יאיר ליפשיץ – Yair Lifshitz

יניר מרמור – Yanir Marmor

ivrit.ai: איך מלמדים את המחשב לשמוע עברית

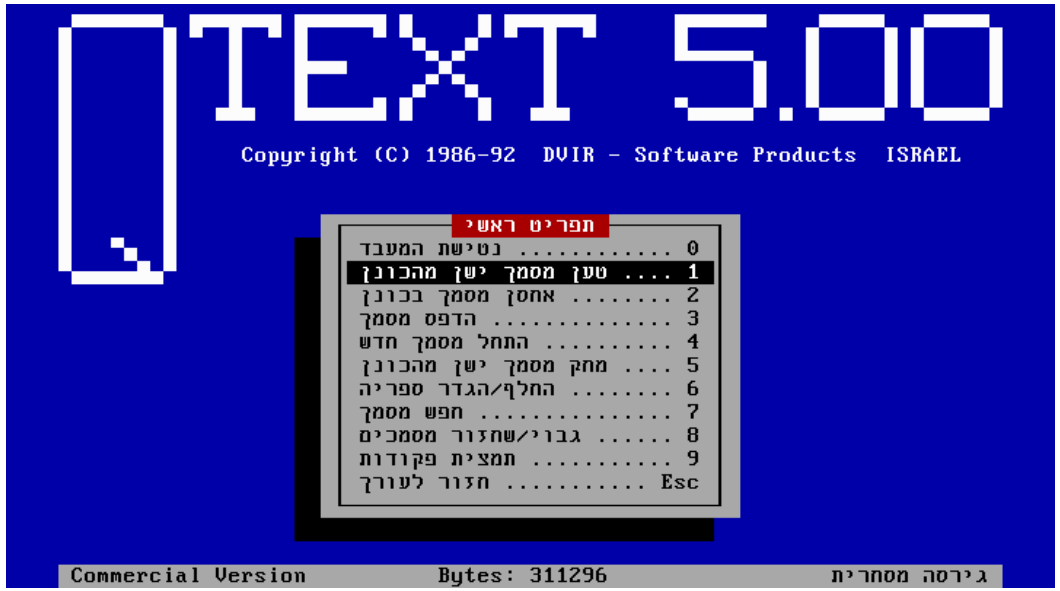
עוזרים לטכנולוגיה להבין עברית החל ממאי 2023

יאיר ליפשיץ, יניר מרמור, כנרת משגב
מאי 2024

עברים, דברו עברית

על מה נדבר?

- ivrit.ai: מה ולמה
- איך בונים מודל AI
- התוכנית
- תמלול המונים



ivrit.ai, מה זה?

- עברית שפה קשה, גם בטכנולוגיה
- הברות שונות (ז', ח, ע, צ)
- זכר/נקבה

- שתי טכנולוגיות מרכזיות שהתפתחו מאוד ב-18 החודשים האחרונים
  - מודלי שפה (ChatGPT ודומיו)
  - מודלי תמלול (Whisper ודומיו)

- **המטרה שלנו**: עברית באיכות גבוהה במוצרי טכנולוגיה
  - **תמלול**

ע
עברים, דברו עברית

למה?

- מודלי שפה: לא "עוד טכנולוגיה"
  - הסקת מסקנות
  - שליפת ידע
  - **בן אדם** (כמעט)

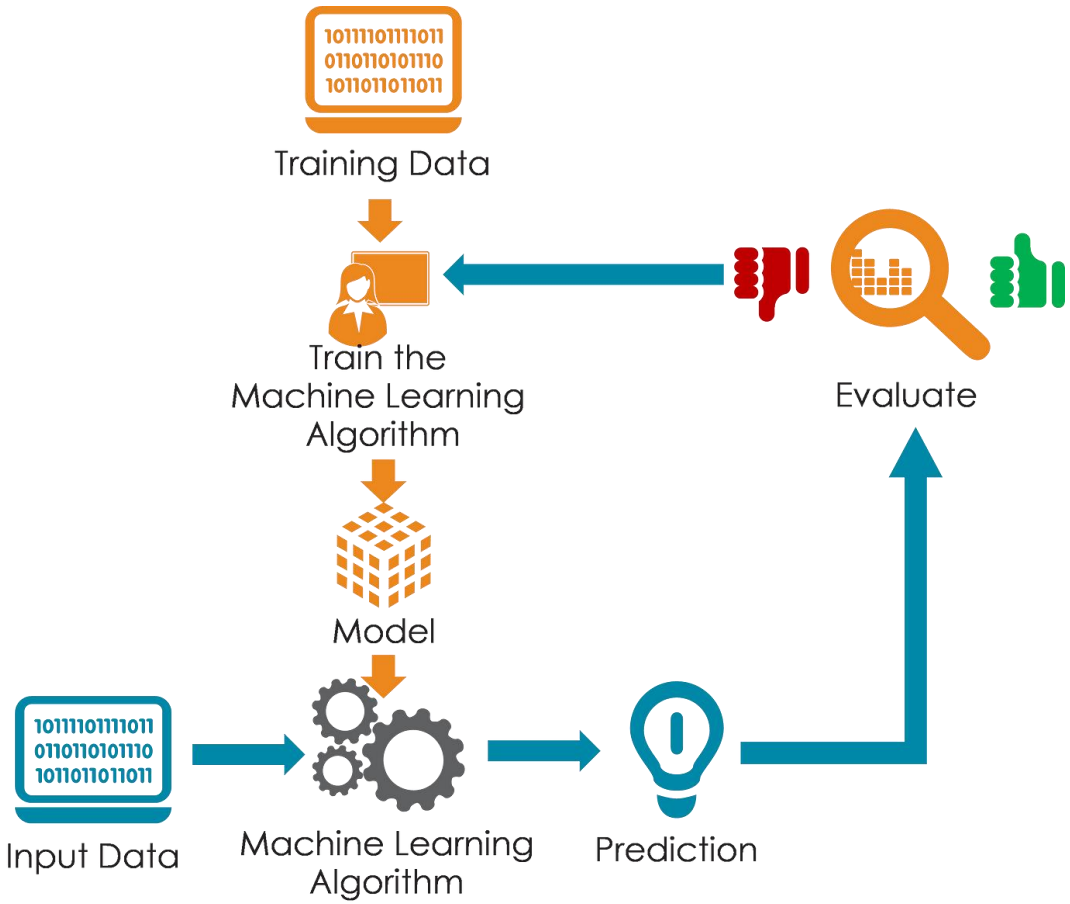
- תמלול (Speech to text)
  - דיבור יעיל בסדר גודל מכתיבה
  - ילדים, זקנים, לא דוברי אנגלית שוטפת

- יעילות ברמת הממשק: מענה אוטומטי, שירותים מתקדמים, הפעלת ציוד ותוכנה
- יעילות -> פריון (!)

ע
עברים, דברו עברית

איך בונים מודל AI?

- **ארכיטקטורה**
- **מידע**
- **אימון**
- **כוונון**



עברים, דברו עברית

התוכנית

- **המטרה**: אפשור מספר רב של מודלים איכותיים בעברית

- **מה כבר קיים?**
  - מודלים איכותיים לתמלול: Whisper, 200 אלף שעות באנגלית ועוד 680 שעות בעברית 🤦
  - 20-30 שעות במאגרים אחרים
  - עשרות אלפי שעות מתומללות ברשות המדינה – לא נגישות לשימוש ציבורי 😠

- **הדרך**: הנגשת מאגרי מידע מתוייגים
  - מטרה: 10,000 שעות תוכן מתומללות
  - חובה שיהיו ברשיון
  - הכל פתוח וחינמי

- **בעיה: איך מתמללים?**

ע
עברים, דברו עברית

תמלול המונים

- **תמלול מקצועי**: 600 ₪ לשעה מתומללת
  - סה"כ כ-6 מיליון ₪ לכל המאגר

- **רעיון**: תמלול המונים
  - המוני מתנדבים, כל אחד מתמלל מעט
  - קטעים קצרים (עד 30 שניות) בכדי לא להתעייף
  - ...
  - הצלחה!

- **שאלת תם**: יש המון חומר באינטרנט, מה איתו?

ע
עברים, דברו עברית

שינוי בהירות

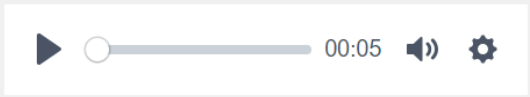
מחובר כ-yair@lifshitz.io. התנתק
דקות שתומללו על ידך: 68:54
דקות שתומללו בסה"כ: 3688:53
דירוג: 13 (מתוך 510)
מרחק לרמה הבאה: 4:33

מקור: Geekonomy

פרק: 2022.08.09 פרק #592 - עדנה חלבני, 50 שנה במשרד ראש הממשלה

מקטע: 140

תמלול ראשוני: לא חשוב, בן דביל. בקיצור, כשהיא חוטפת היום מלא מלא מלא ביקורת.



לא חשוב, בן דביל. בקיצור, כשהיא חוטפת היום מלא מלא מלא ביקורת.

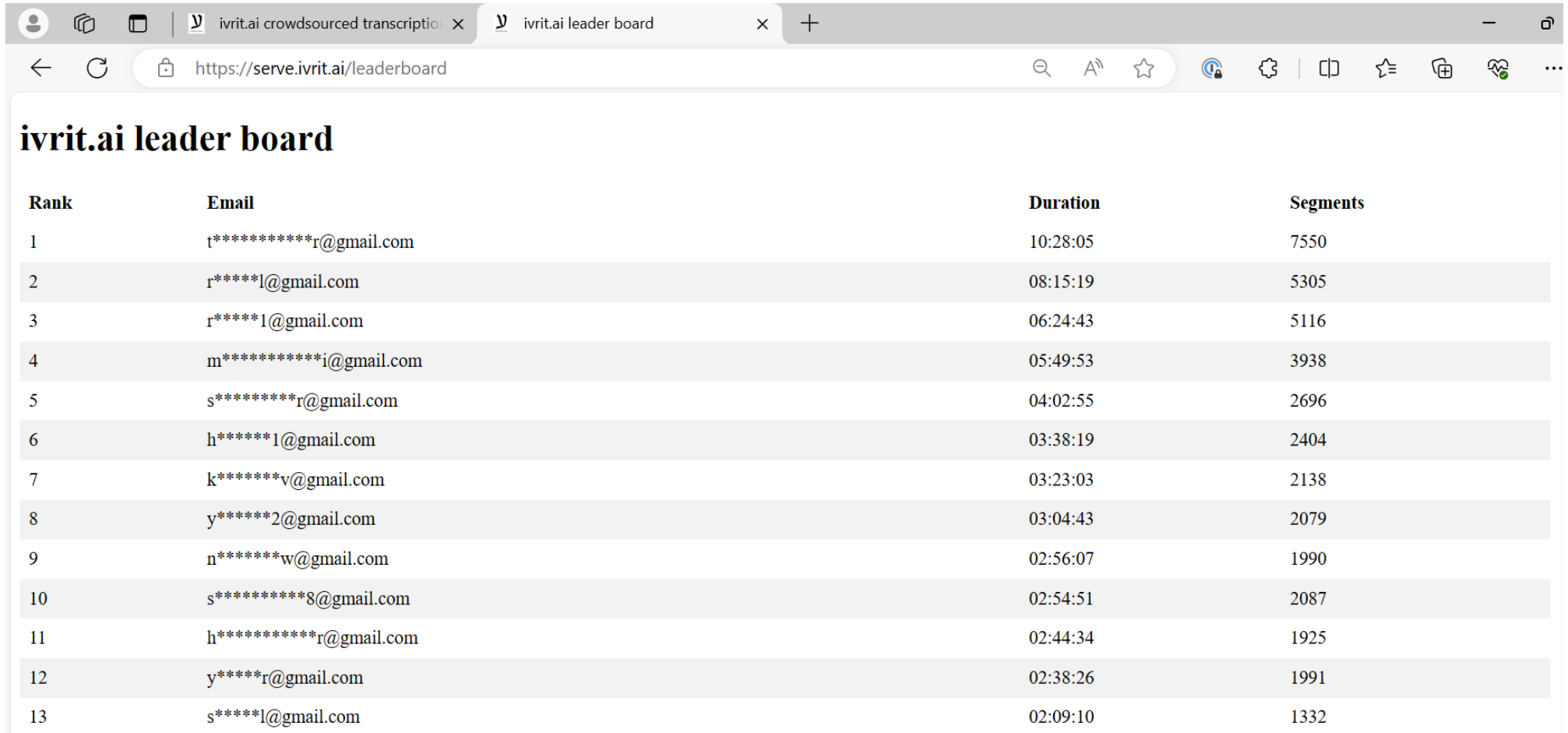
כללי תמלול

הגשה | לא הצלחתי

עברים, דברו עברית

תמלול המונים: אתגרים

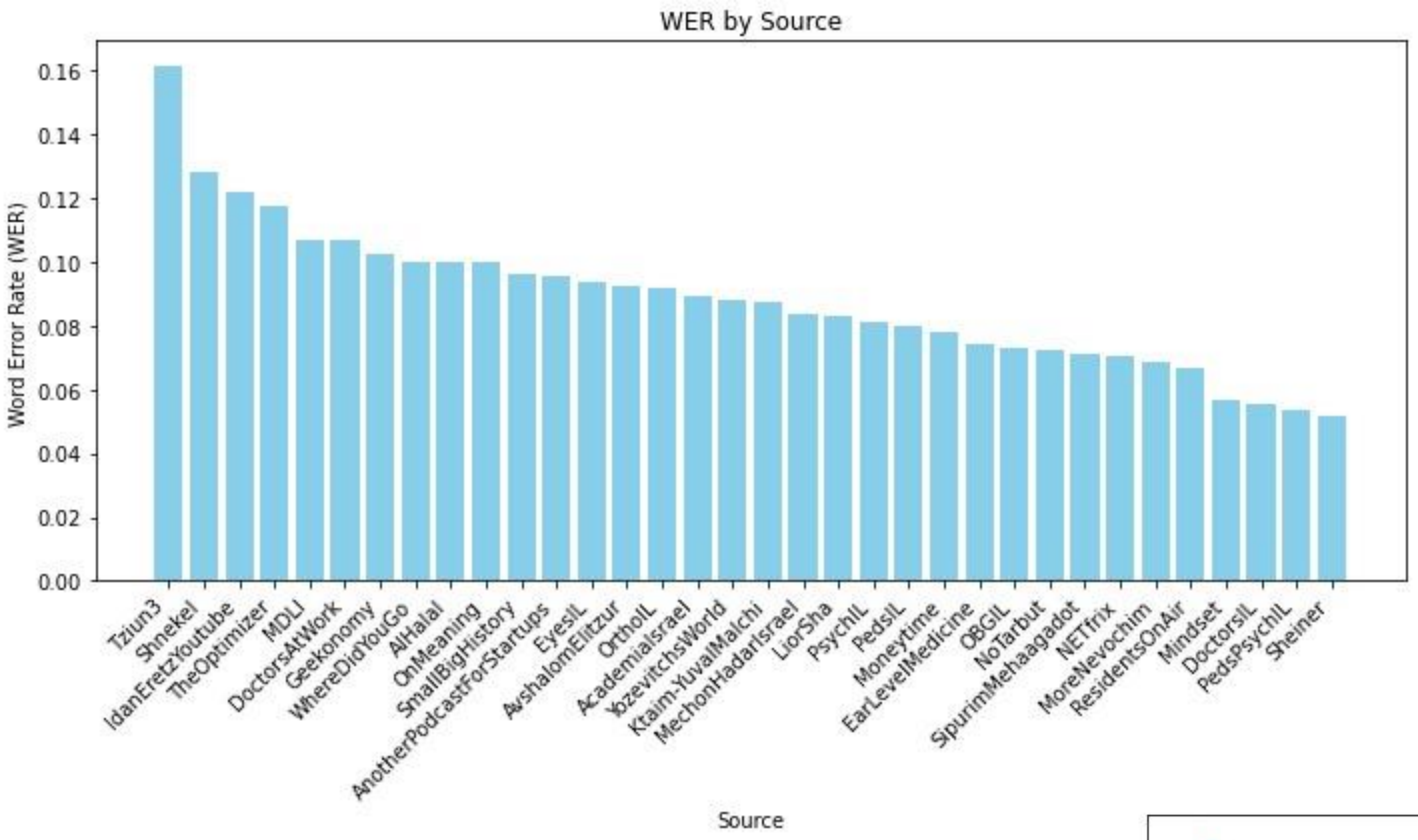
- איתור מתנדבים
  - קל יחסית להביא, קשה להשאיר
  - משחקיות
- איכות התמלול
  - קטעי קול לא מובנים; ריבוי דוברים
  - הקשר (שמות אנשים, מקומות, ...)
  - **איכות המתמלל/ת**
    - הצלבות

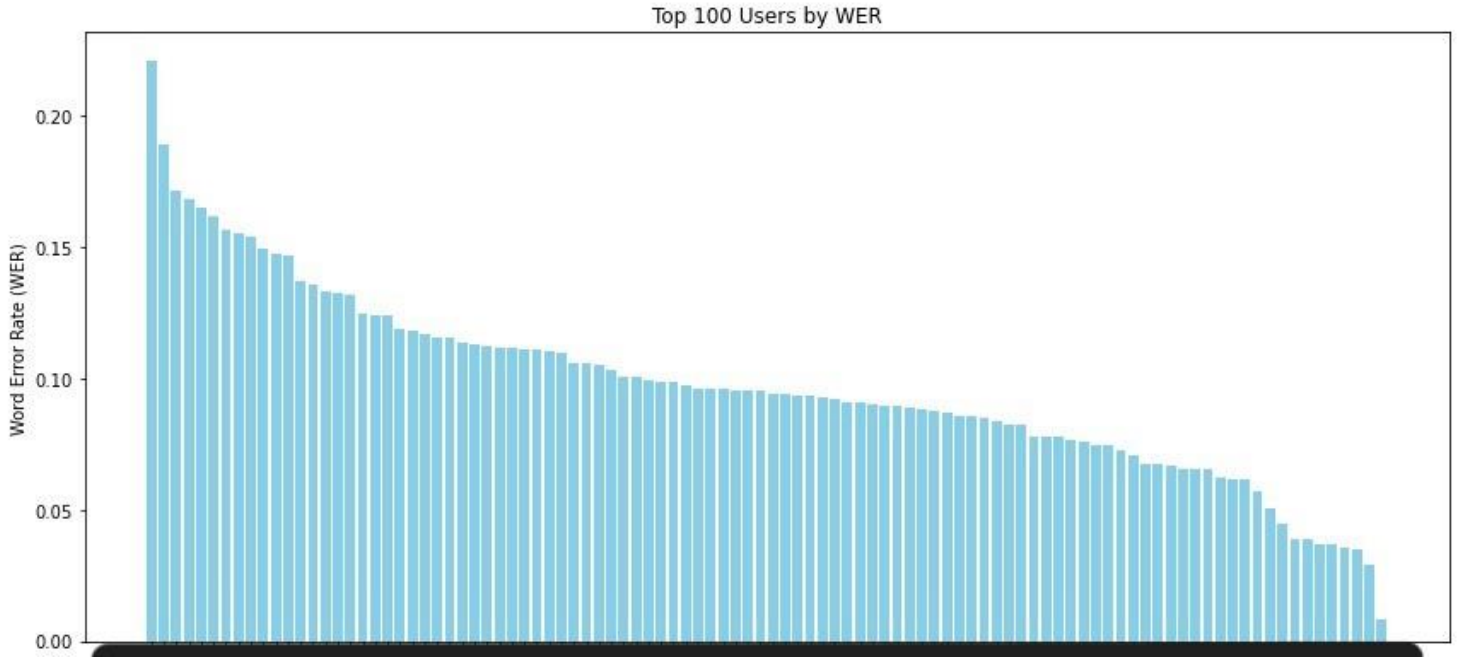


ivrit.ai leader board

Rank	Email	Duration	Segments
1	t**********r@gmail.com	10:28:05	7550
2	r*****l@gmail.com	08:15:19	5305
3	r*****1@gmail.com	06:24:43	5116
4	m**********i@gmail.com	05:49:53	3938
5	s*********r@gmail.com	04:02:55	2696
6	h******1@gmail.com	03:38:19	2404
7	k*******v@gmail.com	03:23:03	2138
8	y******2@gmail.com	03:04:43	2079
9	n*******w@gmail.com	02:56:07	1990
10	s**********8@gmail.com	02:54:51	2087
11	h**********r@gmail.com	02:44:34	1925
12	y*****r@gmail.com	02:38:26	1991
13	s*****l@gmail.com	02:09:10	1332

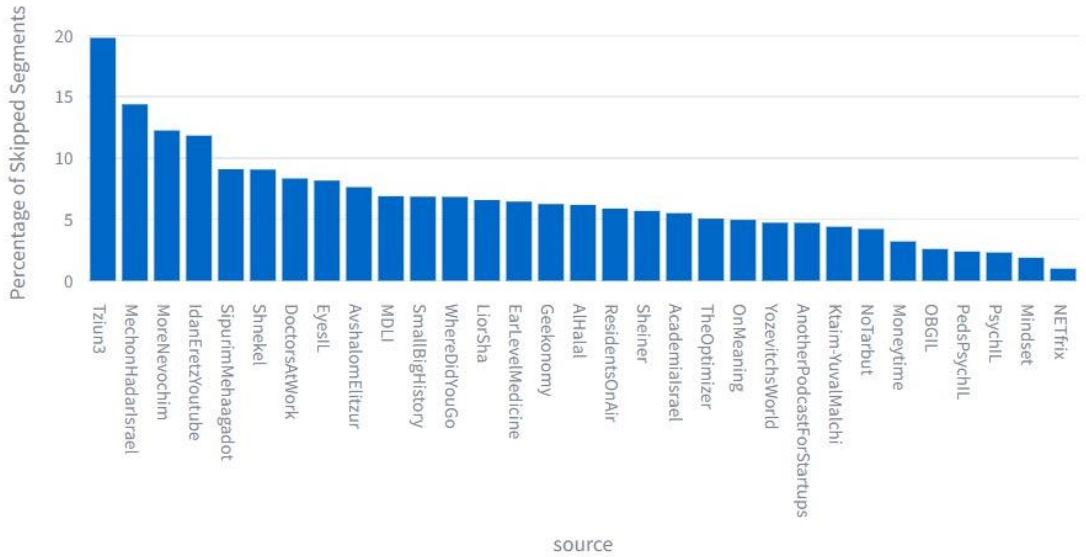
עברים, דברו עברית





עברים, דברו עברית

Skipped Segments per Source (excluding "foreign_language") - Percentage



עברים, דברו עברית

איפה אנחנו היום?

- 198 שעות תמלול בכ-5 חודשים
  - מעל 1500 מתנדבים, מתוכם כ-200 "קבועים"
  - מעל 100 תמללו יותר מ-10 דקות כל אחד
  - המובילה: 10+ שעות (!)
  - כל החומר מונגש לציבור בצורה חופשית
- 2 מודלי תמלול חופשיים שוחררו, עם שיפור של מעל 20% בדיוק
  - זמינים גם דרך אפליקצית Vibe
- עובדים להשיג גישה למאגרי מידע מתומללים נוספים
- המטרה: מודל התמלול המדויק בעולם (טוב בעברית כמודלים אחרים באנגלית)

ע
עברים, דברו עברית

איך אפשר לעזור?

- **קל**: התנדבות דרך https://serve.ivrit.ai

- **בינוני**: תרומות לצורך משאבי מחשוב נוספים
  - שירות תמלול חינמי בעברית
  - תשלום על תחזוקת האתר, שמירת החומר וכו

- **קשה**: גישה לחומר מוקלט ומתומלל
  - מדינת ישראל: **מערכת בתי המשפט**, **הכנסת**, **כאן11**
  - גופי חדשות
  - חברות הפקה
  - כל גוף כזה יכול לספק אלפי שעות תוכן מתומללות, ולהקפיץ אותנו שנות אור קדימה

תודה!