# Adversarial NLI: A New Benchmark for Natural Language Understanding

## הרעיון המרכזי

המאמר מנסה לענות על בעיה קריטית בעולם ה – NLU, מודלים מגיעים לציונים גבוהים על benchmarks , אבל בפועל:

- לא באמת מבינים שפה
- לומדים קיצורי דרך יוריסטיים
- מנצלים biases סטטיסטיים בדאטה

חוקרי המאמר טענו כי ה – benchmarks הסטטיים "נשברים" מהר מדי, כי המודלים לומדים את הטריקים של הדאטה.

## הרעיון החדש Adversarial Data Collection

במקום לאסוף דאטה רגילה, הם עשו : Human-And-Model-in-the-Loop (HAMLET)
כלומר, בני אדם ניסו בכוונה "להפיל" את המודל.

### אופן הפעולה :

1. אימנו מודל NLI על דאטה מתאימה (לדוגמה, על SNLI)
2. לאחר האימון, הנבדק קיבל :
   a. **Context** (Premise)
   b. **Target Label** (contradiction / entailment / neutral)
3. המשתתף האנושי כתב Hypothesis, כך שבינה לבין ה – Premise יש קשר של Target Label
4. שלחו את (Premise, Hypothesis) למודל ה – NLI, וקיבלו את תוצאת הסיווג שלו.
   a. אם המודל צדק – הדוגמה (premise, Hypothesis) לא נכנסה לדאטהסט, ועוברים לדוגמה חדשה.
   b. אם המודל טעה - הדוגמה עוברת verification אנושי נוסף. במטרה לוודא שאין בעיה לוגית בהיפותזה, שגרמה למודל ה – NLI לטעות (וידאו שאין Bias).
      i. בהנחה וה – Hypothesis עברה את ה - Verification- הדוגמאות החדשות שהמודל טעה בהן מתווספות ל – training set שעליו המודל התאמן בשלב 1
5. חוזרים לשלב 1 עם ה – Training Set המעודכן.
   a. חוזרים על התהליך באופן איטרטיבי – במחקר ביצעו זאת ב – 3 איטרציות.

**המטרה -** לייצר benchmark "עוין" (Adversarial) שמתפתח יחד עם המודלים, ולא נשאר סטטי. כזה שנבנה במיוחד כדי לחשוף את החולשות האמיתיות של מודלי ה - NLI

## למה זה חשוב

- בדאטה רגיל - המודל רואה patterns קבועים.
- בדאטה עוין (Adversarial) – בני האדם מחפשים חולשות, יוריסטיקות, קיצורי דרך, הטיות. ולכן, המודל חייב להבין טוב יותר, ולא רק "לנחש".

## ממצאי המאמר

מודלים שנבחנו, כמו BERT, RoBERTa, XLNet – הצליחו על דאטה סטטית "רגילה" (כמו SNLI, MNLI, GLUE), אך **נפלו באופן משמעותי** על הדאטה של ANLI.
החוקרים זיהו מספר סוגי reasoning ו - language phenomena (תופעות שפה) שגרמו למודלים להיות שבירים במיוחד:

- **numerical reasoning** - reasoning המבוסס על מספרים וכמויות.
  **לדוגמה**
  - "less than 5" למול "more than 10"

- **Negation** - שימוש בשלילה שמשנה את משמעות המשפט.
  **לדוגמה**
  - "He is running" למול "He is not running"

- **lexical tricks** - ניסוחים מבלבלים המבוססים על מילים דומות או patterns שטחיים.

- **Presuppositions** - מידע שמונח כ"ברור מאליו" בתוך המשפט.
  **לדוגמה:**
  - "The king stopped smoking" ← מניח שהמלך עישן בעבר.

- **Coreference** - הבנה למי מתייחס כינוי גוף.
  **לדוגמה:**
  - "Dan told Tom that he won."

- **word overlap traps** - משפטים עם מילים משותפות רבות, אך משמעות שונה או סותרת.

- **commonsense reasoning** - צורך בידע עולם בסיסי והיגיון אנושי.
  **לדוגמה**
  - אם מישהו נרטב בגשם, אז כנראה ירד גשם.

- **Ambiguity** - עמימות במשמעות המשפט.

- **pragmatic reasoning** - הבנת משמעות לפי הקשר וכוונת הדובר, ולא רק לפי המילים עצמן.

**לסיכום**, המאמר מראה ש:
High benchmark score ≠ Real language understanding

וזו אחת הטענות המרכזיות ביותר בעולם ה - NLU המודרני.