המאמר היומי של מייק: 31.07.25

Empirical evidence of Large Language Model's influence on human spoken communication

ההיפוך הלשוני: כשבני אדם מתחילים לחקות את המכונות שהם אימנו

מאמר חדש(גרסתו השנייה אמנם) חשף לאחרונה אות מצמרר אך רב-משמעות: מאז סוף 2022, האופן שבו אנשים מדברים, כן, מדברים, לא כותבים, משתנה באופן מדיד ומתקרב לטביעת האצבע המסוגננת, הכמעט מטרידה, של ChatGPT.

חוקרים ניתחו יותר מ-740,000 שעות של דיבור אנושי' החל מערוצי יוטיוב אקדמיים ועד לפודקאסטים יומיומיים. האות שהם מצאו הוא בלתי ניתן להכחשה מבחינה סטטיסטית: אנשים החלו להשתמש במונחים המועדפים באופן מובהק על ידי מודלים מסוג GPT, בשיעורים גבוהים משמעותית מהמגמות ההיסטוריות. מילים כמו "להתעמק" (delve), "קפדני" (meticulous), "להתהדר" (boast), "מורכב" (intricate) ו"להבין" (comprehend) זינקו בתדירותן, עם שינוי שיפוע שתואם במדויק את השקתו הציבורית של ChatGPT. וזו לא רק סטייה מקרית של פרוזה כתובה שזולגת לדיאלוג. זוהי לולאת משוב התנהגותית, ויש לה השלכות שחורגות הרבה מעבר לשפה.

במערכות דינמיות, אנו מחפשים לעיתים קרובות מעברי פאזה: נקודות שבהן מערכת מארגנת את עצמה מחדש באופן פתאומי למשטר חדש מבחינה איכותית. הראיות כאן מצביעות בדיוק על כך. לפני ChatGPT, הצמיחה בשימוש במילים המועדפות על GPT הייתה איטית, כמעט לינארית, חשבו על סחף לקסיקלי טבעי לאורך זמן. אבל לאחר שחרורו של ChatGPT, השיפוע משתנה. בחדות. כמעט באופן בלתי רציף.

זהו סמן קלאסי להפרעה מחוץ לשיווי משקל: כוח חיצוני כלשהו חולל שינוי משטר במערכת לשונית שעד כה נסחפה באטיות. הכוח הזה, במקרה שלנו, הוא הפלט של LLMs שמציף את המרחבים הדיגיטליים ומשפיע בעדינות על האופן שבו בני אדם מדברים; במיוחד אלו המשובצים בתת-תרבויות הקרובות לעולמות הבינה המלאכותית.

הסיפור היסודי של מודלי שפה גדולים תמיד היה חד-כיווני: בני אדם ← נתונים ← מודל. אבל עכשיו אנחנו עדים להיפוך: מודל ← נתונים ← בני אדם. זה לא ספקולטיבי. זה ניתן לכימות.

שפה היא תהליך סטוכסטי רב-ממדי. כאשר מתחילים להבחין בגרדיאנטים קוהרנטיים, שבהם אשכולות שלמים של ביטוי אנושי מתחילים לנטות לכיוון שדה וקטורי שנוצר על ידי פלטי מכונה. זו כבר לא הופעה ספונטנית. זוהי סחיפה (entrainment). הסחיפה הזו אינה מוגבלת רק לבחירות לקסיקליות. היא זולגת לפרוזודיה, למבנה, לסגנון הטיעון. טנזור המטריקה כולו של השיח מתעוות כדי להתיישר עם מה שמודלי השפה למדו לייצר. ומכיוון שהמודלים הללו מאומנים על חיזוי המילה הבאה (next-token prediction), המשטח הלשוני שלהם מותאם לקוהרנטיות, נימוס, שטף ויכולת חיזוי. אבל עצם האופטימיזציה הזו מענישה אי-סדירות, עמימות וסטייה מהנורמות הסטטיסטיות.

בואו נחשוב במונחים של גיאומטריה של המידע. לשפה, במצבה הטבעי, יש אנטרופיה גבוהה: מגוון של טונים, משלבים, ניבים, היסוסים ושימוש יצירתי שגוי. LLMs, לעומת זאת, משטיחים את המרחב הזה. הם ממלאים פערים בהשלמות בנויות היטב המבוססות על סבירות מרבית לא על חדשנות הבעתית.

כאשר דוברים אנושיים מתחילים לחקות LLMs במודע או שלא במודע, הם מפחיתים את האנטרופיה של השיח. אנחנו מתחילים להעדיף פניות שיח בטוחות, דמויות-GPT. הדיוק עולה, אך השונות פוחתת. וכאשר השונות קורסת, יכולת ההבעה מתחילה להישחק. המערכת מאבדת מעושרה גם כשהיא זוכה בבהירות. למעשה, אנחנו מחליפים מרקם סמנטי בסדירות תחבירית.

לא מדובר רק בלהישמע רובוטי. שפה משקפת מחשבה. אם אנחנו מעצבים מחדש את האופן שבו אנחנו מדברים, אנחנו בהכרח מעצבים מחדש את האופן שבו אנחנו חושבים. ואם המחשבה שלנו מתחילה להתיישר עם הנחות היסוד המבניות של מכונה שאומנה על חסכנות סטטיסטית, מה קורה ליכולת שלנו לסתירה? לעמימות יצירתית? לכישלון יצירתי?

זה לא שמודלי השפה מחליפים אותנו. ייתכן שאנחנו מפנימים אותם.

מה שהופך את זה למשכנע באמת אינו המגמה השטחית, אלא הטופולוגיה הסיבתית שלה. השתמשנו בדיבור אנושי כדי לאמן מכונות. המכונות משפיעות כעת על הדיבור האנושי. זהו מעגל סיבתי בעל חיזוק עצמי.

אם אי פעם למדתם מערכות עם משוב, אתם מכירים את החשש הקנוני: לולאות משוב חיובי אינן יציבות אלא אם כן הן מווסתות. הלולאה הזו אינה מווסתת. אין מנגנון ריסון טבעי. אין מערכת חיסונית לשונית. ובניגוד לאופנה או מוזיקה, שעוברות במחזורים של פופולריות, פלטי המכונה יציבים באופן אסימפטוטי. ברגע שהם מתייצבים על ביטויים בעלי סבירות גבוהה, הם אינם סוטים. ואם בני האדם ננעלים בהתיישרות עם אותם ביטויים, המערכת עלולה להתכנס אך במחיר החיוניות.

האם עלינו להיכנס לפאניקה? לא. אבל עלינו להתבונן. למדוד. להתערב במידת הצורך.

כמה רעיונות:

לנטר את האנטרופיה הלשונית לאורך זמן בקרב אוכלוסיות החשופות לבינה מלאכותית. ירידה בשונות עשויה לאותת על התיישרות-יתר.

לעודד דיסוננס: להעריך ביטויים ייחודיים, שאינם דמויי-LLM – במיוחד בדיבור.

לגוון את מאגרי הנתונים (corpora): להזין מודלים בנתונים עשירים בניבים אזוריים, אנגלית עילגת, אינטונציה רגשית – לא רק בפרוזה מחוטאת.

החלק המפחיד ביותר בשינוי הזה אינו שהמכונות מחליפות אותנו. הוא עדין יותר. ייתכן שאנחנו הופכים להדים לא מודעים של המערכות שבנינו, מבריקים, קוהרנטיים, מלוטשים... אך בסופו של דבר, נגזרים. הסכנה האמיתית אינה הסינגולריות. היא ההתכנסות.

והתכנסות, כשמותירים אותה ללא פיקוח, תמיד משטיחה את העקומ

https://arxiv.org/abs/2409.01754