⚡️🚀המאמר היומי של מייק 12.07.24: ⚡️🚀

To Believe or Not to Believe Your LLM

מאמר מאוד מעניין מבית גוגל. המאמר מנסה להבין איך ניתן לזהות עד כמה המודל בטוח בתשובתו לשאלה. כלומר המאמר עוסק בכימות של אי ודאות של תשובות המודל. המאמר מנסה בין שני סוגים של אי-וודאות הידועים בתורת השערוך: אלטורי (aleatoric) אפיסטמי (epistemic). אי-הוודאות האפיסטמית מתרחשת כאשר המודל לא יודע מה התשובה לשאלה ומתחיל לאלתר (כלומר להוציא הזיות או hallucinations). לעומת זאת אי הוודאות אלאטורית מתרחשת כאשר יש כמה תשובות לשאלה נתונה והמודל בוחר אחת התשובות הנכונות.

המאמר מציע שיטת פרומפטינג המאפשרת להבדיל בין שני סוגי אי-וודאות. מאוד בגדול לשאלה נתונה מזינים למודל תשובות אחרות (לאו דווקא) נכונות לשאלה (other response is…). לאחר מכן בודקים האם ההסתברות של התשובה הנכונה מושפעת מכמות התשובות האחרות המוזנות למודל. אם הסתברות זו מתחילה לרדת זה הסימן שמודל שפה לא כזה ״יודע מה התשובה״ ואי הוודאות האפיסטמית הינה גבוהה.

המאמר גם מציע פריימוורק מתמטי המבוסס על כלים מתורת המידע לאנליזה של אי-הוודאות האלו. נשמע מאמר שווה להתעמק בו.

https://arxiv.org/pdf/2406.02543