**האם ה-"Trade-Off" בין חיפוש לניצול הוא רק בעיית מדידה?, סקירה 520**

**סקירת המאמר היומית של מייק: 06.10.25  
BEYOND THE EXPLORATION-EXPLOITATION TRADE-OFF: A HIDDEN STATE APPROACH FOR LLM REASONING IN RLVR**

הסיפור של למיד חיזוק עבור תגמולים ודאיים (RLVR) מוסגר כאקט של איזון הרואי. המודל חייב לנווט על חבל דק בין *חיפוש* (exploration) – איתור נתיבי הסקה חדשים ומגוונים, לבין *ניצול* (exploitation), התבססות על אסטרטגיות מוכרות ואמינות. נרטיב זה בנוי כמעט כולו על מדדים מרמת הטוקן, שם חיפוש מזוהה עם חיזוי טוקנים בעלות אנטרופיה גבוהה, וניצול עם חיזויים בעלי אנטרופיה נמוכה וביטחון גבוה. ההנחה הנובעת מכך היא אינטואיטיבית: מודל אינו יכול להיות ודאי ולא ודאי בו-זמנית, ולכן ה-trade-off הוא בלתי נמנע.

המאמר שואל שאלה פרובוקטיבית: מה אם הטרייד-אוף הבסיסי הזה אינו חוק יסוד של הסקת מסקנות, אלא פשוט תוצר לוואי של התבוננות במקום הלא נכון? מה אם, על ידי הזזת העדשה האנליטית שלנו מהשכבה השטחית של הטוקנים אל הגיאומטריה הסמנטית העמוקה של הייצוגים הפנימיים (internal representations) של המודל, הטרייד-אוף פשוט נעלם?

החידוש המרכזי של המחברים הוא שינוי פרספקטיבה זה. במקום לשאול כיצד נראית התפלגות הטוקנים בשכבה הסופית של המודל, הם שואלים כיצד נראית ה*גיאומטריה* של "תהליך החשיבה" הפנימי שלו. כדי לעשות זאת, הם מציגים סט כלים חדש לניתוח מטריצת המצבים החבויים, הייצוג המספרי של מסלול ההסקה של המודל.

ראשית הם מציגים את **הדרגה האפקטיבית (Effective Rank - ER)**. טכנית, ER היא הפרפלקסיות (האקספוננט של אנטרופיית שאנון) של הערכים הסינגולריים המנורמלים של הייצוג הפנימי של שרשרת חשיבה. באופן אינטואיטיבי יותר, דמיינו את המצבים החבויים של המודל כ"מרחב קונספטואלי" עצום ורב-ממדי. דרגה קונבנציונלית הייתה פשוט סופרת את המספר הכולל של המימדים, או המושגים, הזמינים במרחב זה. ER הוא מדד מתוחכם בהרבה. הוא מודד את המספר ה*אפקטיבי* של הממדים שהמודל משתמש בהם בפועל במהלך תהליך הסקה ספציפי. ER גבוה פירושו שהמודל ממנף סט עשיר ומגוון של תכונות פנימיות, ופורס את "מחשבתו" על פני מושגים רבים ושונים באופן מאוזן. זוהי חתימה ישירה להתנהגות חיפוש אמיתית. לעומת זאת, ER נמוך מסמל מצב פנימי "שקרס", בו המודל מסתמך על מושגי ליבה מעטים בלבד, גם אם רבים אחרים זמינים.

שנית, כדי למדוד ניצול, הם מציעים את **מהירות הדרגה האפקטיבית (Effective Rank Velocity - ERV)**, מדד זה לוכד את הדינמיקה של צבירת מידע. אם ER מודד את ה*נפח* של המרחב הקונספטואלי בתשובת המודל, ואילו ERV מודד את ה*קצב* שבו הנפח משתנה לאורך זמן. ERV גבוה מציין שבכל שלב חדש, המודל מוסיף בהצלחה מורכבות חדשה ומשמעותית לייצוג הפנימי שלו, בקצב העולה על הממוצע ההיסטורי שלו עצמו. זהו סימן לתהליך הסקה פרודוקטיבי ומואץ. ERV נמוך, לעומת זאת, מעיד על כך שהנתיב הנוכחי הופך פחות יעיל, והמודל מגיע לנקודה של תפוקה שולית פוחתת.

חמושים בכלים מתוחכמים יותר אלה, המחברים מוצאים משהו מדהים: ברמת המצב החבוי, לחיפוש (ER) ולניצול (ERV) יש מתאם קרוב לאפס. הטרייד-אוף נעלם. הדבר מרמז שבהחלט ייתכן שמודל ירחיב את ארגז הכלים הקונספטואלי שלו ובמקביל ישפר את יכולתו להשתמש בו. תובנה זו מולידה את השיטה שלהם (Velocity-Exploiting Rank-Learning (VERL אינו עוד רגולריזטור המנסה באיטיות לאזן בין שני כוחות מנוגדים. במקום זאת, הוא פועל כמכוונן אדפטיבי המעצב ישירות את פונקציית היתרון של למידת החיזוק כדי לשפר באופן יעיל את שתי היכולות.

החידוש המרכזי הוא השימוש בנגזרת מסדר שני(ERV היא נגזרת ראשונה), תאוצת הדרגה האפקטיבית (ERA), כמטא-בקר (meta-controller) של החיזוי. מכיוון ש-ERA מספק אות יציב לגבי האם תהליך ההסקה מאיץ או קופא על שמריו, VERL יכול להשתמש בו כדי ליצור מבנה תמריצים דו-ערוצי. אם ERA מאותת שהמודל הופך בטוח יתר על המידה ומסתכן באוברפיט, VERL מגביר את התגמול עבור ה-exploration על ידי העלאת ER. אם ERA מאותת שתהליך ההסקה מאבד מומנטום(כלומר יותר מדי ״מתפרע״ מבחינת אנטרופיה), הוא מחזק את רווחי הניצול (ERV) כדי לגבש את מה שהמודל כבר הבין. כל זה נעשה דרך כוונון של פונקציית היתרון (advantage) על ידי משקול של ER ו-ERV.

החידוש כאן הוא עמוק. VERL מתקדם מעבר למשיכת החבל ברמת הטוקן, ובמקום זאת פועל לטיפוח גיאומטריית הסקה פנימית חזקה יותר. על ידי הצעת דרך חדשה למדוד ולהבין מה מודלי שפה גדולים עושים כשהם "חושבים", עבודה זו מרמזת שהאילוצים המקובלים ביותר בתחום עשויים להיות אשליות שיצרנו בעצמנו.

[https://arxiv.org/abs/2509.23808](https://www.google.com/search?q=https://arxiv.org/abs/2509.23808)