

Hebrish v1.0: כוונן עדין של Parakeet להכתבה באנגלית משולבת עברית

איך אומן (Jachnun) v1.0, איך יוצא ואיך נמדד, במה הוא מצליח, ומה המספרים לא מראים.

דניאל רוזהיל • 2026-10-19 • hebrish.app

מודל: [danielrosehill/Parakeet-TDT-0.6B-v2-English-Hebrew-Codeswitching](https://huggingface.co/danielrosehill/Parakeet-TDT-0.6B-v2-English-Hebrew-Codeswitching), גרסה v1.0, ריצה r6-en30

תקציר

דוברי אנגלית בישראל משלבים מילים בעברית באנגלית שלהם. מערכות זיהוי דיבור באנגלית כמעט לא שמעו את המילים האלה, ולכן כותבות במקומן את המילה האנגלית שנשמעת הכי קרוב. Hebrish v1.0 הוא כוונן עדין מלא (full fine-tune) של NVIDIA Parakeet TDT 0.6B v2, על 339.2 דקות של דיבור מוקרא בהחלפת קוד (code-switching) ועוד כ-30% אנגלית. המודל לומד לכתוב כל מילה עברית בתעתיק לטיני קנוני אחד (*Bituach Leumi, makolet*). זוהי נקודת הביקורת (checkpoint) השישית והגרסה היציבה הראשונה; חמש נקודות הביקורת הקודמות נקראות כעת PoC-1 עד PoC-5. במבחן הייחוס הציבורי להחלפת קוד אנגלית-עברית (סט מבחן v1, תת-הסט הנקי, 84 קטעים) משיג v1.0 שיעור זיהוי מונחים עבריים (term recall) של 73.3%, 70.1% על מונחים שלא הופיעו באימון, WER של 3.3% ו-WER של 2.2% בבקרת האנגלית, לעומת 32.6%, 32.5%, 9.5% ו-2.7% ב-Parakeet v2 המקורי. האנגלית הרגילה לא נפגעה. כל ההקלטות, באימון ובמבחן, הן בקולו של דובר אחד, ולכן המספרים האלה מדרגים מודלים עבור הדובר הזה בלבד.

1.7%

WER באנגלית של הדובר
עצמו (v2 המקורי: 3.1%)

3.3%

WER, מבחן ייחוס ציבורי (v2)
המקורי: 9.5%

70.1%

זיהוי מונחים שמעולם לא
הופיעו באימון

73.3%

זיהוי מונחים, מבחן ייחוס
ציבורי (v2 המקורי: 32.6%)

1. הבעיה

דוברי אנגלית רבים בישראל מדברים אנגלית שמשובצות בה מילים בעברית: "*grab some bamba from the makolet*", "*I have to go to Bituach Leumi*". זו החלפת קוד (code-switching). מודלים להכתבה באנגלית אומנו על אנגלית, וכמעט לא שמעו את המילים האלה, אם בכלל. לכן הם מחליפים אותן במילה האנגלית הקרובה ביותר בצליל: *Macaulay* במקום *makolet*. לא פעם דווקא המילה שנושאת את המשמעות היא זו שיוצאת שגויה.

Hebrish הוא כוונון עדין שמשאיר את ההכתבה באנגלית כפי שהייתה, וכותב נכון את המילים בעברית. צריך להגדיר מה זה "נכון", כי למילים עבריות בטקסט אנגלי אין איות מוסכם אחד. Hebrish משתמש באיות לטיני קנוני אחד לכל מילה, מתוך לקסיקון איות ציבורי (`terms.csv`), וכותב עברית באותיות לטיניות בלבד. הטוקניזר של Parakeet הוא SentencePiece BPE בן 1,024 טוקנים, עם אותיות רישיות וסימני פיסוק אך בלי תווים עבריים, כך שכתב עברי מעולם לא היה אפשרות. התעתיק הלטיני גם מתאים לשימוש עצמו, שהוא טקסט באנגלית.

מודל הבסיס הוא NVIDIA Parakeet TDT 0.6B v2 (רישיון CC-BY-4.0), מודל FastConformer TDT בן 0.6B פרמטרים. הוא קטן מספיק כדי לרוץ על מעבד של מחשב נייד ועל טלפון, ושם בדיוק מכתיבים.

2. מערך הנתונים

נתוני האימון הם `Read-Aloud-Voice-Sentences` (ציבורי, CC-BY-4.0), בגרסה `550bfcaa0f84` (revision). כל קטע הוא **דובר אחד** (דניאל רוזהיל, יליד אירלנד, דובר אנגלית, גר בירושלים) שמקריא בקול משפט מוכן מראש, רובם על אותו טלפון, כשמע מונו 16 kHz עם תמלול כולל אותיות רישיות ופיסוק ורשימת מונחי היעד של הקטע. משפטי החלפת הקוד מכילים את המונחים העבריים; משפטים באנגלית רגילה מוקלטים באותו קול, לבדיקת השכחה ולתמהיל האימון.

חלוקה בגרסה <code>550bfcaa</code>	קטעים	דקות	שימוש
אימון החלפת קוד (<code>cs_train</code>)	3,224	339.2	אימון
אימות החלפת קוד (<code>cs_val</code>)	218	22.6	בחירת האפוק הטוב ביותר
מאגר האנגלית של הדובר	751	95.3	אימון (כולו)
מאגר ריענון LibriSpeech-PC	900	190.7	אימון (נדגמו 48.5 דקות)
מבחן החלפת קוד (<code>test_cs</code>)	387	43.9	הערכה
מבחן האנגלית של הדובר	73	9.2	בדיקת שכחה
LibriSpeech test-clean	200	24.6	בדיקת שכחה (דוברים אחרים)
מבחן הקלטות ארוכות	קבצים 15	44.6	הקלטות ארוכות

תמהיל האימון

הריצה אומנה על 4,202 שורות, 483.0 דקות: **339.2** דקות של החלפת קוד, **95.3** דקות של האנגלית של הדובר עצמו ו-**48.5** דקות של ריענון (replay) מ-LibriSpeech-PC. היעד לחלק האנגלית היה 30%, והתוצאה בפועל 29.8%. היחס הזה נשען על הספרות בתחום החלפת הקוד והלמידה המתמשכת: כוונון עדין על נתונים מהתחום החדש בלבד הוא מה שגורם לשכחה קטסטרופלית, והפתרון המקובל הוא לערבב נתונים מהתחום המקורי ומעט ריענון של נתונים כלליים.

מונחים מוחזקים בצד (held-out)

החלוקה **זרה במונחים ביחס לרשימת מונחים מוחזקים**, לא ביחס לכל מונח. בגרסה הזו יש 362 מונחים ברשימה; משפט שמכיל אחד מהם עובר לסט המבחן, כך שהמונחים האלה לעולם לא מופיעים באימון. משפט מבחן יכול להכיל גם מונחים אחרים שאינם ברשימה, ואלה יכולים להופיע באימון. לכן ההערכה מדווחת שני מספרים: זיהוי של מונחים מוחזקים (293 מופעים של 118 מונחים ב-`test_cs`), שמודד הכללה, וזיהוי של מונחי מבחן שהופיעו גם באימון (278 מופעים), שמודד למידה. רק המספר הראשון מעיד שהמודל כותב מילה שמעולם לא לימדו אותו.

3. נקודת ביקורת 6 לאימון (v1.0)

v1.0 הוא הריצה r6-en30, שאומנה ב-2026-10-19. זהו כונון עדין מלא מהמודל המקורי nvidia/parakeet-tdt-0.6b-v2 על מערך הנתונים כולו כפי שהיה באותו רגע, ולא המשך אימון של PoC-5.

הגדרה	ערך
סביבת עבודה	NeMo 2.7.3 (nemo_toolkit[asr,cu12]), torch 2.11.0, נעולים בתמונת Modal
נקודת מוצא	nvidia/parakeet-tdt-0.6b-v2, הטוקניזר נשמר
אופטימיזר	AdamW, lr 3e-5, betas 0.9 / 0.98, weight decay 1e-3
תזמון	Cosine annealing, 10% חימום (warmup), min lr = lr / 20
אפוקים / אצווה	20 אפוקים, גודל אצווה 16
דיוק / seed	דיוק מעורב bf16, seed 7
בחירה	האפוק הטוב ביותר על cs_val : אפוק 18 מתוך 20, WER אימות 5.0% (גולמי, 0.0497)
חומרה / זמן	NVIDIA L40S אחד (48 GB) ב-Modal, 41.0 דקות

למה כונון מלא מצטבר ולא המשך אימון

כל גרסה מתחילה מ-Parakeet המקורי ומתאמנת על כל ההקלטות שנאספו עד אז. המשך אימון של הכונון הקודם על נתונים חדשים בלבד הוא מה שגורם לשכחה קטסטרופלית, והמשך אימון על כל הנתונים נותן לשורות הישנות עוד סבב של אפוקים בכל גרסה, כך שהסחף מצטבר. ריצה עולה בערך US\$0.80, כך שהמשך אימון לא חוסך כלום. האימון והבנצ'מרק הם שלבים נפרדים: האפוק הטוב ביותר נבחר על cs_val (חלק מהאימון, אף פעם לא מהמבחן), והבנצ'מרק רץ אחר כך על סט המבחן.

למה 30% אנגלית

מודל הכתבה שלומד מילים בעברית אבל מאבד את האנגלית לא שווה כלום. חלק האנגלית שומר על המשימה המקורית בכל אצווה: האנגלית של הדובר עצמו מלמדת שאנגלית רגילה בקול הזה צריכה להישאר אנגלית רגילה, וריענון LibriSpeech שומר על קולות אחרים. פרק 6 בודק אם זה עבד.

שושלת

נקודת ביקורת	ריצה	דקות CS	שוחר	נמדד בבנצ'מרק
PoC-1 · Amba	pilot-en30	77.4	2026-10-15	כן
PoC-2 · Bamba	r2-en30	120.1	2026-10-15	כן
PoC-3 · Chamin	r3-en30	196.1	2026-10-18	כן
PoC-4 · Ful	r4-en30	269.3	2026-10-18	לא
PoC-5 · Halva	r5-en30	294.2	2026-10-18	לא
v1.0 · Jachnun	r6-en30	339.2	2026-10-19	כן, 2026-10-19

PoC-1 עד PoC-5 פורסמו ב-Hugging Face בשמות v1 עד v5. נקודת הביקורת השישית נקראת v1.0 כי זו הגרסה הראשונה שהפרייקט עומד מאחוריה. גרסת סטרימינג מבוססת Nemotron (ריצה r6-nemotron-en30, בסיס nvidia/nemotron-speech-streaming-en-0.6b, אפוק מיטבי 18, WER אימות 6.3% גולמי, 45.8 דקות) אומנה על אותם נתונים ובאותו מתכון.

4. ייצוא

v1.0 מפורסם בפורמטים שבהם משתמשות סביבות ההרצה על המכשיר. כל ייצוא נבדק על חמישה קטעים מול ONNX ב-fp32 לפני כל בנצ'מרק.

פורמט	היכן	רץ ב-	בדיקה (5 קטעים מול fp32)
nemo	מאגר המודל הראשי	NeMo (PyTorch)	ייחוס
GGUF F16, Q8_0, Q4_K_M	Hebrish/...-GGUF	,Handy ,transcribe.cpp Offline Voice Input	F16 5/5, Q8_0 5/5, Q4_K_M 4/5
ONNX ב-fp32 וב-int8	Hebrish/...-ONNX	,Voctype ,onnx-asr ההדגמה בדפדפן	int8 4/5
sherpa-onnx int8	Hebrish/...-sherpa-onnx-int8	sherpa-onnx, אפליקציית Hebrish לאנדרואיד	4/5
MLX	Hebrish/...-mlx	parakeet-mlx (Apple Silicon)	עדיין PoC-3; לא יוצא מחדש

הטעויות ב-int8 הן מילה אחת: tzatzumim במקום שבו fp32 כתב satzumim; התמלול המקורי הוא tza'atzuim. תוצאות הבנצ'מרק שלהלן הן של קובץ nemo ב-PyTorch, לא של הגרסאות המכונצות (quantized).

5. הערכה

v1.0 נמדד ב-2026-10-19 על שני סטים של מבחן, שעונים על שאלות שונות.

מבחן הייחוס הציבורי

מבחן הייחוס הציבורי לזיהוי דיבור בהחלפת קוד אנגלית-עברית (מערך נתונים - Hebrish/English-Hebrew-Codeswitching ASR-Benchmark, סט מבחן v1, מוקפא) כולל 145 קטעים. תת-הסט הנקי, שעליו מדווחות התוצאות המרכזיות, הוא 84 הקלטות טלפון (11.8 דקות) של משפטים שאינם בשום סט אימון שפורסם, עם 172 מופעים של מונחים עבריים (139 מונחים שונים). תת-סט נפרד של בקרת אנגלית, בן 24 קטעים, הוא אותו דובר מקריא אנגלית בלבד.

- זיהוי מונחים (term recall): שיעור מופעי המונחים שהאיות הקנוני שלהם מופיע בתמלול כמילה שלמה, בלי תלות באותיות רישיות, כשמקף ורווח שקולים. זיהוי מונחים לא מוכרים מצמצם את המדד למונחים שאינם בחלוקת האימון הציבורית (117 מתוך 172 המופעים).
- WER: אותיות קטנות, מקפים הופכים לרווחים, פיסוק מוסר מלבד גרש בתוך מילה (jiwer 3.1.0).
- בדיקת זליגה: אף אחד מ-145 קטעי הבנצ'מרק, לא לפי מזהה ולא לפי טקסט, אינו נמצא ברשימת האימון של v1.0 (-runs/r6) en30/train.jsonl (4,202 שורות), וגם לא בזו של ריצת Nemotron.

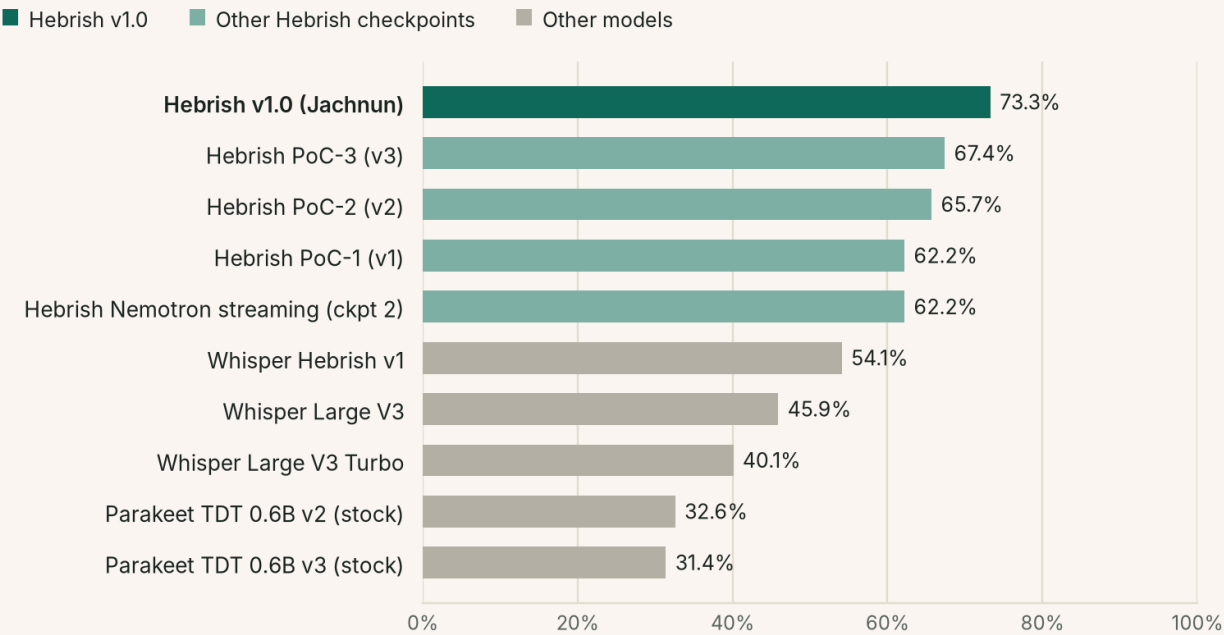
כל קובצי התוצאות נוקדו עם אותו מנרמל ואותו מתאם מונחים; הכוונון של Nemotron נמדד במצב לא מקוון (הקשר מלא), לא במצב סטרימינג.

סט המבחן המוחזק

סט המבחן של הפרויקט עצמו בגרסה 550bfcaa0f84 (פרק 2) גדול יותר: 387 קטעים בהחלפת קוד, זיהוי המונחים המוחזקים והמונחים המוכרים שתוארו למעלה, בדיקות השכחה על האנגלית של הדובר ועל LibriSpeech, ו-15 הקלטות ארוכות. v2 המקורי, PoC-3 (הגרסה האחרונה שנמדדה), v1.0 וריצת Nemotron נוקדו כולם מחדש על הגרסה הזו, כך שההשוואה ביניהם הוגנת. ה-WER כאן מנורמל. המספרים האלה אינם ברי השוואה לתוצאות ה-PoC הישנות, שהשתמשו בגרסה b47310621f79.

Hebrew term recall, public benchmark

Public benchmark test set v1, clean subset: 84 clips, 172 term occurrences. Scored 2026-10-19.



Source: English-Hebrew-Codeswitching-ASR-Benchmark, round 3 (results/*.json). Nemotron scored offline. hebrish.app/benchmarks
One speaker's voice: every test clip is read by the same speaker, so these numbers rank models on that speaker only.

איור 1. שיעור זיהוי המונחים העבריים במבחן הייחוס הציבורי, תת-הסט הנקי, כל המודלים שנמדדו.

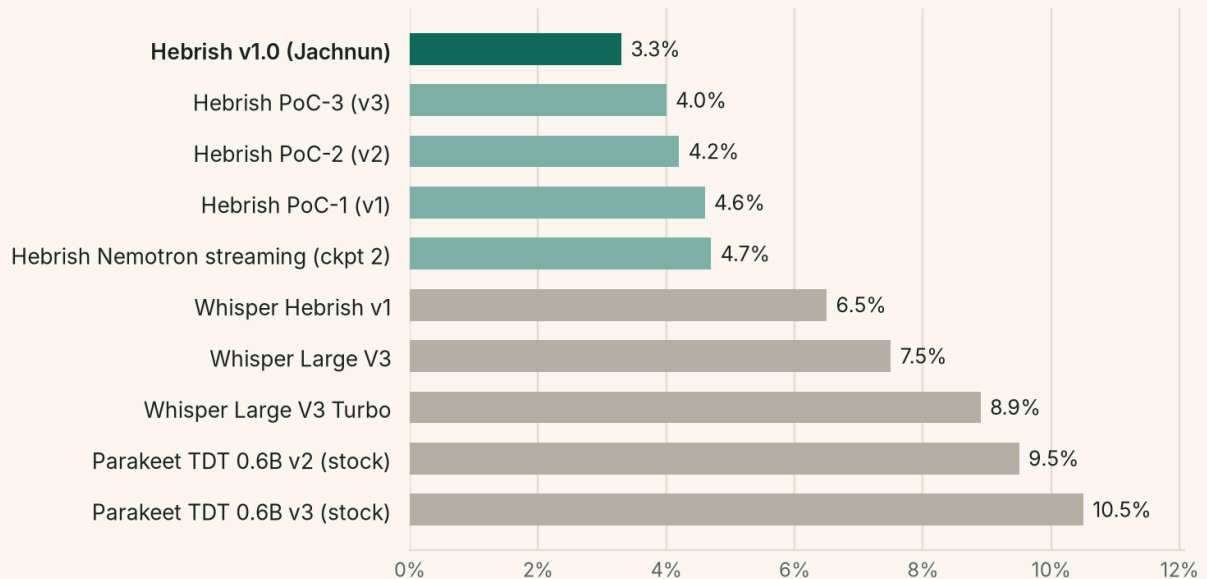
מודל (מבחן ייחוס ציבורי, תת-הסט הנקי)	זיהוי מונחים	מונחים לא מוכרים	WER	WER בבקרת אנגלית
Hebrish v1.0 (Jachnun)	73.3%	70.1%	3.3%	2.2%
Hebrish PoC-3	67.4%	63.2%	4.0%	2.5%
Hebrish PoC-2	65.7%	60.7%	4.2%	2.4%
Hebrish PoC-1	62.2%	61.5%	4.6%	2.5%
Hebrish Nemotron סטרימינג, נקודת ביקורת 2 (לא מקוון)	62.2%	54.7%	4.7%	2.5%
Whisper Hebrish v1	54.1%	56.4%	6.5%	4.7%
Whisper Large V3	45.9%	46.2%	7.5%	2.4%
Whisper Large V3 Turbo	40.1%	39.3%	8.9%	2.4%
Parakeet TDT 0.6B v2 (מקורי)	32.6%	32.5%	9.5%	2.7%
Parakeet TDT 0.6B v3 (מקורי)	31.4%	31.6%	10.5%	2.7%

סבב 3 של טבלת הדירוג, נוקד ב-2026-10-19. על 172 מופעים, הפרש של 2 עד 3 נקודות בזיהוי הוא בערך 4 עד 5 פגיעות, בתוך טווח הרעש.

Word error rate, public benchmark

Public benchmark test set v1, clean subset: 84 clips, 172 term occurrences. Scored 2026-10-19.

■ Hebrish v1.0 ■ Other Hebrish checkpoints ■ Other models



Source: English-Hebrew-Codeswitching-ASR-Benchmark, round 3 (results/*.json). Nemotron scored offline. hebrish.app/benchmarks

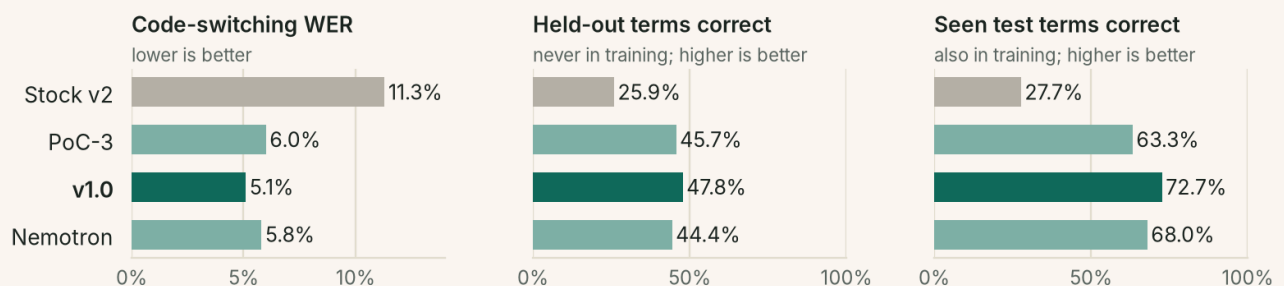
One speaker's voice: every test clip is read by the same speaker, so these numbers rank models on that speaker only.

איור 2. שיעור שגיאות מילים (WER) במבחן הייחוס הציבורי, תת-הסט הנקי.

Same split, four models

Held-out split at dataset revision 550bfcaa0f84: 387 code-switched clips. Scored 2026-10-19.

■ Hebrish v1.0 ■ Other Hebrish checkpoints ■ Stock



Source: Parakeet-V2-Personal-Finetune data/evals/r6-*/summary.json. Normalised WER. hebrish.app/benchmarks

One speaker's voice: every test clip is read by the same speaker, so these numbers rank models on that speaker only.

איור 3. v2 המקורי, PoC-3, v1.0 ו-Nemotron על אותו סט מבחן מוחזק (גרסה 550bfcaa0f84).

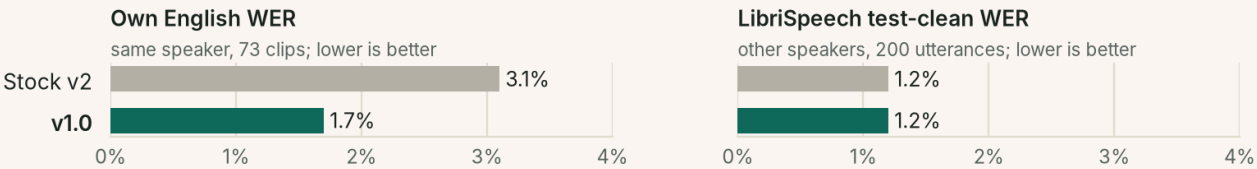
סט מוחזק, 550bfcaa	CS-ב WER	מונחים מוחזקים (293)	מונחים מוכרים (278)	אנגלית הדובר	LibriSpeech	הקלטות ארוכות
Parakeet v2 המקורי	11.3%	25.9%	27.7%	3.1%	1.2%	9.6%
Hebrish PoC-3	6.0%	45.7%	63.3%	2.3%	1.0%	5.8%
Hebrish v1.0	5.1%	47.8%	72.7%	1.7%	1.2%	5.0%
Hebrish Nemotron (לא מקוון)	5.8%	44.4%	68.0%	3.3%	2.0%	5.8%

מה התוצאות אומרות

- מילים בעברית.** v1.0 מכפיל יותר מפי שניים את שיעור זיהוי המונחים של Parakeet המקורי במבחן הייחוס הציבורי (73.3% לעומת 32.6%) ומוביל על פני כל המודלים שנמדדו, כולל Whisper Large V3 (45.9%) וכוונן ה-Whisper הקודם (54.1%). זיהוי המונחים הלא מוכרים (70.1%) קרוב לזיהוי הכולל, מה שמרמז שהשיפור אינו רק שינון של מונחי האימון.
- WER בהחלפת קוד** ירד מ-11.3% (מקורי) ל-5.1% בסט המוחזק, ומ-9.5% ל-3.3% במבחן הייחוס הציבורי.
- לעומת PoC-3,** v1.0 טוב יותר בכל מדד של מבחן הייחוס הציבורי. בסט המוחזק, השיפור במונחי מבחן שהופיעו באימון (+9.4 נקודות) אמיתי; השיפור במונחים המוחזקים (+2.1 נקודות, בערך 6 פגיעות מתוך 293) בתוך טווח הרעש. יותר נתונים לימדו את המילים שהיו בהם; הם לא שיפרו באופן ברור את היכולת לנחש מילים שהמודל מעולם לא שמע.
- Nemotron,** גרסת הסטרימינג שנמדדה במצב לא מקוון, מפגרת אחרי v1.0 בכל מדד, ולא כתבה נכון אף אחד מ-8 אסימוני המספרים במבחן. אין לה פסק דין לגבי שכחה, כי Nemotron המקורי לא נוקד מחדש על הגרסה הזו.

English did not get worse

Forgetting check on dataset revision 550bfcaa0f84. Scored 2026-10-19.



Source: Parakeet-V2-Personal-Finetune data/evals/r6-*-summary.json. LibriSpeech unrounded: 1.20% vs 1.23%, within noise.

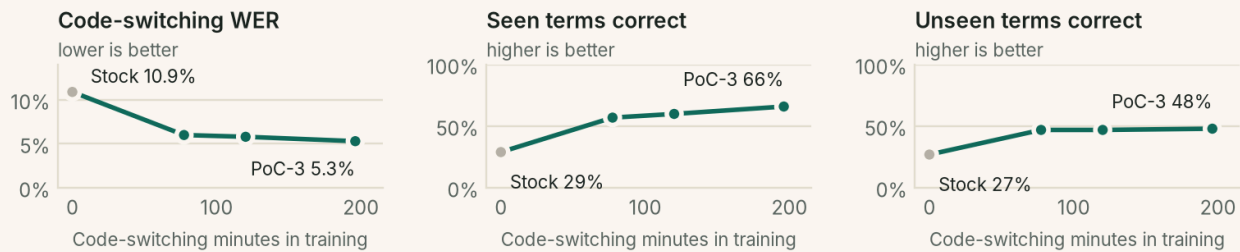
One speaker's voice: every test clip is read by the same speaker, so these numbers rank models on that speaker only.

איור 4. בדיקת שכחה. "אנגלית הדובר" היא אותו דובר; LibriSpeech הוא דוברים אחרים.

האנגלית לא נפגעה. ה-WER באנגלית של הדובר עצמו ירד מ-3.1% ל-1.7%, ובבקרת האנגלית של מבחן הייחוס הציבורי מ-2.7% ל-2.2%. ב-LibriSpeech, כלומר באנגלית של אנשים אחרים, v1.0 השיג 1.20% לעומת 1.23% במודל המקורי: הפרש ברמת הרעש, וזו בדיוק הנקודה. v1.0 עבר את סף השכחה של הפרויקט.

Proof-of-concept checkpoints

Stock, PoC-1, PoC-2 and PoC-3 on the older test split, revision b47310621f79 (219 clips).



Source: tracker README 'Results' (data/results.json). Not comparable with the 550bfcaa0f84 split. PoC-4, PoC-5 not benchmarked.

One speaker's voice: every test clip is read by the same speaker, so these numbers rank models on that speaker only.

איור 5. נקודות הביקורת של שלב הוכחת ההיתכנות על סט המבחן הישן (גרסה b47310621f79, 219 קטעים). אינו בר השוואה לאיורים 3 ו-4.

7. מגבלות

- קול של דובר אחד.** כל קטע, באימון ובמבחן, הוא של אותו דובר: יליד אירלנד, דובר אנגלית, גר בירושלים. התוצאות מדרגות מודלים עבור הדובר הזה. הן עדיין לא אומרות דבר על מבטאים אחרים. קולות נוספים הם הצורך העיקרי של הפרויקט.
- דיבור מוקרא, רובו מאותו טלפון.** משפטים מוכנים שמוקראים בקול נקיים יותר מהכתבה ספונטנית.
- סטים קטנים של מבחן.** 84 קטעים נקיים ו-172 מופעי מונחים במבחן הייחוס הציבורי; הפרש של 2 עד 3 נקודות בזיהוי הוא בתוך טווח הרעש. הסט המוחזק גדול יותר (387 קטעים, 293 מופעים מוחזקים) אבל עם אותו דובר.
- לא כל מונח בבנצ'מרק הוא עברי.** 22 מתוך 139 המונחים בתת-הסט הנקי הם מונחים טכניים באנגלית או שמות מוצרים (*Proxmox*, *grep*, *JSON*, ...), 26 מתוך 172 המופעים. הוצאתם לא משנה אף דירוג, אבל מורידה את שיעור הזיהוי של כל מודל ב-2 עד 6 נקודות. סט מבחן עם תיוג מתוקן ייצא כגרסה חדשה.
- נבדקת זליגה אפשרית של מונחים מוחזקים.** שני מונחים מרשימת המוחזקים, *mandatim* ו-*Beit Safafa*, מופיעים כמונחי יעד באימון בגרסה מאוחרת יותר של מערך הנתונים (*7bfcea1d*). עדיין לא אומת לאן נותבו בגרסה *550bfcaa*. אם אומנו עליהם, הפגיעות שלהם בסט המוחזק אינן "לא מוכרות".
- Nemotron נמדד במצב לא מקוון,** עם הקשר מלא, ולא במצב הסטרימינג שבשבילו הוא קיים. הדיוק שלו בסטרימינג לא נמדד.
- השיפור על פני PoC-3 בסט המוחזק בתוך טווח הרעש** (פרק 6). PoC-4 ו-PoC-5 מעולם לא נמדדו בבנצ'מרק.
- הייצואים המכווצים** נבדקו על חמישה קטעים כל אחד, ולא נמדדו בבנצ'מרק. MLX עדיין PoC-3.
- האיות הקנוני הוא מוסכמה.** מודל שכותב *tsaharon* במקום *tzaharon* אינו טועה במובן מוחלט; הבנצ'מרק מדווח לצד זה גם זיהוי מקל (*lenient recall*).

8. עלות

שלב	משאבי חישוב	עלות
ריצת אימון r6-en30	41.0 דקות על L40S אחד (Modal)	בערך US\$0.80 (חיוב Modal על ריצות קודמות מהסוג הזה: US\$0.71–0.86)
סבב בנצ'מרק, 2026-10-19	שש משימות L40S במקביל, 19:44–19:46 UTC, בערך 2.5 דקות זמן אמת כל אחת, פחות מ-15 דקות GPU בסך הכול	לכל היותר בערך US\$0.50

צוואר הבקבוק הוא ההקלטה, לא החישוב. אימון מחדש ומדידה של מודל בגודל הזה עולים בערך US\$1.30 בחישוב, ולכן הפרויקט מאמן מחדש על הנתונים המצטברים בכל פעם שהוקלט מספיק אודיו חדש בהחלפת קוד.

9. קישורים

- אתר ובנצ'מרקים: hebrish.app/benchmarks · מערכי נתונים: hebrish.app/datasets
- מודל: huggingface.co/danielrosehill/Parakeet-TDT-0.6B-v2-English-Hebrew-Codeswitching
- ייצואים: huggingface.co/Hebrish (GGUF, ONNX, int8, sherpa-onnx, MLX, Nemotron)
- נתוני אימון: [Read-Aloud-Voice-Sentences](https://hebrish.app/datasets#read-aloud-voice-sentences)
- טבלת דירוג: [English-Hebrew-Codeswitching-ASR-Leaderboard](https://hebrish.app/leaderboard) · נתוני הבנצ'מרק: [Hebrish/English-Hebrew-Codeswitching-ASR-Benchmark](https://hebrish.app/benchmark)
- המסמך הזה (עברית): hebrish.app/whitepaper/hebrish-v1.0-whitepaper-he.pdf · המקור באנגלית: hebrish.app/whitepaper/hebrish-v1.0-whitepaper.pdf

דניאל רוזהיל, 2026-10-19. תרגום לעברית: 2026-10-20; במקרה של סתירה, הנוסח האנגלי קובע. הטקסט מורשה ברישיון **CC-BY-4.0**. המודל יורש את רישיון **CC-BY-4.0** מ-[nvidia/parakeet-tdt-0.6b-v2](https://nvidia.com/en-us/parakeet-tdt-0.6b-v2); LibriSpeech-PC, ששימש לריענון, מורשה ב-CC-BY-4.0. כל התוצאות הן מקולו של דובר אחד. המספרים הועתקו מקובצי התוצאות של הפרויקט בלי עיגול מעבר לדיוק שבו פורסמו. התרשימים זהים לאלה של הנוסח האנגלי, והתוויות בהם באנגלית.