⚡️🚀המאמר היומי של מייק 16.07.24: ⚡️🚀

How Does Quantization Affect Multilingual LLMs?

היום נסקור קצרות מאמר שחוקר נושא חשוב לכל מי שעוסק במודלי שפה. הנושא הזה הוא קוונטיזציה או קווינטוט של מודלי שפה שמאפשר לנו גם להקטין את כמות הזכרון הנדרש לאחסון של המודל וגם מזרז את האינפרנס של המודל. אבל כמובן שזה לא בא בלי המחיר והמחיר הוא ביצועיי המודל. המאמר חוקר עד כמה חמורה פגיעה בביצועי המודלי לכמה רמות ושיטות קווינטוט(ניתן לקוונטט שכבות שונות ברמות שונות וגם ניתן לקוונטט משקלי המודל והאקטיבציות ברמות שונות של קווינטוט).

המאמר נכתב על ידי מדעני חברת cohere ובאופן טבעי מתמקד במודלי שלהם. המחברים לקחו מודלים בגדלים שונים ובדקו אותם במספר בנצ'מארקים שונים וגם ביצעו אבלואציה של ביצועי המודלים על ידי בודקים אנושיים. המחברים הגיעו למספר מסקנות מעניינות:

הפגיעה מהקווינטוט הנמדדת על הבנצ'מארקים משמעותית קטנה יותר מזו הנעשית על ידי בודקים אנושיים.

הפגיעה לרוב מחמירה ככל שקווינטוט נהיה יותר קשוח כלומר לפחות ביטים

מודלים גדולים בד״כ עמידים יותר לקווינטוט מאשר מודלים קטנים יותר

מודלים מולטי-שפתיים (multilingual) סובלים יותר מקווינטוט מאשר מודלים חד שפתיים והביצועים על השפות הפחות נפוצות נפגעות יותר מאשר על שפות נפוצות יותר

היכולת של המודלי ל-reasoning (למשל יכולת לפתור שאלות מתמטיות) נפגעת מאוד מהקוויטוט.

יש עוד כמה מציאות מעניינות…

https://arxiv.org/pdf/2407.03211