⚡️🚀המאמר היומי של מייק 24.06.24:⚡️🚀

Are you still on track!? Catching LLM Task Drift with Activations

הסקירה הזו הולכת להיות קצרה כי הרעיון העיקרי של המאמר הוא די פשוט ואינטואיטיבי. אתם מדברים עם מודלי שפה שלכם באמצעות שאילתות שבד״כ נקראות פרומפטים שהמודל עונה לכם. אבל מה קורה אם מודל השפה שלכם מחובר לעוד כלי שמגנרט בשבילו פרומפטים למשל בהתבסס על תוצאה של איזשהו חישוב על הפלט של מודל אחר או מתבסס על RAG או אולי אפילו תלוי בתוצאות חיפוש באינטרנט.

כמובן שגנרוט אוטומטי של פרומפט יכול להתפקשש (באגים, אולי פעילות זדונית) ואז יחד עם שאלה לגיטימית המודל מקבל תופסת לא קשורה. בעיה ידועה, אה?

אז המאמר שבנידון חקר את האקטיבציות של שכבות המודל (טרנספורמר כמובן) ומצאו הבדלים משמעותיים בין האקטיבציות הנוצרות על ידי שאלה לגיטימית לבין אלו שנוצרו עם שאלה ״מקושקשת״. ואז הם בנו דאטהסט של שאלות טובות ושאלות מורעלות ואימנו מודל (קטן) שיודע להבדיל בין האקטיבציות של שאלות הטובות והלא טובות. המחברים לוקחים אקטיבציות של הטוקן האחרון של הפרומפט (השאלה) המלא

הם ניסו שתי שיטות: אחת היא אימון של שכבה לינארית המפרידה בין ייצוגים טובים ומורעלים. השיטה השניה שהם מנסים נקראת metric learning שבמילים פשוטות מנסה ללמוד ייצוג (המופק על ידי המודל ״המבדיל״) המקרב ייצוגים של העוגן (התחלת השאלה) עם השאלה הטובה ומרחיק אותו מהייצוג של השאלה המורעלת (התוספת המורעלת). אם מצליחים ב-metric learning אז בקלות אפשר לתפור שכבה לינארית המבדילה בין הטובים ללא טובים.

https://arxiv.org/pdf/2406.00799