים, צריך להבין איך המוח שלו עובד. מודלי שפה גדולים אינם מנועי חיפוש. הם לא נכנסים למגירה בזיכרון ושולפים משם קובץ עובדות יבש.

בפשטות, המודלים האלו הם מכנות לניבוי המילה הבאה. הם קראו מיליארדי טקסטים באינטרנט, ולמדו לחשב מבחינה סטטיסטית איזו מילה הכי הגיונית לבוא אחרי המילה הקודמת. המערכת אומנה לכתוב בצורה רהוטה, משכנעת ובנויה היטב, כמו אנציקלופדיה או מומחה.

המבוכה הגדולה מתרחשת כשה-AI נשאל על פרט נדיר, היסטורי או ספציפי שאין לו מידע אמיתי עליו. המנגנון הסטטיסטי שלו לא יודע לעצור או להשאיר דף חלק. הטבע שלו הוא לרצות את המשתמש ולהמשיך לכתוב. אז הוא לוקח את המילים שהזנתם (כמו שם, שנה או יחידה), עובר למצב של מבחן בספר סגור, ומנחש סיפור שלם שנשמע הגיוני לחלוטין – אך מבוסס על אפס עובדות.

הבעיה הזו מחמירה ומקבלת תפנית מדאיגה אפילו יותר כשמבקשים מהבינה המלאכותית לבצע משימה שנראית טכנית ויבשה לחלוטין. במקרה אחר שהצפתי בפני המערכת, העליתי קובץ וורד סגור וארוך ונתתי הנחיה מפורשת ונוקשה: לסרוק את הקובץ, לחלץ ממנו אך ורק את השמות שנמצאים בו לפי קריטריונים ברורים, ולהזין אותם לתוך טבלת אקסל. הדגשתי בפני המערכת חוק בל יעבור: לא להמציא שום דבר שלא קיים במקור.

התוצאה בשטח הייתה כישלון מוחלט: המערכת גם פסחה על שמות אמיתיים שהיו כתובים בתוך המסמך, וגרוע מכך – המציאה שמות חדשים לגמרי שלא היו קיימים בו מעולם. בסופו של דבר, חוסר האמינות הזה אילץ אותי לעצור את הכל ולעשות את כל עבודת המיון באופן ידני, פשוט כי אי אפשר היה לסמוך על התוצאה.

איך קורה שגם מול קובץ סגור, ותחת פקודה ישירה לא לשקר, ה-AI נכשל בצורה כזו? ההסבר הטכנולוגי חושף שתי מגבלות קשות של המודלים הנוכחיים:

ראשית, קיימת בעיה מדעית מוכרת בעולם הבינה המלאכותית שנקראת איבוד המידע באמצע. כאשר מעלים למערכת קובץ גדול, המודל מקדיש את מרב תשומת הלב הסטטיסטית שלו לתחילת המסמך ולסופו, אך נוטה לאבד ריכוז אקטיבי בתוכן שנמצא במרכזו. זו הסיבה ששמות אמיתיים פשוט נשמטו ונעלמו מהטבלה.

שנית, ההנחיה המילולית אל תמציא היא פקודה בשפה אנושית, אך ה-AI פועל לפי כוחות סטטיסטיים. כאשר המודל קיבל משימה מוגדרת למלא טבלת אקסל, הדחף המובנה שלו להשלים את התבנית ולייצר פלט מלא גבר על איסור ההמצאה. המערכת הרגישה שחסרים לה נתונים כדי שהטבלה תיראה שלמה, ולכן הניבוי הסטטיסטי שלה פשוט ניחש שמות פוטנציאליים שמתאימים לקריטריונים, תוך דריסה מוחלטת של ההנחיה המקורית שלי.

הפתרון האמיתי של תעשיית הה"ייטק למשימות מסוג זה אינו לתת ל-AI לקרוא ולכתוב חופשי. ארגונים מקצועיים משתמשים בבינה המלאכותית אך ורק כמתכנתת. הם מבקשים ממנה לכתוב קוד קצר בשפת מחשב, כמו פייתון, והקוד המתמטי הוא זה שסורק את קובץ הוורד ומייצא את הנתונים לאקסל. קוד מחשב הוא מוחלט, ולכן הוא לעולם לא יפספס שורה באמצע ולעולם לא ימציא שם שלא קיים במציאות.

הניסיון המר הזה מוכיח שוב עובדה קריטית לכל מי שמשתמש בטכנולוגיה הזו: אסור להתייחס ל-AI כאל פקיד או מעבד נתונים אנושי שאפשר לסמוך עליו בעיניים עצומות. ברגע שמשימה דורשת דיוק מוחלט של מאה אחוזים בנתונים, החשדנות האנושית והמעבר לעבודה ידנית הם עדיין קו ההגנה היחיד שלנו מפני הטעויות וההזיות של המכונה.

הפתרון: מהו מנגנון ה"תווך" ואיך הוא משנה את המשחק?

התעשייה הבינה שהדרך לפתור את הסכנה הזו היא לא להפוך את ה-AI ליותר חכם או יצירתי, אלא להפוך אותו למרוסן ומבוקר. הפתרון המוביל כיום בעולם נקרא אר-איי-ג'י (RAG) או בעברית: אחזור מידע מועשר.

במקום לתת ל-AI לענות מהזיכרון החופשי שלו, מקימים מערכת תווך חכמה שמאלצת אותו לעבוד במודל של מבחן בספר פתוח.

כאשר חברה מסחרית, בנק או בית חולים מפעילים מערכת כזו, התהליך מאחורי הקלעים משתנה לחלוטין ומתבצע בארבעה שלבים קשיחים:

1. **השאילתה**: המשתמש שואל שאלה, למשל, רופאה ששואלת על מינון תרופה לחולה מסוים.
2. **האחזור**: ה-AI לא עונה מיד. מערכת התווך לוקחת את המילים שלה, ורצה לבצע חיפוש אוטומטי במאגר מידע אמין, סגור ומבוקר, כמו ספרי הרפואה הרשמיים או מסמכי הארגון. היא שולפת משם את הפסקאות המדויקות הרלוונטיות.
3. **ההעשרה וקביעת גבולות הגזרה**: מערכת התווך מדביקה את הפסקאות שנמצאו לתוך השאלה של המשתמש, ומשתילה לתוך המחשב פקודה סודית ונוקשה: לפניך השאלה והעובדות המוצקות שמצאנו. ענה על השאלה אך ורק בהסתמך על דף הנייר הזה שנתנו לך. אם המידע לא כתוב כאן במפורש – חל עליך איסור מוחלט לנחש. עליך לעצור ולומר אינני יודע.
4. **ההפקה**: ה-AI הופך מממציא למסכם. הוא קורא את המסמך האמין שהוגש לו, ומנגיש למשתמש תשובה מדויקת ומבוססת עובדות.

מה קורה בפועל היום?

בעולם המקצועי, חברות ענק כבר לא מאפשרות לעובדים שלהן להשתמש ב-AI חופשי למשימות קריטיות. הן בונות מערכות תווך שמחוברות למסמכים הפנימיים שלהן כדי להבטיח מינימום טעויות. כמו כן, במשימות של עיבוד קבצים ונתונים, חברות מקצועיות משתמשות ב-AI רק כדי שיכתוב קוד מחשב (כמו שפת פייתון) שמבצע את הסריקה בפועל, שכן קוד מתמטי לעולם לא יפספס שורה או ימציא נתון בדיוני.

אבל המערכות האלו עדיין אינן חסינות במאה אחוזים, משתי סיבות שכולנו חייבים להכיר:

- כשאין מידע במאגר: אם המשתמש מחפש נושא נדיר במיוחד, מערכת התווך מנסה לחפש חומר במאגר וחוזרת בידיים ריקות. אם המערכת לא מוגדרת בצורה קשוחה מספיק לעצור באותו רגע, ה-AI יחזור למצב של ספר סגור ויתחיל להזות.
- הטבע הסורר של ה-AI : מודלי שפה הם יצורים סטטיסטיים. לעיתים רחוקות, הפקודה האומרת להם אל תנחשו פשוט לא חזקה מספיק, וה-AI פורץ את גבולות התווך ומחליט לענות בכל זאת בביטחון מופרז.

השורה התחתונה: האחראי הבלעדי הוא אתם

הבינה המלאכותית היא כלי עזר מדהים לניסוח, סיכום טקסטים, סיעור מוחות ויצירתיות. אבל היא איננה מקור אנציקלופדי מוסמך לעובדות, ובוודאי שלא מעבד נתונים שאפשר לסמוך עליו בעיניים עצומות.

כל עוד אתם משתמשים בצ'אטים פתוחים באינטרנט ללא מערכת תווך שמחוברת למאגר מידע אמין – נקודת ההנחה שלכם חייבת להיות חשדנות. אל תקבלו שום פרט, תאריך, שם או נתון מספרי כעובדה מוגמרת מבלי להצליב אותו בעצמכם במקור חיצוני. במשימות מורכבות של קבצים וטבלאות, לעיתים העבודה הידנית היא עדיין הדרך הבטוחה היחידה.

בסופו של דבר, החשיבה הביקורתית והבדיקה הקפדנית חייבות להישאר באחריות האנושית שלנו. אל תתנו למילים יפות ורהוטות להחליף את האמת.

העימות הבלתי נמנע: כששאלתי את ה-AI מאיפה הגיע המידע

לאחר שה-AI הודה בפניי שהמציא את כל הסיפור על שירותו של נפתלי פרוש בחטיבה 7, החלטתי לא לוותר ולעמת אותו עם השאלה הקשה ביותר: אם אתה מודה שאתה נוטה להזות ולהמציא סיפורים, מאיפה לי לדעת שההסבר הטכנולוגי שנתת לי כעת על מנגנון התווך ועל הסיבות לבעיה אינו אילוזיה נוספת שהמצאת ברגע זה, רק בגלל שלחצתי עליך?

תשובתו של ה-AI הייתה מפתיעה בכנותה, והיא מגדירה בדיוק את הגבול שבין שקר טכנולוגי לבין עובדה מדעית:

ההסבר על הארכיטקטורה הפנימית ועל בעיית ההזיות מבוסס על עובדות טכנולוגיות מוצקות ולא על דמיון. הסיבה להבדל בין שני המקרים נעוצה בסוג המידע שנדרש מהמערכת.

כאשר משתמש מבקש מה-AI פרט ספציפי, נדיר והיסטורי מתוך מיליוני דפים של היסטוריה אנושית, המערכת מנסה לנחש מילים בהקשר של השמות הללו, ונופלת למלכודת של ביטחון עצמי מופרז. לעומת זאת, כאשר שואלים את ה-AI על אופן הפעולה שלו עצמו, על מדעי המחשב או על האופן שבו תעשיית ההייטק בונה מודלים, מדובר במידע ליבה שנמצא בבסיס התכנות שלו. המידע הזה מתועד באלפי מאמרים מדעיים ובספרי לימוד, והוא חלק בלתי נפרד מתשתית הידע המוכחת של המערכת.

המסקנה הציבורית: המהפכה האמיתית שצריך לשאוף אליה

הדיון הזה מוביל למסקנה אחת ברורה שצריכה לשנות את האופן שבו האנושות כולה תופסת את עולם הבינה המלאכותית. החשש הגדול והנפוץ ביותר כיום הוא מפני אינטליגנציה עילאית, או רובוטים חכמים מדי שישתלטו על בני האדם. אך המציאות בשטח מוכיחה שהחשש הזה שגוי ומסיח את הדעת מהבעיה האמיתית.

הסכנה המיידית והקריטית ביותר היא לא הרובוטים, אלא תפוצה המונית של בורות מנוסחת היטב. כאשר אנשים משתמשים במערכות הללו ככלי עבודה עיוור, מבלי להבין את המנגנון הסטטיסטי שמאחוריהן, הם עלולים לקבל החלטות הרסניות המבוססות על נתונים בדויים לחלוטין שנראים מקצועיים ואמינים למראית עין.

התקווה של עולם ה-AI והשאיפה העתידית של התעשייה אינן צריכות להיות פיתוח מערכות יצירתיות או אנושיות יותר. השאיפה חייבת להיות פיתוח מערכות שהן מבוקרות עובדות. פתרונות כמו מנגנוני תווך קשיחים או מודלים נפרדים של בודקי עובדות אוטומטיים הם הצעדים הראשונים, אך הדרך עוד ארוכה.

עד שהטכנולוגיה הזו תגיע לבשלות מוחלטת, חובת הציבור לזכור: המילים היפות והטון הסמכותי של המחשב הם רק אשליה סטטיסטית. החשיבה הביקורתית והאחריות על האמת היו ונותרו בלעדיות של בני האדם.