🚀המאמר היומי של מייק 24.09.24: ⚡️🚀  
LLMs Still can’t Plan; can LRMs? A PRELIMINARY EVALUATION OF OPENAI’S O1 on PLANBENCH

סקירה של מאמר שלא מכיל מתמטיקה בצורה מפורשת…מאמר זה בוחן את יכולות התכנון של מודלי שפה גדולים (LLMs) ומודלי חשיבה גדולים (LRMs) כמו משפחת o1 באמצעות סדרת מבחנים הנקראת PlanBench.

PlanBench הוא מערך מבחנים מקיף שפותח ב-2022 להערכת יכולות התכנון של LLMs. מרכיביו העיקריים:

מערכת סטטית של 600 בעיות Blocksworld הכוללות 3 עד 5 קוביות.

גרסה מוסתרת (Mystery Blocksworld) של אותן בעיות, שבה המונחים והפעולות מוחלפים במילים אקראיות כדי לבחון הבנה מופשטת.

בעיות Blocksworld מורכבות יותר עם 6 עד 20 קוביות, הדורשות תוכניות ארוכות יותר של 20 עד 40 צעדים.

בעיות בלתי פתירות, שנוצרו על ידי הוספת ״יעד״ בלתי אפשרי לבעיות קיימות.

PlanBench נועד להיות כלי גמיש ומקיף להערכת יכולות תכנון של מודלי שפה תוך בחינת היבטים שונים של תכנון כמו הבנה מופשטת, התמודדות עם מורכבות, וזיהוי בעיות בלתי פתירות.

החוקרים מצאו כי LLMs השתפרו בביצועי תכנון בסיסיים, כאשר המודל הטוב ביותר, LLaMA 3.1 405B, השיג דיוק של 62.5% במשימות Blocksworld פשוטות. עם זאת, LLMs נכשלו במשימות בעלי פתרון סבוך יותר.

לעומת זאת, מודל ה-LRM החדש של OpenAI, o1, הציג שיפור משמעותי, עם דיוק של כמעט 98% במשימות Blocksworld פשוטות ו-52.8% במשימות עם פתרון סבוך. למרות זאת, הביצועים של o1 ירדו משמעותית במשימות מורכבות יותר ובבעיות בלתי פתירות.

עם זאת החוקרים מדגישים את החשיבות של הטרייד-אופים הכוללים יעילות, עלות וערבויות לנכונות הפתרון (ככה כתוב במאמר) בהערכת מודלים אלה. הם מציינים כי o1 יקר משמעותית להפעלה ואינו מספק ערבויות לנכונות, בניגוד למתכנני AI קלאסיים. המסקנה היא שבעוד LRMs כמו o1 מציגים התקדמות, הם עדיין רחוקים

מלהיות פתרון כללי ואמין לבעיות תכנון.

https://arxiv.org/abs/2409.13373