Stress Test Evaluation for Natural Language Inference

הרעיון המרכזי

המאמר טוען ש - accuracy גבוה על datasets כמו SNLI ו – MNLI לא באמת מוכיח שהמודל מבין שפה.

לכן החוקרים מציעים מבחן שנקרא - **Stress Tests**
מבחנים שמטרתם לבדוק איך מודלי NLI מתמודדים עם תופעות לשוניות קשות, מבלבלות, או מטעות .

הבעיה שהמאמר מזהה

מודלים מצליחים על benchmarks רגילים, אבל ייתכן שהם פשוט:

- מזהים patterns
- מסתמכים על lexical overlap
- או משתמשים ב - shortcuts סטטיסטיים

במקום לבצע reasoning אמיתי.

מתודולוגיה

הם בנו סדרת Stress Tests אוטומטיים, כאשר כל test בודק חולשה אחרת של המודל.
הם חילקו את המבחנים ל - 3 קבוצות:

- **Competence Tests**
  - מבחנים הבודקים האם המודל באמת יודע לבצע reasoning
  - למשל:
    - **Antonyms** – בדיקה האם המודל מבין שמילים מסוימות הן הפכים במשמעות.
      - love ↔ hate
    - **numerical reasoning** - בדיקה האם המודל יודע להבין ולהשוות כמויות ומספרים בצורה לוגית.
      - "less than 5" מול "more than 10".

- **Distraction Tests**
  - מבחנים הבודקים האם המודל נופל בגלל "הסחות דעת" שטחיות.
  - למשל:
    - **Overlap** - גבוה בין מילים.
      - לדגמה, הוספת מילים שלא מוסיפות למשמעות של המשפט (אך גם לא פוגעות בו).
    - **מילת שלילה** כמו - not
      - למשל, במקום לכתוב true, ירשמו not false – במטרה לבחון האם המודל "נלחץ" ממילת השלילה, גם כשאינה משנה את המשמעות.
    - **משפט ארוך ומבלבל**

  - **Noise Tests**
    - מבחנים הבודקים robustness לרעש בטקסט.
    - למשל:
      - **שגיאות כתיב**
      - **החלפת אותיות**

הממצא המרכזי
כל המודלים חוו ירידה משמעותית בביצועים על ה - Stress Tests, למרות שהם הצליחו מאוד על דאטה סט ה MNLI

ממצא חשוב נוסף - המודלים נטו מאוד לחזות entailment, כאשר היה Lexical Overlap גבוה – גם אם זה היה שגוי לוגית.
זה חיזק את הטענה שהם משתמשים ב - heuristics שטחיים, ולא reasoning אמיתי.

המסר המרכזי של המאמר
מודלי NLI יכולים להיראות חזקים מאוד על benchmarks רגילים, אבל להיכשל ברגע שמפעילים עליהם "מבחני לחץ" שבודקים reasoning אמיתי ו – robustness.