⚡️🚀המאמר היומי של מייק 16.09.24: ⚡️🚀  
Rethinking Benchmark and Contamination for Language Models with Rephrased Samples

חתיכת נושא זה. לאחרונה אני ניהלתי מספר שיחות עם אנשי NLP לא מעטים על הנושא הזה. מי שעוקב אחריי ברשתות החברתיות אולי שם לב כי אני בד״כ לא מתלהב ממודל שפה שניצח את כל המודלים הקיימים בכל הבנצ'מרקים. הסיבה לכך היא די טבעית ונובעת מכך שבלא מעט מקרים לא מפרסמים באופן גלוי את כל הדאטה שעליה המודל אומן.

כמובן שהחשד שלי הוא הדאטה(משימות) האימון יהיו דומות מדי לאלו שמופיעות בבנצ'מרקים האלו. כמובן אני לא בא להאשים אנשים על כך שהם מרמים בכוונה (למרות שבטח יש מקרים כאלו) אלא אני בא להגיד שזיהוי דוגמאות בדאטהסט הדומות מדי לבנצ'מרקים אינן מצליחות לפלטר את הדוגמאות האלו. והתוצאה היא מודל שהוא אוברפיט על בנצמרק כזה או אחר.

כאמור יש שיטות די בסיסיות הבודקות את הדמיון בין הדוגמאות בדאטהסט לדוגמאות בבנצ'מארק מבוססות על n-grams ועל דמיון סמנטי המחושב באמצעות מרחק בין הייצוגי של הדוגמאות בדאטהסט ובבנצ'מרק. המאמר המסוקר טוען שזה לא מספיק וצריך לעשות בדיקה נוספת לזיהוי של דוגמאות אלו. בגדול המאמר מציע בנוסף לבדיקה הסמנטית לרתום איזה LLM עוצמתי לבדיקה של דמיון דוגמאות.

בגדול מזהים K דוגמאות הכי דומות סמנטית לכל דוגמא בבנצ'מרק ואז מפעילים LLM חזק כמו GPT4 עם איזה פרומפט מתוחכם כדי לזהות את הדוגמאות הבאמת דומות. המאמר מראה כי בצורה כזו הצליחו לתפוס דוגמאות שלמרות שנראות שונה מהוות rephrasing של דוגמא מסוימת מהבנצ'מרק. ואז מעיפים את הדוגמה הזו מהדאטהסט.

המאמר טוען כי ללא שימוש בשיטה שלהם ניתן ״לאמן״ מודל 13B כדי ש״ינצח״ את GPT4 על כל הבנצ'מרקים - נצחון לא אמיתי אמנם.

מאמר ללא יותר מדי חדשנות אך מעלה נושא מאד מעניין

https://arxiv.org/pdf/2311.04850