**לא כל “חשיבה ארוכה” היא חשיבה טובה.**

**המאמר היומי של נורית ומייק: 14.10.25, סקירה 526  
Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and Correctness in LLMs**

במודלי שפה אנחנו נוטים לשלם בטוקנים מתוך אמונה שיותר חשיבה תוביל לדיוק גבוה יותר. בפועל, במאמר שנסקור הפעם, מתגלה תבנית עקבית: יש אורך חשיבה אופטימלי (טוקנים). קצר מדי - המודל מפספס. ארוך מדי - המודל מתחיל להסתבך בעצמו. התוצאה נראית כמו גבעה: דיוק עולה… ואז יורד.

המאמר האמפירי שנסקור היום בוחן באופן שיטתי את הקשר בין אורך החשיבה לנכונות התשובה. אז איך בנו את הניסוי?

בדקו שני מודלי Reasoning קטנים (1.5B~ פרמטרים) על שתי משימות דאטהסטים מתמטיים (GSM8K ו- MATH). לכל שאלה דוגמים 10 תשובות עם תנאי הרצה זהים (כגון טמפרטורה) קבועים ומנתחים את הדיוק כתלות באורך של הטוקנים. בנוסף, הם תייגו את רמת הקושי של השאלות לפי מידת ההצלחה בתשובות והציגו ממצאים.

מה יצא בניסויים (כאמור, זהו מאמר אמפירי):

* עקומה לא-מונוטונית: הדיוק מטפס עם האורך עד נקודה מסוימת, ואז יורד (תופעת overthinking).
* וגם ההיפך קורה (underthinking): כאשר שרשרת החשיבה קצרה מדי המודל מפספס שלבי ביניים הכרחיים, במיוחד בבעיות קשות, ונותן תשובה בטוחה אך שגויה כי לא הוקצב לו “אורך נשימה” מספק.
* תשובות שגויות נוטות להיות ארוכות יותר מתשובות נכונות. התשובה הנכונה לעיתים קרובות מופיעה כבר בדגימות הקצרות.
* “התשובה הנכונה הראשונה” לעיתים קרובות קצרה: עבור חלק גדול מהשאלות, אחת התשובות הקצרות ביותר כבר נכונה. הארכת השרשור לאחר נקודה זו מגדילה את הסיכון להחמרת טעויות.
* במדידה למול קושי: המודלים “מושכים” קצת אורך כשקשה להם יחסית בשאלות קלות או בינוניות, אבל בשאלות קשות במיוחד, אין רמזים למהו אורך ״נכון״, המודל אינו מצליח לאפטם את עצמו, וכאמור, הדיוק יורד.
* ומה לגבי הביטחון של המודל? כששאלות “קלות” יותר למודל, הפרפלקסיטי יורד (יותר ביטחון). בשאלות הקשות מאוד, ירידה עקבית לא תמיד נצפית, עוד אינדיקציה לכך שהארכת החשיבה לא בהכרח משפרת דיוק.

מה כן עובד (פרקטיקה שאנו מאמצים בעבודה עם המודל):

* קצר-קודם: להתחיל בתשובה תמציתית, ולהאריך רק לפי צורך אמיתי.
* עצירה מודעת-אורך: להחליש או לפסול שרשראות חריגות באורכן בזמן self-consistency.

יש למאמר הזה כמה חולשות, הראשונה זה המודלים הקטנים מדי שנבחנו, והדאטהסטים המוגבלים. השניה, האופן שבו הם ״תייגו״ קושי של בעיות, יש בזה הטייה מובנית שמעידה כבר על התוצאה. ובכל זאת, המסר המרכזי של המאמר: לא לשלם בטוקנים “על אוטומט”. לשלם על היגיון נקי, אז להאריך רק כשזה מוסיף ערך.

arxiv.org/abs/2505.00127