⚡️🚀המאמר היומי של מייק 02.07.24:⚡️🚀

From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data

היום סוקרים מאמר קליל שלא דורש כל התעמקות מתמטית אבל עדיין יש בו רעיון נחמד. המאמר מציע גישה מאוד פשוטה לשיפור יכולת של מודל שפה להפיק מידע מטקסט בצורה מדויקת. למשל בהינתן טקסט ארוך המוזן למודל, המודל נדרש לענות נכון על שאלות עליו (הטקסט) בלי קשר לאיפה נמצא פיסת הטקסט הרלוונטית לשאלה. מודלי שפה בד״כ מתקשים במשימה זה בהעדר אימון ייעודי.

שיטת פיינטיון מקובלת לתת למודל טקסטים ארוכים ולאמן אותו לענות על מגוון שאלות בטקסט הזה (למשל לוקחים פסקה לא קשורה, משתילים אותה לטקסט ושואלים אתה המודל לגביה. גישה זו מביאה לשיפור בביצועי המודל במשימה אבל כמה מחקרים הצביעו על כך שבמהלכה המודל למד ״מידע ועובדות מיותרים״ שהרע את יכולת ה-reasoning שלו.

המחברים הציעו שיטה כדי להקל הבעיה זו. הם בנו דאטהסט שהוא הרבה מאוד מילונים שהמפתחות והערכים שבהם הם מספרים. המודל מאומן להפיק נכון ערך של מפתח נתון. משימה יותר קשה להפיק ערך של מפתח מסיום המורכב מכמה מספרים כאשר אני מעבירים את המספרים מהפתח למודל בסדר שונה מאשר הם מופיעים באחד המילונים. היופי כאן שהדאטהסט הזה לא מכיל מידע עובדתי בכלל והמודל לא יכול ללמוד אותו (המידע). ככה מונעים את ״הרעלת המודל״ במידע זר…

https://arxiv.org/abs/2406.19292