**תמחור**   
התאם למידע שסיפקת, יש לך 2,986 מילים ב-23 פוסטים. בהנחה שמילה מכילה בממוצע 5 תווים, זה מסתכם לכ-14,930 תווים. בהמרה לטוקנים, כאשר כל טוקן מכיל בממוצע 4 תווים, מתקבלים כ-3,733 טוקנים.

מודל GPT-4o mini מתומחר ב-0.15 דולר למיליון טוקני קלט ו-0.60 דולר למיליון טוקני פלט.

[OpenAI](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/?utm_source=chatgpt.com)

בהנחה שהקלט והפלט מכילים כל אחד 3,733 טוקנים, החישוב הוא:

* **עלות קלט**: (3,733 / 1,000,000) \* 0.15 דולר ≈ 0.00056 דולר
* **עלות פלט**: (3,733 / 1,000,000) \* 0.60 דולר ≈ 0.00224 דולר

**סה"כ עלות משוערת**: 0.00056 + 0.00224 = 0.0028 דולר

כלומר, העלות הכוללת לעיבוד 23 הפוסטים היא כ-0.0028 דולר, שהם פחות מאגורה אחת.

שים לב שהחישוב מבוסס על הנחות כלליות. לקבלת הערכה מדויקת יותר, מומלץ להשתמש בכלים המחשבים את מספר הטוקנים בפועל עבור הקלט והפלט.

**תוצאות חיזוי**   
  
  
**Detailed Report on Crowdsourcing Results**

**Overview**

This report evaluates the labeling performance of human raters (Danielle and Hadas) and compares it to the outputs of a machine learning model (LLM). The evaluation involves 23 posts annotated across three metrics:

1. **Stress Intensity**: A measure of the intensity of stress described in a post.
2. **Emotional Overload**: A measure of the mental exhaustion or emotional overload in a post.
3. **Pregnancy Week**: Identification of the pregnancy week mentioned in a post.

For each metric, the report explores the level of agreement between human raters and between the human average and the model. Several statistical methods were used to assess these agreements.

**Methods**

**Questions Asked:**

1. **Stress Intensity**:
   * *Question for Humans*: "Rate the intensity of the stress described in the post on a scale from 1 to 10, where 1 = no stress, 10 = extreme stress, and -1 = no mention of stress."
   * *Model Prediction*: Predicted intensity of stress on the same scale.
2. **Emotional Overload**:
   * *Question for Humans*: "Evaluate the level of emotional overload or mental exhaustion described in the post on a scale from 1 to 10, where 1 = no overload, 10 = extreme overload, and -1 = no mention of overload."
   * *Model Prediction*: Predicted emotional overload on the same scale.
3. **Pregnancy Week**:
   * *Question for Humans*: "Identify the week of pregnancy mentioned in the post as a number (e.g., 20 for week 20). If no week is mentioned, return -1."
   * *Model Prediction*: Predicted pregnancy week or -1 if none is mentioned.

**Statistical Metrics:**

1. **Pearson Correlation**: Measures the linear correlation between two sets of ratings.
2. **Spearman Correlation**: Measures the monotonic relationship between two sets of ratings.
3. **Intraclass Correlation Coefficient (ICC)**:
   * *ICC (Humans and Model)*: Agreement between both human raters and the model.
   * *ICC (Humans Only)*: Agreement between the two human raters.
4. **Scatter Plots**: Visualizations to assess relationships between human raters and model predictions.

**Results**

**1. Stress Intensity**

* **Pearson (Humans)**: 0.918 — Strong linear agreement between Danielle and Hadas.
* **Spearman (Humans)**: 0.925 — Strong monotonic agreement between human raters.
* **Pearson (Humans vs Model)**: 0.812 — Moderate agreement between the human average and the model.
* **Spearman (Humans vs Model)**: 0.840 — Moderate monotonic agreement with the model.
* **ICC (Humans and Model)**: 0.245 — Low agreement between human raters and the model.
* **ICC (Humans Only)**: 0.048 — Very low agreement between Danielle and Hadas.

**2. Emotional Overload**

* **Pearson (Humans)**: 0.920 — Strong linear agreement between Danielle and Hadas.
* **Spearman (Humans)**: 0.930 — Strong monotonic agreement between human raters.
* **Pearson (Humans vs Model)**: 0.750 — Moderate agreement between the human average and the model.
* **Spearman (Humans vs Model)**: 0.781 — Moderate monotonic agreement with the model.
* **ICC (Humans and Model)**: 0.280 — Low agreement between human raters and the model.
* **ICC (Humans Only)**: 0.014 — Very low agreement between Danielle and Hadas.

**3. Pregnancy Week**

* **Pearson (Humans)**: 1.000 — Perfect linear agreement between Danielle and Hadas.
* **Spearman (Humans)**: 0.999 — Near-perfect monotonic agreement between human raters.
* **Pearson (Humans vs Model)**: 0.768 — Moderate agreement between the human average and the model.
* **Spearman (Humans vs Model)**: 0.743 — Moderate monotonic agreement with the model.
* **ICC (Humans and Model)**: -0.475 — Negative agreement, indicating significant deviation between human raters and the model.
* **ICC (Humans Only)**: -0.276 — Negative agreement between Danielle and Hadas.

**Discussion**

**Why These Metrics Were Chosen:**

1. **Pearson and Spearman Correlations**:
   * These metrics evaluate different aspects of relationships: linear (Pearson) and monotonic (Spearman). They provide a robust analysis of agreement even if the data deviates from linearity.
2. **ICC**:
   * Intraclass Correlation Coefficient measures the consistency and reliability of ratings among multiple raters.
   * The inclusion of ICC helps assess the degree to which the model aligns with human annotations, which is critical for understanding model performance in real-world contexts.

**Insights:**

1. **Stress Intensity and Emotional Overload**:
   * Human raters showed strong agreement (high Pearson and Spearman correlations), indicating consistent labeling.
   * Moderate agreement with the model suggests that the LLM captures some patterns but struggles to fully replicate human judgment.
   * Low ICC values imply variability in ratings, especially when including the model.
2. **Pregnancy Week**:
   * Perfect agreement between human raters reflects the objective nature of this task.
   * Moderate model agreement suggests that while the model performs reasonably well, there is room for improvement in identifying pregnancy weeks.
   * Negative ICC values indicate disagreement between the model and humans, particularly in edge cases where no week is mentioned.

**Visualizations**

* Scatter plots have been generated to visualize the relationships:
  1. Human vs. Human.
  2. Human average vs. Model.
* These plots provide a clear depiction of where agreements and disagreements lie.

**Conclusion**

1. **Human Raters**:
   * Danielle and Hadas demonstrated strong agreement in subjective metrics (Stress Intensity, Emotional Overload).
   * Perfect agreement was achieved in the objective metric (Pregnancy Week).
2. **Model Performance**:
   * Moderate agreement with the human average highlights areas where the model aligns with human intuition but also exposes its limitations.
3. **Recommendations**:
   * Refine the model using more diverse and annotated datasets.
   * Address edge cases explicitly (e.g., posts with no explicit mention of stress or pregnancy week).
   * Consider additional metrics for evaluation, such as Bland-Altman plots, to assess systemic biases.

This analysis sets a solid foundation for scaling up the labeling process to larger datasets while refining model accuracy through iterative improvements.

**דוח מפורט על תוצאות Crowdsourcing**

**סקירה כללית**

דוח זה מעריך את ביצועי התיוג של מתייגים אנושיים (דניאל והדס) ומשווה אותם לתוצאות של מודל למידת מכונה (LLM). ההערכה כוללת 23 פוסטים שתויגו לפי שלושה מדדים:

1. **עוצמת לחץ**: מדד לעוצמת הלחץ המתואר בפוסט.
2. **עומס רגשי**: מדד לעומס רגשי או תשישות נפשית המתוארת בפוסט.
3. **שבוע ההיריון**: זיהוי שבוע ההיריון המוזכר בפוסט.

עבור כל מדד, הדוח חוקר את רמת ההסכמה בין המתייגים האנושיים ובין ממוצע התיוגים האנושיים למודל. שיטות סטטיסטיות שונות שימשו להערכת ההסכמה.

**שיטות**

**שאלות שנשאלו:**

1. **עוצמת לחץ**:
   * *שאלה לאנושיים*: "דרגו את עוצמת הלחץ המתוארת בפוסט בסולם מ-1 עד 10, כאשר 1 = אין לחץ, 10 = לחץ קיצוני, ו--1 = אין אזכור ללחץ."
   * *חיזוי המודל*: עוצמת הלחץ החזויה על ידי המודל באותו סולם.
2. **עומס רגשי**:
   * *שאלה לאנושיים*: "העריכו את רמת העומס הרגשי או התשישות הנפשית המתוארת בפוסט בסולם מ-1 עד 10, כאשר 1 = אין עומס, 10 = עומס קיצוני, ו--1 = אין אזכור לעומס."
   * *חיזוי המודל*: עומס רגשי חזוי על ידי המודל באותו סולם.
3. **שבוע ההיריון**:
   * *שאלה לאנושיים*: "ציינו את שבוע ההיריון המוזכר בפוסט כמספר (למשל, 20 לשבוע 20). אם אין אזכור לשבוע, חזרו -1."
   * *חיזוי המודל*: שבוע ההיריון החזוי או -1 אם אין אזכור.

**מדדים סטטיסטיים:**

1. **קורלציה של פירסון**: מודדת קשר ליניארי בין שתי קבוצות של דירוגים.
2. **קורלציה של ספירמן**: מודדת קשר מונוטוני בין שתי קבוצות של דירוגים.
3. **מקדם קורלציה תוך-מחלקתי (ICC)**:
   * *ICC (אנושיים ומודל)*: מדד להסכמה בין המתייגים האנושיים והמודל.
   * *ICC (אנושיים בלבד)*: מדד להסכמה בין שני המתייגים האנושיים.
4. **תרשימי פיזור**: המחשות חזותיות להערכת היחסים בין תיוגי האנושיים והמודל.

**תוצאות**

**1. עוצמת לחץ**

* **פירסון (אנושיים)**: 0.918 — הסכמה ליניארית חזקה בין דניאל להדס.
* **ספירמן (אנושיים)**: 0.925 — הסכמה מונוטונית חזקה בין המתייגים האנושיים.
* **פירסון (אנושיים לעומת מודל)**: 0.812 — הסכמה בינונית בין ממוצע האנושיים למודל.
* **ספירמן (אנושיים לעומת מודל)**: 0.840 — הסכמה מונוטונית בינונית עם המודל.
* **ICC (אנושיים ומודל)**: 0.245 — הסכמה נמוכה בין המתייגים האנושיים והמודל.
* **ICC (אנושיים בלבד)**: 0.048 — הסכמה נמוכה מאוד בין דניאל להדס.

**2. עומס רגשי**

* **פירסון (אנושיים)**: 0.920 — הסכמה ליניארית חזקה בין דניאל להדס.
* **ספירמן (אנושיים)**: 0.930 — הסכמה מונוטונית חזקה בין המתייגים האנושיים.
* **פירסון (אנושיים לעומת מודל)**: 0.750 — הסכמה בינונית בין ממוצע האנושיים למודל.
* **ספירמן (אנושיים לעומת מודל)**: 0.781 — הסכמה מונוטונית בינונית עם המודל.
* **ICC (אנושיים ומודל)**: 0.280 — הסכמה נמוכה בין המתייגים האנושיים והמודל.
* **ICC (אנושיים בלבד)**: 0.014 — הסכמה נמוכה מאוד בין דניאל להדס.

**3. שבוע ההיריון**

* **פירסון (אנושיים)**: 1.000 — הסכמה ליניארית מושלמת בין דניאל להדס.
* **ספירמן (אנושיים)**: 0.999 — הסכמה מונוטונית כמעט מושלמת בין המתייגים האנושיים.
* **פירסון (אנושיים לעומת מודל)**: 0.768 — הסכמה בינונית בין ממוצע האנושיים למודל.
* **ספירמן (אנושיים לעומת מודל)**: 0.743 — הסכמה מונוטונית בינונית עם המודל.
* **ICC (אנושיים ומודל)**: -0.475 — הסכמה שלילית, מציינת סטייה משמעותית בין המתייגים האנושיים והמודל.
* **ICC (אנושיים בלבד)**: -0.276 — הסכמה שלילית בין דניאל להדס.

**דיון**

**מדוע נבחרו מדדים אלו:**

1. **קורלציות פירסון וספירמן**:
   * מדדים אלו בוחנים היבטים שונים של קשרים: ליניארי (פירסון) ומונוטוני (ספירמן). הם מספקים ניתוח מקיף גם אם הנתונים חורגים מליניאריות.
2. **ICC**:
   * מקדם קורלציה תוך-מחלקתי מודד את העקביות והאמינות של דירוגים בין מספר מתייגים.
   * השימוש ב-ICC מסייע להעריך עד כמה המודל מתיישר עם תיוגי האנושיים, מה שחשוב להבנת ביצועי המודל בהקשר מציאותי.

**תובנות:**

1. **עוצמת לחץ ועומס רגשי**:
   * המתייגים האנושיים הציגו הסכמה חזקה (קורלציות פירסון וספירמן גבוהות), מה שמעיד על עקביות בתיוג.
   * הסכמה בינונית עם המודל מצביעה על כך שהמודל תופס חלק מהתבניות אך מתקשה לשחזר באופן מלא את השיפוט האנושי.
   * ערכי ICC נמוכים מרמזים על שונות בדירוגים, במיוחד כשכוללים את המודל.
2. **שבוע ההיריון**:
   * הסכמה מושלמת בין המתייגים האנושיים משקפת את האופי האובייקטיבי של משימה זו.
   * הסכמה בינונית עם המודל מצביעה על כך שבעוד המודל מבצע ביצועים סבירים, יש מקום לשיפור בזיהוי שבועות ההיריון.
   * ערכי ICC שליליים מציינים חוסר התאמה בין המודל לאנושיים, במיוחד במקרי קצה שבהם אין אזכור לשבוע.

**ויזואליזציות**

* תרשימי פיזור נוצרו כדי להמחיש את היחסים:
  1. אנושיים מול אנושיים.
  2. ממוצע האנושיים מול מודל.
* תרשימים אלו מספקים תיאור ברור של אזורי הסכמה וחוסר הסכמה.

**מסקנות**

1. **מתייגים אנושיים**:
   * דניאל והדס הציגו הסכמה חזקה במדדים הסובייקטיביים (עוצמת לחץ, עומס רגשי).
   * הושגה הסכמה מושלמת במדד האובייקטיבי (שבוע ההיריון).
2. **ביצועי המודל**:
   * הסכמה בינונית עם ממוצע האנושיים מדגישה את התחומים שבהם המודל מתיישר עם השיפוט האנושי, אך גם חושפת את מגבלותיו.