⚡️🚀המאמר היומי של מייק 12.09.24: ⚡️🚀

Learning to reason with LLMs

היום במקום הסקירה אשתף איתכם את מחשבותיי על המודל החדש של openai שקיבל שם o1. אני בדרך כלל נמנע מלהגיב ולכתוב פוסטים על כל מודל חדש שמנצח את כל ה-benchmarks בעולם אבל הפעם אחרוג ממנהגי. ולא מהסיבה שמהמודל הזה השאיר אבק לרוב ה-benchmarks אלא בגלל שאני זיהיתי כאן שינוי מסוים בפרדיגמה בעולם ה-llms.

השינוי בפרדיגמה בא בדמות של שינוי היחס בכמות הקומפיט המוקדש ללמידה ולהסקה (אינפרנס). אנחנו רגילים למודל שמצריכים כמות אדירה של קומפיוט במהלך הלמידה (אימון מקדים, SFT, יישור המודל וכדומה) כאשר האינפרנס הוא די זול (כמובן יחסית לאימון כי גם בהסקה יש עלויות די גבוהות בשל עצמם). O1 לעומות זאת מאתגר את ההנחה הזו ושואל את השאלה: האם זה אופטימלי? אולי אנו צריכים לאמן את המודל שלנו פחות ולהשקיע יותר קומפיט בהסקה.

לפני כמה זמן סקרתי מאמר שדי שינה (או לכל הפחות רענן) את תפיסתי בעניין זה (Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters). המאמר הזה היה של deepmind אולם הייתה לי תחושה שהם לא היחידים שהגיעו לתובנה הדי לא טריוויאלית הזה.

בעקרון הכל מסתכם לשתי הנקודות הבאות:

אולי אתה לא צריך מודל שפה ענק להסקה. חלק ניכר מהפרמטרים כנראה ממשמשים לאחסון עובדות, כדי שהמודל לא ידבר שטויות לשאלות לידע כללי (כמו מתי נולד מוצרט). לדעתי ניתן להפריד בין הסקה לידע, כלומר אפשר להסתפק ב"ליבה להסקה" קטנה שיודעת איך להשתמש בכלים כמו וולפרם, בראוזר ובודק קוד כלומר המשימות הדורשות סוג של ידע עובדתי (ידע בשפת תכנות). ככה ניתן להפחית את כמות החישוב המוקדשת לאימון המוקדם.

כמות משמעותית של קומפיט מועברת להסקה בזמן הרצת המודל ולא לאימון המודל. ניתן לחשוב על מודלי שפה בתור סימולטורים מבוססי טקסט. על ידי הרצת תרחישים ואסטרטגיות רבות (גנרוט טקסט), המודל יגיע בסופו של דבר לפתרונות reasoning טובים. התהליך בחירת הפתרון נראה די דומה לבעיות שנחקרו היטב כמו חיפוש העץ של מונטה קרלו (MCTS) ב-AlphaGo.

כמובן שאם יש שימוש בטכניקות כמו MCTS אנו צריכים את פונקציית ה-reward. בניית פונקצייה כזו היא לא טריוויאלית כאן כי אין לנו דרך טובה (אלא אם כן יש לנו דאטהסט reasoning מגוון ועצום שניתן לאמן עליו מודל כזה) לשערך את איכות ה-reasoning. כמובן שניתן לנצל מודלי שפה אחרים, בדיקות עצמיות על ידי מודלי שפה וכדומה אבל עדיין לא ברור ב-100% איך לעשות את זה (ד״א אני בכלל לא בטוח שהם השתמשו ב-mcts). אולי הם פיתחו שיטה מגניבה לעקוף את ה-reward כמו שנעשה ב-dpo וב-orpo שעשו זאת עבור ppo -אין לדעת.

בקיצור מחכה לדוח הטכני שבתקווה ישפוך אור על הסיפור הזה (גם בזה אני לא בטוח בכלל)....

https://openai.com/index/learning-to-reason-with-llms/