המאמר היומי של מייק: 29.08.25

DEEP THINK WITH CONFIDENCE

סקירה מס 497 - עוד 3 סקירת בדרך ל-500 והיום סקירה קצרה של מאמר בעל שם מפוצץ (שאכן נהנה מהייפ משמעותי) עם רעיון די אינטואיטיבי שגרם לי לתהות איך אף אחד לא עשה את זה קודם (אם זה נכון). המאמר מציע שיטה מבוססת אנטרופיה לדגימה ממודלי שפה אוטורגרסיביים (למרות שלדעתי אפשר יחסית בקלות להרחיב את הגישה המוצעת למודלים שמגנרטים פלט בצורה לא אוטורגרסיבית כמו מודלי שפה מבוססי דיפוזיה). כמו שאתם בטח יודעים אנטרופיה הינו מדד לאי ודאות וניתן להשתמש בו במודלי שפה למטרת שערוך של ״מידת הביטחון״ של המודל בפלט שהוא מג'נרט.

מודלי שפה אוטורגרסיביים מגנרטים כל טוקן בהתבסס על ההתפלגות של הטוקן הזה בהינתן ההקשר הקודם לו. ככל שהאנטרופיה של הטוקן הנחזה, השווה למינוס לוג של ההסתברות שלו, גבוהה יותר האי ודאות שלו גבוה יותר. כלומר ככל שההסתברות הטוקן יורדת, אי הוודאות עולה הקשורה בבחירותו עולה. כאמור המחברים מציעים שיטת דגימה מבוססת אנטרופיה ממוצעת של הטוקנים בטקסט מג'ונרט.

בפרט אם במקרים שהמודל מגנרט כמה כמה תשובות לשאלה מתמטית ואז אנו בוחרים את התשובה הנכונה לא עם ה-majority vote פשוט (כלומר התשובה הסופית שרוב התשובות התכנסו אליה) אלא על ידי משקול תשובה עם הודאות שלה כלומר עם הממוצע של האנטרופיה של כל הטוקנים שלה. כך תשובות שהמודל ממש לא בטוח בהם מפולטרות. המחברים גם מציעים לקבוע סף של אי וודאות מקסימלית של תשובת המודל. אם אי הוודאות הממוצעת השוטפת של התשובה (מחושבת מחדש עבור כל טוקן מג'ונרט), התשובה נפסלת והמודל מספיק לגנרט אותה. הסף נקבע בתור אחוזון של אי הוודאויות של התשובות הנכונות בשלב ה-warmup.

בנוסף המחברים מציעים לקבוע את מספר תשובות הנדגמות מהמודל בהתבסס על קושי השאלה. ככל ש״יש מעט מדי הסכמה״ בין התוצאות של התשובות השונות המודל מגנרט יותר תשובות כאשר התשובות בעלות אי וודאות גבוהה מדי מפולטרות כאמור.

מאמר נחמד אבל משאיר תחושה שכבר ראיתי משהו כזה בעבר….

https://arxiv.org/abs/2508.15260

שחר ה-AI המקיאווליאני? חשיפת התבונה האסטרטגית של LLMS