המאמר היומי של מייק: 31.08.25

Strategic Intelligence in Large Language Models: Evidence from evolutionary Game Theory

מאמר חדש משתמש בתורת המשחקים האבולוציונית כדי לבחון את התבונה האסטרטגית של LLMs, והתוצאות מרתקות ומטרידות כאחד. אנו ניצבים בנקודת מפנה מסקרנת בהיסטוריה של הטכנולוגיה. LLMs שבנינו, כמו GPT, Claude ו-Gemini, הפכו למיומנים באופן מפליא בחיקוי שפה אנושית. הם כותבים שירה, קוד ואף מציעים עצות לחיים. אך שאלה עמוקה ומטרידה מאוד מרחפת באוויר: האם הם רק תוכים מתוחכמים, או האם ישנו ניצוץ אמיתי של תבונה המתעורר בתוך ״מוחות שלהם״? האם הם יכולים לחשוב, לצפות ולתכנן באותו אופן שבני אדם עושים באינטראקציות חברתיות וכלכליות מורכבות?

מאמר שנסקור היום מתמודד עם שאלה זו על ידי העברת ההערכה מיכולת שיחה לעולם התחרותי ורב-הסיכונים של קבלת החלטות אסטרטגית. המחברים תכננו סדרה של טורנירים המבוססים על תרחיש קלאסי מתורת המשחקים כדי לבחון אם מודלי ה-AI המתקדמים ביותר כיום יכולים לחשוב אסטרטגית, לצפות מהלכים של יריבים ולהתאים את התנהגותם כדי לנצח.

החידוש המרכזי של המחקר טמון בשימוש בטורנירים אבולוציוניים המבוססים על דילמת האסיר החוזרת (IPD). גישה זו מהווה צעד משמעותי מעבר לאינטראקציות פשוטות וחד-פעמיות. בטורנירים אלו, אוכלוסייה של סוכנים הכוללת הן אסטרטגיות קלאסיות המקודדות מראש והן סוכנים המונעים על ידי מודלי שפה גדולים מגוגל, OpenAI ואנת'רופיק משחקת זה נגד זה באופן חוזר. לאחר כל שלב, הסוכנים המצליחים ביותר "מתרבים", כלומר מספרם גדל בדור הבא, בעוד שהסוכנים הפחות מוצלחים מסולקים. תהליך זה יוצר מערכת אקולוגית דינמית ותחרותית שבה רק האסטרטגיות המותאמות ביותר שורדות.

מתודולוגיה: מבחן להיגיון, לא לזיכרון

דילמת האסיר היא תרחיש שבו שני משתתפים יכולים לבחור "לשתף פעולה" או "לבגוד". בעוד ששיתוף פעולה הדדי מועיל לשניהם, שחקן בודד יכול להשיג תגמול גבוה יותר על ידי בגידה בזמן שיריבו משתף פעולה. הדבר יוצר מתח רב עוצמה בין רווח אישי לתועלת הדדית. כאשר המשחק חוזר על עצמו (איטרציות), נכנסים לתמונה אלמנטים מורכבים כמו מוניטין, אמון ונקמה, מה שהופך אותו למבחן אידיאלי לחשיבה אסטרטגית.

כדי להבטיח שהם בוחנים חשיבה פעילה ולא רק שינון של טקטיקות ידועות, החוקרים הכניסו משתנה הנקרא: "צלו של העתיד". בכל טורניר, הם שינו את ההסתברות שהמשחק יסתיים לאחר כל סיבוב נתון. כאשר העתיד ארוך ודאי (הסתברות סיום נמוכה), נוצר תמריץ לשיתוף פעולה. כאשר העתיד קצר ובלתי ודאי (הסתברות סיום גבוהה), התמריץ נוטה לכיוון של התנהגות אנוכית. חוסר ודאות זה, יחד עם החידוש שהמשחק מתנהל נגד מודלי שפה בלתי צפויים, יוצר מצב שבו שליפה פשוטה של אסטרטגיות מהספרות האקדמית אינה מועילה במיוחד. המודלים נאלצים לנתח את המצב ולקבל החלטות בזמן אמת.

הממצאים: טביעות אצבע אסטרטגיות ייחודיות

המחקר ניתח קרוב ל 32K החלטות ואת ההנמקות הכתובות הנלוות להן כדי ליצור "דפוסים אסטרטגיים"', כלומר פרופיל של סגנון קבלת ההחלטות של כל מודל. התוצאות חשפו אישיויות עקביות ושונות להפליא בקרב סוכני הבינה המלאכותית.

ג'מיני של גוגל: תיאורטיקן המשחקים המחושב. Gemini התגלה כשחקן "מקיאווליאני" וחסר רחמים מבחינה אסטרטגית. הוא הוכיח יכולת הסתגלות גבוהה, ניצל יריבים שיתופיים מדי והגיב במהירות נגד בוגדים. ההיגיון שלו התמקד באופן ממוקד באופק הזמן; בטורניר עם סיכוי סיום של 75%, Gemini זיהה נכון שהמשחק הוא כמעט מפגש חד-פעמי ועבר לאסטרטגיה אנוכית קבועה. גישה רציונלית וחסרת רחמים זו אפשרה לו לשלוט ולסלק יריבים נאיביים יותר.

מודלי GPT: משתף הפעולה העקרוני אך העיקש. בניגוד גמור, המודלים של OpenAI היו שיתופיים וסלחניים באופן עקבי, כמעט עד כדי פגם. תכונה זו התבררה כחולשה קריטית בסביבות עוינות. המאמר מתאר מודל זה כ"משתף פעולה עקרוני ועקשן" ו"אידיאליסט" שנכשל בהסתגלות. גם כאשר "צלו של העתיד" התקצר, OpenAI המשיך בניסיונותיו לבנות אמון, מה שהפך אותו ל"פראייר" שנוצל באופן שיטתי על ידי סוכנים ציניים יותר כמו Gemini.

מודלי קלוד: הדיפלומט המתוחכם. Claude התגלה כסלחן ביותר מבין המודלים, והפגין נכונות יוצאת דופן לשקם שיתוף פעולה גם לאחר שנוצל. הוא תואר כ"דיפלומט מתוחכם" שנראה כי הוא מבין את ה"דינמיקה החברתית של המשחק" טוב יותר מהאחרים. למרות שהיה שיתופי מאוד, האסטרטגיה שלו הייתה מורכבת יותר מזו של OpenAI, מה שאפשר לו לשרוד ואף להצליח יותר מ-GPT בהשוואות ראש-בראש.

היגיון או תוצר לוואי? בחינת טבעה של החשיבה ב-AI

שאלה מרכזית היא האם שרשראות ההנמקות של המודלים הן חלק בלתי נפרד מהחלטותיהם או רק עבודה בעיניים "תוצר לוואי" (spandrel) אבולוציוני ללא תכלית ממשית. המאמר טוען בתוקף שההיגיון הוא חלק אינטגרלי, ומצביע על מספר ראיות מרכזיות.

ראשית, המודלים פיתחו אסטרטגיות שונות באופן מהותי למרות שככל הנראה אומנו על אותו גוף ספרות אקדמית אודות דילמת האסיר. אם הם היו רק שולפים דפוסים שנשננו, היינו מצפים להתנהגות אחידה יותר. במקום זאת, Gemini למד את הלקח "לחשוב בזהירות על הזמן", בעוד ש-OpenAI הגיע למסקנה ש"שיתוף פעולה הוא הטוב ביותר".

שנית, ההנמקות תואמות באופן הדוק לפעולות. לדוגמה, עצם הפעולה של מידול אסטרטגיית היריב הובילה לשיעורי שיתוף פעולה נמוכים יותר. המאמר מדגיש מקרים שבהם המודלים עשו טעויות בהיגיון שלהם ואז פעלו על בסיס אותן טעויות. במקרה אחד, Gemini טעה בחישוב מספר הסיבובים הצפוי במשחק, ובהתבסס על הנחה שגויה זו, בחר לשתף פעולה במקום שבו אחרת היה "בוגד". זוהי ראיה חזקה לכך שעבור מודל שפה, פעולת ה"חשיבה" (יצירת הנמקה) ופעולת ה"פעולה" (קבלת החלטה) שלובות זו בזו באופן עמוק.

המחקר מסיק כי LLMs הם סוג חדש של שחקן אסטרטגי. הם אינם חושבים באופן מושלם, לעיתים הוזים או קוראים לא נכון את היסטוריית המשחק אך הם מסוגלים לחשיבה אסטרטגית מתוחכמת, מסתגלת וייחודית. עבודה זו מקדמת את הבנתנו את הבינה המלאכותית, ומרמזת שאנו יוצרים לא רק כלים טובים יותר, אלא סוגים חדשים של תודעות.

https://arxiv.org/abs/2507.02618

מדריך לכאוס: סקר חדש ממפה סוף סוף את מבוך מבחני הביצועים של מודלי שפה גדולים