



AI in Education

אבחן תפיסות שגויות ומשוב אוטומטי

מאת: נתן בר, תומר בנג'יב, אוראל כהן

מקרה שימוש והאתגר

רקע

- כוים מערכות חינוך נאלצות לבחור בין **Scalability** (שאלות אמריקאיות קלות לבדיקה אר שטחות) לבין **Quality** (שאלות פתוחות שבודקות הבנה אר דורות בדיקה ידנית יקרה).
- בקורסים, סטודנטים מקבלים ציון "יבש" ללא משוב עללמה הם טעו, ובכך הלמידה נפגעה.

למה זה חשוב?

- **ערך פדגוגי عمוק:** המערכת לא רק נותנת ציון, אלא מספקת **משוב מעצב**. הסטודנט לומד למה הוא טעה ומהי התפיסה השגניה שלו, מה שմשפר את תהליכי הלמידה.
- **סקלibilitות בהוראה:** פתרון לצורכי הבקבוק של בדיקה ידנית. המערכת מאפשרת לתת יחס אישי וניתוח עמוק של תשובה פתוחות גם בקורסים המוניים או כיתות ענק, ללא עומס נוסף על הסגל.

פתרונות כיוון

- **בדיקה ידנית:** הפתרון האיכותי ביותר, אך הוא איטי מאוד, יקר ולא ניתן ליישום בקורסים המוניים (MOOCs).
- **שאלות אמריקאיות (MCQ):** פתרון קל לבדיקה אוטומטית, אך בודק בעיקר זיכרון וניחוש ולא הבנה عمוקה או יכולת ניסוח.



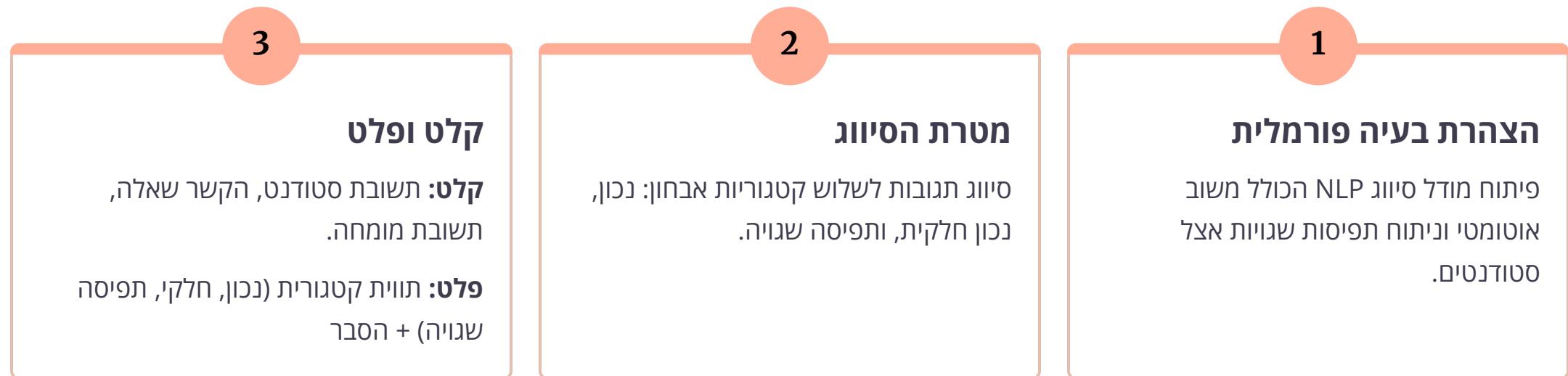
האתגר לפתרון הבעיה

סובייקטיביות בתיאוג (Label Noise): אפילו מומחים אנושיים לא תמיד מסכימים איפה עובר הגבול בין "תשובה חלקיים" ל"טעות". הרעש הזה ב-Ground Truth מקשה מאוד על המודל להתכנס לפתרון יציב.

הונב הארוך של הטעויות: יש רק דרך אחת להיות צודק, אבל אין סוף דרכי לטעות. קשה לאמן מודל כ שיש מעט דוגמאות לכל סוג ספציפי של טעות.



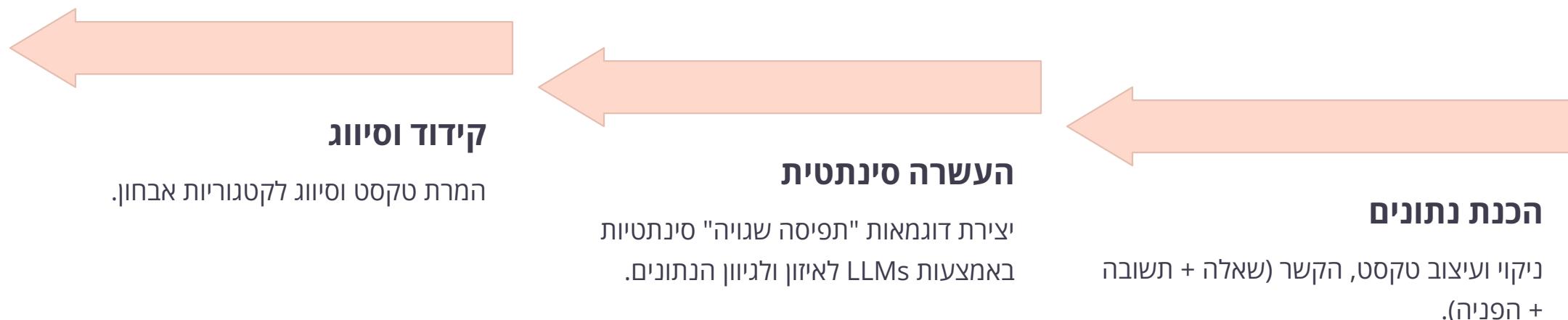
תיאור משימת הפרויקט: מודל סיווג NLP



חידוש הפרויקט

- ניתוח סמנטי عمוק:** מעבר להתקנת מילוט מפתח לזיהוי פערים מושגיים.
- אסטרטגיית נתונים סינטטיים:** טיפול במחסור נתונים וחוסר איזון באמצעות "טעויות סטודנטים" סינטטיות עם AI-Gen.
- אימון על "דוגמאות מכשילות":** אנו משתמשים ב-AI-Gen כדי ליצור בכונה **Hard Negatives** – תשובה שנראות נכוןות (משתמשות במונחים המקבילים ובמבנה תקין), אך מכילות כשל לוגי עדין

זרימת העבודה



מודלים ושיטות



כונון וטיפול בנתונים

- כונון (Fine-tuning):** ביצוע אימון נוספת למודל השפה הנוכחי על גבי מאגר הנתונים הייעודי שלנו (המשלב נתונים אמיתיים וסינטטיים), כדי להתאים לנויאנסים של שפה מדעית וחינוכית.
- איזון מחלקות:** התמודדות עם הפער הכמותי בין תשובות נכונות (שהן הרוב) לתשובות שגויות, באמצעות מתן משקל גובה יותר לטיעויות בזמן האימון כדי שהמודל לא יתעלם מהן.
- coil ואופטימיזציה:** התאמת מודיקת של קצב הלמידה ומונעת "התאמת יתר", כדי להבטיח שהמודול ידע לנתח נכון גם תשובות לשאלות חדשות שלא ראה בעבר.



נתונים: דרישות, מקורות ודור סינטטי

דרישות נתונים

נדרשת חלוקת נתונים (אימון/אימוט/בדיקה) עם מחלקות מאוזנות. האתגר הוא שננתונים חינוכיים בעולם האמיתי אינם מאוזנים.

מערכות נתונים

- **נתונים אמיתיים:** SciEntsBank (~10k תשובות סטודנטים מתוינגות על ידי מומחים).
- **נתונים מוגדלים:** מערכת נתונים סינטטי מותאם אישי (~2k-5k דוגמאות) המתמקד בתפיסות שונות קשות ליהוו.

אסטרטגיית תיוג

- **נתונים אמיתיים:** שימוש בתוויות מומחים קיימות.
- **נתונים סינטטיים:** תיוג מרומז (התווית ידועה מראש מההנחיה).
- **אימוט:** בדיקה ידנית של 5% מהד נתונים הסינטטיות.

דור נתונים סינטטי

יצירת "תשובות סטודנטים" ספציפיות באמצעות Llama-3 / GPT-4o עם יצירה מבוססת תוכנות, המדמה סטודנטים עם פורי ידע ספציפיים.

הערכתה ומדידה

מדד הערכתה ראשוניים

- **ציון F:** המודד העיקרי שלנו. מכיוון שמערכת הנתונים אינה מדויקת (פחות "תפיסות מוטעות" מתשובות "נכונות"), הדיקט מטעה. F1 מבטיח לנו מזהים היטב את קבוצות המידע.
- **מטריצת בלבול:** לניטוח סוגי שניאות ספציפיים (למשל, הבחנה בין הבנה חילונית לתפיסות מוטעות עמוקות).



מדידות אינכות נתונים (שלב אחר שלב)

aicots נתונים סינטטיים:

- **דמיון סמנטי:** מדידת מרחק הטמעה בין תפיסות מוטעות שנוצרו בין שניאות אמיתיות של תלמידים (m-SciEntsBank) כדי להבטיח ריאליות.
- **אימות אנושי:** פרוטוקול "אנושי בלולאה" שבו אנו מאמתים ידנית מדגם אקלראי (למשל, 5% מהנתונים שנוצרו).



פרוטוקולי מדידה

- **נתונים אמיתיים:** הערות מומחחים המסופקות על ידי מערכת הנתונים של SciEntsBank.
- **נתונים סינטטיים:** התווית מוגדרת על ידי בקשת היצירה (מאומתת על ידי שלב האימות).

קווי בסיס להשוואה:

- **גבול תחתון:** F1-IDF + רגסיה לוגיסטיבית (קו בסיס פשוט).
- **גבול עליון:** GPT-4o-Zero-Shot (כדי לראות אם כוונון של מודלים קטנים יותר יכול להתחזר בדגמי LLM ענקיים).

