# Review 112: [Short] Skeleton-of-Thought: Large Language Models Can Do Parallel Decoding, 31.07.23 https://huggingface.co/papers/2307.15337

**Paper: https://arxiv.org/abs/2307.15337v3**

אתם בטח יודעים שמודלי שפה יוצרים טקסט טוקן לאחר טוקן. כלומר כדי ליצור את המילה השלישית אנו צריכים לגנרט את המילה הראשונה ואת השנייה. גנרוט סדרתי שכזה כמובן מאט את גנרוט הטקסט על ידי מודלי שפה ואנו נאלצים לחכות יותר זמן בשביל לקבל את התשובה. נשאלת השאלה האם ניתן להאיץ את הגנרוט?  
  
היום ב- #shorthebrewpapereviews אנו סוקרים מאמר שמצליח לזרז את גנרוט הטקסט על יד מודלי שפה באמצעות טריק מאוד פשוט ואלגנטי. במקום ליצור את כל התשובה יחד מודל שפה קודם כל יוצר את הסקיצה של התשובה (נגיד בתור רשימת נושאים) ואז מעביר את השרביט לכמה מודלי שפה שכל אחד מהם מגנרט תשובה עבור כל אחד מהנושאים שנוצרו.

למשל אם אתם שואלים מודל שפה על מאכלים סיניים אז בשלב הראשון הוא יוצר רשימת של שמות המאכלים בלבד ובשלב השני כמה (רפליקות) מודלי שפה מרחיבות על כל סוג של מאכל. כמו שאתם כבר מבינים ככה ניתן למקבל את גנרוט הטקסט כי יצירת סקיצה קצרה אמורה לקחת מעט מאוד זמן. כמובן יש כמה טריקים איך לגרום למודל שפה ליצור סקיצת תשובה קצרה (אחד מהם הוא פשוט להגיד לו ליצור רשימה של תשובות קצרות). עם הטריק הפשוט זה המחברים הצליחו לזרז גנרוט עד פי 2.39!