נייר עמדה טכני

Hebrish, מהדורת מודלים 1: Jachnun v1.0 ו-Sabich v1.0

המודלים היציבים הראשונים של Hebrish: שני כוונונים עדינים של NVIDIA Parakeet להכתבה באנגלית משולבת עברית. איך אומנו, איך יוצאו ואיך נמדדו, במה הם מצליחים, ומה המספרים לא מראים. בצירוף הערה על Shakshuka, קו הסטרימינג, שעדיין בעבודה.

מחבר
דניאל רוזהיל

תאריך
21/10/2026

מודלים: `Hebrish/Jachnun` (ריצה `r6-en30`) ו-`Hebrish/Sabich` (ריצה `r7-all`). מחליף את נייר העמדה "Hebrish v1.0" מ-2026-10-19 ואת הנספח שלו מ-2026-10-20.

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. מהדורת המודלים 1 של Hebrish היא סט המודלים היציב הראשון של הפרויקט: שני כוונונים עדינים מלאים (full fine-tune) של NVIDIA Parakeet TDT 0.6B v2, שכל אחד מהם למד לכתוב כל מילה עברית בתעתיק לטיני קנוני אחד (*makolet*, *Bituach Leumi*). **Jachnun v1.0** למד מ-339.2 דקות של דיבור מוקרא בהחלפת קוד (code-switching) ועוד כ-30% אנגלית. **Sabich v1.0** למד מאותו סוג הקלטות ועוד 187.4 דקות של הקלטות טכנולוגיה ומשימות, והוא המודל שאפליקציות Hebrish ממליצות עליו. במבחן הייחוס הציבורי להחלפת קוד אנגלית-עברית (סט מבחן v1, תת-הסט הנקי, 84 קטעים) משיג Jachnun שיעור זיהוי מונחים עבריים (term recall) של 73.3%, 70.1% על מונחים שלא הופיעו באימון ו-WER של 3.3%, ו-Sabich משיג 73.8%, 68.4% ו-3.3%, לעומת 32.6%, 32.5% ו-9.5% ב-Parakeet v2 המקורי. על 422 קטעים מוחזקים בצד הם שקולים בהחלפת קוד (WER של 5.42% ו-5.30%, המקורי 11.82%), Sabich טוב יותר בהכתבה טכנולוגית, והאנגלית הרגילה לא נפגעה. קו שלישי, **Shakshuka**, אותו מתכון על מודל הסטרימינג Nemotron של NVIDIA, הוא הוכחת היתכנות ואינו חלק מהמהדורה הזו. כל ההקלטות, באימון ובמבחן, הן בקולו של דובר אחד, ולכן המספרים האלה מדרגים מודלים עבור הדובר הזה בלבד.

73.8%
זיהוי מונחים של Sabich, מבחן ייחוס ציבורי (v2 המקורי: 32.6%)

73.3%
זיהוי מונחים של Jachnun; 70.1% על מונחים שמעולם לא הופיעו באימון

3.3%
WER, מבחן ייחוס ציבורי, שני המודלים (v2 המקורי: 9.5%)

1.54%
WER של Sabich בהכתבה טכנולוגית (Jachnun 1.74%, המקורי 2.51%)

1. מהדורה 1 במבט אחד

Hebrish הוא הפרויקט, לא מודל: הכתבה באנגלית שכותבת נבון את המילים בעברית, אפליקציות חינמיות לאנדרואיד, ל-Ubuntu ול-Windows, והמודלים והנתונים הפתוחים שמאחוריהן. קווי המודלים שלו נקראים על שם

מאבלים ישראליים. **מהדורת מודלים 1** היא המהדורה היציבה הראשונה של שני הקווים שבנויים על Parakeet; כל אחד מהם בגרסה **v1.0**. לפניה היו חמש נקודות ביקורת (checkpoints) של הוכחת היתכנות, PoC-1 עד PoC-5 (פורסמו ב-Hugging Face בשמות v1 עד v5 של `danielrosehill/Parakeet-TDT-0.6B-v2-English-Hebrew-Codeswitching`).

מודל	בסיס	אומן על	ריצה · תאריך	באפליקציות
Sabich v1.0	**Parakeet TDT 0.6B v2**	**החלפת קוד ועוד הקלטות טכנולוגיה ומשימות (623 דקות)**	**2026-10-20** · `r7-all`	**מומלץ (מאז 2026-10-21)**
Jachnun v1.0	**Parakeet TDT 0.6B v2**	**החלפת קוד ועוד אנגלית (483 דקות)**	**2026-10-19** · `r6-en30`	**בחירה שנייה**
Shakshuka PoC-2	Nemotron Speech Streaming EN 0.6B	ההקלטות והמתכון של Jachnun	`r6-nemotron-en30` · 2026-10-19	עדיין לא (פרק 9)

אימונים מחדש על הקלטות נוספות ייקראו Jachnun v1.1, v1.2 וכן הלאה (וכך גם Sabich). האימון מחדש הראשון של Shakshuka ייקרא Shakshuka v1.0.

2. 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 משאיר 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. הטוקנייזר 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.

3. מערך הנתונים

נתוני האימון הם Read-Aloud-Voice-Sentences (CC-BY-4.0) בבקשת גישה מאז 2026-10-20, ראו hebrish.app/datasets). 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 ומשפטי משימות מוקלטים באותו קול. הטבלה מציגה את החלוקה בגרסה (revision) `550bfcaa0f84`, זו שעליה Jachnun אומן ונמדד לראשונה.

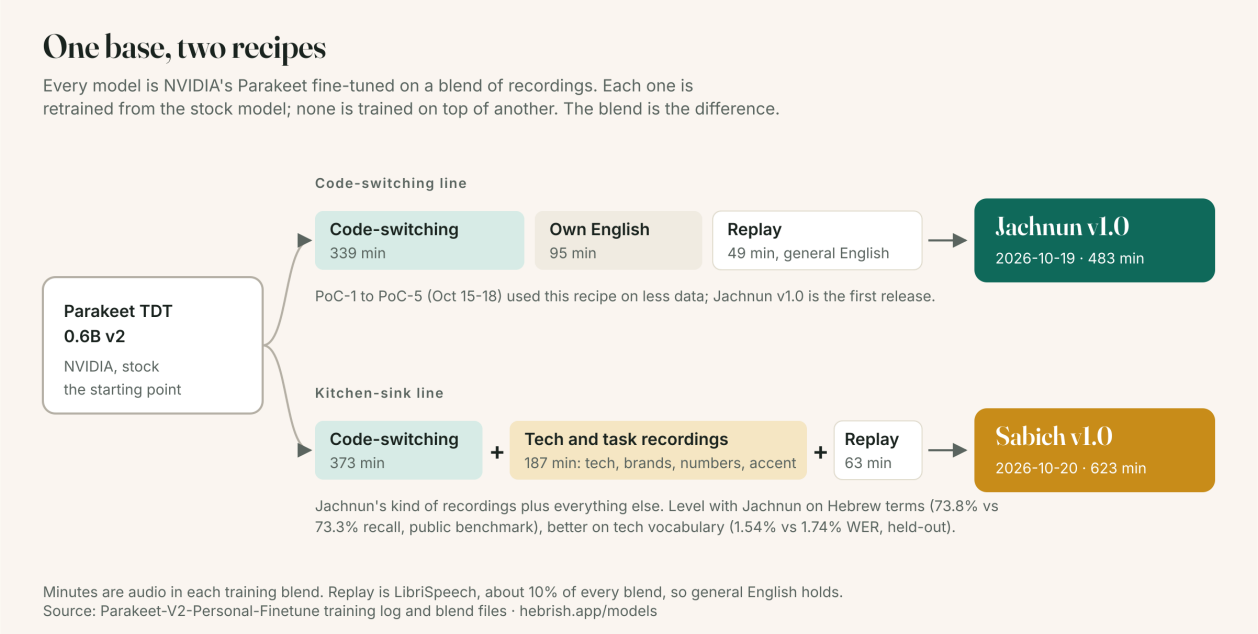
חלוקה בגרסה 550bfcaa	קטעים	דקות	שימוש
אימון החלפת קוד (cs_train)	3,224	339.2	אימון
אימות החלפת קוד (cs_val)	218	22.6	בחירת האפוק הטוב ביותר
מאגר האנגלית של הדובר	751	95.3	אימון (בולו)
מאגר ריענון LibriSpeech-PC	900	190.7	אימון (נדגמו 48.5 דקות)
מבחן החלפת קוד (test_cs)	387	43.9	הערכה
מבחן האנגלית של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

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.

4. שני מתכונים מבסיס אחד

שני המודלים מאומנים מ-nvidia/parakeet-tdt-0.6b-v2 המקורי; אף אחד מהם אינו מאומן על גבי השני. רק התמהיל שונה. Jachnun לומד מדיבור בהחלפת קוד ומהאנגלית הרגילה של הדובר. Sabich, שנקרא על שם הסביח, הפיתה שממולאת בקצת מכל דבר, לומד מכל משימה שהדובר הקליט.



איור 1. בסיס אחד, שני מתכונים. הדקות הן במות האודיו בכל תמהיל אימון.

תמהיל האימון (דקות)	Jachnun v1.0 (r6-en30)	Sabich v1.0 (r7-all)
החלפת קוד	339.2	373.1
הקלטות אחרות של הדובר	95.3 (אנגלית של הדובר)	187.4 (טכנולוגיה ומשימות)
ריענון LibriSpeech-PC	48.5	62.5 (10%)
סך הכול	483.0 (4,202 שורות)	623.0

187.4 דקות הקלטות המשימות של Sabich כוללות אוצר מילים טכנולוגי, שמות מותגים, ראשי תיבות, מספרים ומטבעות, אנגלית אירית, היסוסים ומילות מילוי, תיקונים עצמיים ומשימות נוספות; יחד עם החלפת הקוד אלה 560.5 דקות של הקלטות הדובר עצמו.

למה 30% אנגלית

מודל הכתבה שלומד מילים בעברית אבל מאבד את האנגלית לא שווה כלום. היעד לחלק האנגלית של Jachnun היה 30%, והתוצאה בפועל 29.8%: האנגלית של הדובר עצמו מלמדת שאנגלית רגילה בקול הזה צריכה להישאר אנגלית רגילה, וריענון LibriSpeech (replay) שומר על קולות אחרים. 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. Sabich שומר על ריענון של 10%; הקלטות המשימות שלו הן ברובן אנגלית באותו קול. פרק 8 בודק אם זה עבד.

למה כוונון מלא מצטבר ולא המשך אימון

כל גרסה מתחילה מ-Parakeet המקורי ומתאמנת על כל ההקלטות שנאספו עד אז. המשך אימון של הכוונון 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 נמשכת פחות משעה על GPU אחד, כך שהמשך אימון לא חוסך כלום. האימון והבנצ'מרק הם שלבים נפרדים: האפוק הטוב ביותר נבחר על נתוני אימות (חלק מהאימון, אף פעם לא מהמבחן), והבנצ'מרק רץ אחר כך על סט המבחן.

5. אימון

הגדרה	Jachnun v1.0	Sabich v1.0
סביבת עבודה	`nemo_toolkit[asr,cu12]` (NeMo 2.7.3), torch 2.11.0, נעולים בתמונת Modal	
נקודת מוצא	`nvidia/parakeet-tdt-0.6b-v2`, הטוקנייזר נשמר	
אופטימייזר	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 מעורב bf16, seed 7	
בחירה	אפוק 18 מתוך 20, WER אימות 5.0% (גולמי, 0.0497)	אפוק 19 מתוך 20, WER אימות 4.63%
חומרה / זמן	NVIDIA L40S אחד (48 GB) ב-Modal, 41.0 דקות	NVIDIA L40S אחד ב-Modal, 52.1 דקות

מהוכחת היתכנות למהדורה 1

נקודת ביקורת	ריצה	דקות CS	שוחרר	נמדד בבנצ'מרק
PoC-1	`pilot-en30`	77.4	2026-10-15	כן
PoC-2	`r2-en30`	120.1	2026-10-15	כן
PoC-3	`r3-en30`	196.1	2026-10-18	כן
PoC-4	`r4-en30`	269.3	2026-10-18	לא
PoC-5	`r5-en30`	294.2	2026-10-18	לא
Jachnun v1.0	`r6-en30`	**339.2**	**2026-10-19**	**כן, 2026-10-19**
Sabich v1.0	`r7-all`	**373.1**	**2026-10-20**	**כן, 2026-10-20**

PoC-1 עד PoC-5 השתמשו במתכון של Jachnun על פחות נתונים. Jachnun v1.0 הוא נקודת הביקורת השישית בקו הזה; הוא נקרא v1.0 כי זו הגרסה הראשונה שהפרויקט עומד מאחוריה. Sabich v1.0 אינו נקודת ביקורת מאוחרת יותר, אלא מתכון שני מהמודל המקורי.

6. ייצוא

שני המודלים מפורסמים בפורמטים שבהם משתמשות סביבות ההרצה על המכשיר (Hugging Face, בבקשת גישה). הייצואים של Jachnun נבדקו על חמישה קטעים מול ONNX ב-fp32 לפני כל בנצ'מרק.

פורמט	מאגרים	רץ ב-	בדיקה מול דיוק מלא
`.nemo`	`Hebrish/Jachnun` `Hebrish/Sabich`	NeMo (PyTorch)	ייחוס
sherpa-onnx int8	`Hebrish/Jachnun-sherpa-onnx-int8` `Hebrish/Sabich-sherpa-onnx-int8`	sherpa-onnx, אפליקציות Hebrish	Jachnun 4/5
ONNX ב-fp32 וב-int8	`Hebrish/Jachnun-ONNX` `Hebrish/Sabich-ONNX`	onnx-asr, Voxtype, ההדגמה בדפדפן	Jachnun int8 4/5 Sabich fp32 5/5, int8 3/5
GGUF F16, Q8_0, Q4_K_M	`Hebrish/Jachnun-GGUF` `Hebrish/Sabich-GGUF`	transcribe.cpp, Handy, Offline Voice Input	Jachnun F16 5/5, Q8_0 5/5, Q4_K_M 4/5
MLX	`Hebrish/Jachnun-MLX` `Hebrish/Sabich-MLX`	parakeet-mlx (Apple Silicon, עם `--fp32`)	Jachnun: 8/10 קטעים, Sabich 9/10

הטעויות של Jachnun ב-int8 הן מילה אחת: *tzatzumim* במקום שבו fp32 כתב *satzumim* התמלול המקורי הוא *tza'atzuim*. הטעויות של Sabich ב-int8 היו אות רישית אחת ופסיק אחד; הטעות שלו ב-MLX, פסיק אחד. ב-bf16, ברירת המחדל של parakeet-mlx, Jachnun פענח קטע אחד באסימונים לא ידועים, ומכאן fp32. תוצאות הבנצ'מרק שלהלן הן של קובץ `.nemo` ב-PyTorch, לא של הייצואים המכווצים (quantized).

7. הערכה

שני סטים של מבחן עונים 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. Jachnun נוקד בסבב 3 (2026-10-19), ו-Sabich בסבב 4 (2026-10-20, נעוץ לגרסה `640c1483` ב-Hugging Face).

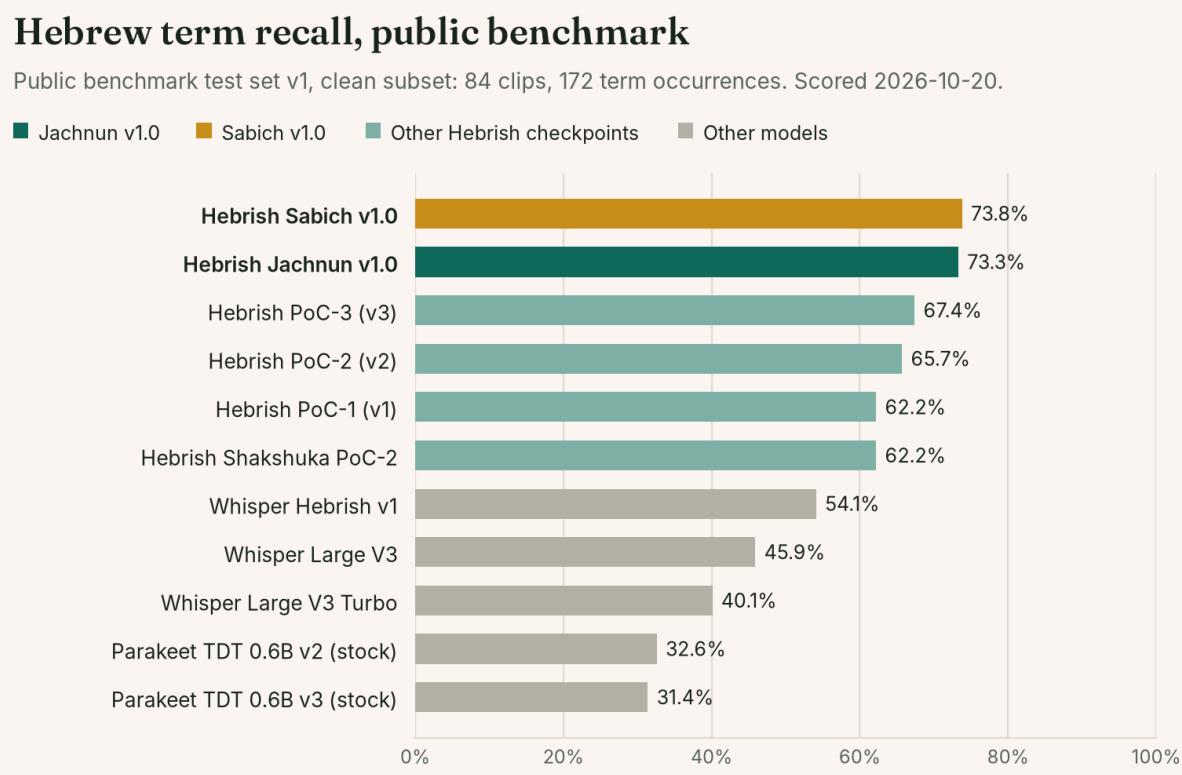
- **זיהוי מונחים (term recall)**: שיעור מופעי המונחים שהאיות הקנוני שלהם מופיע בתמלול כמילה שלמה, בלי תלות באותיות רישיות, בשמקף ורווח שקולים. **זיהוי מונחים לא מוכרים** מצמצם את המדד למונחים שאינם בחלוקת האימון הציבורית (117 מתוך 172 המופעים).
- **WER**: אותיות קטנות, מקפים הופכים לרווחים, פיסוק מוסר 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 Jachnun (`runs/r6-en30/train.jsonl`, 4,202 שורות), בנתוני האימון או האימות של Sabich, או בנתונים של ריצת Shakshuka.

סטי המבחן המוחזקים

סט המבחן של הפרויקט עצמו גדול יותר. משתמשים בשתי גרסאות שלו, ומשווים מספרים רק בתוך אותה גרסה: `550bfcaa0f84` (387 קטעים בהחלפת קוד, פרק 3), שעליה נוקדו v2 המקורי, PoC-3, Jachnun ו-Shakshuka ב-PoC-2 ב-2026-10-19; והגרסה החדשה יותר `f956dbb184c2` (סבב R7: 422 קטעים בהחלפת קוד, 84 קטעים באנגלית של הדובר, מהם 57 בלי עברית, 20 קטעי הכתבה טכנולוגית, 200 קטעי LibriSpeech ו-17 הקלטות ארוכות), שעליה נוקדו v2 המקורי, Jachnun ו-Sabich ב-2026-10-20. ה-WER מנורמל.

8. תוצאות

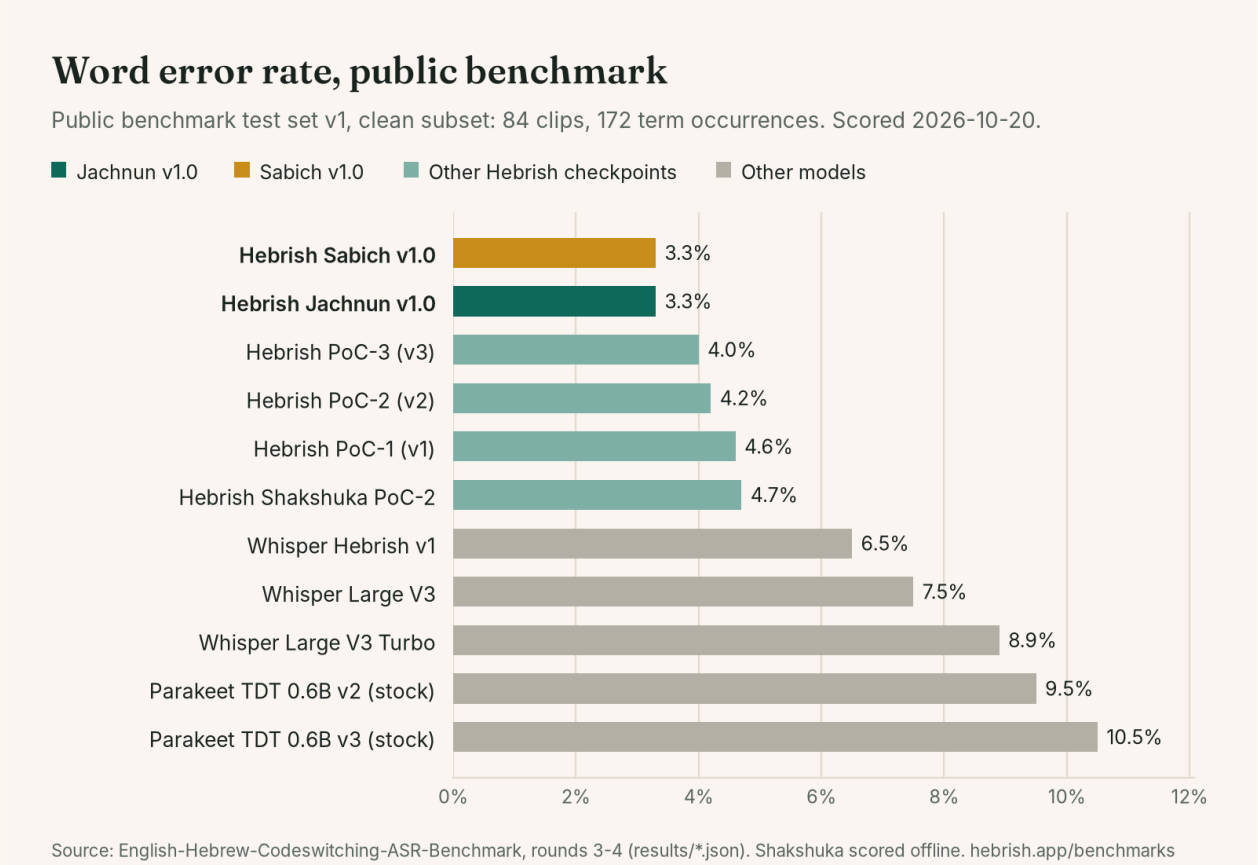
מבחן הייחוס הציבורי



איור 2. שיעור זיהוי המונחים העבריים במבחן הייחוס הציבורי, תת-הסט הנקי, כל אחד-עשר המודלים שנמדדו.

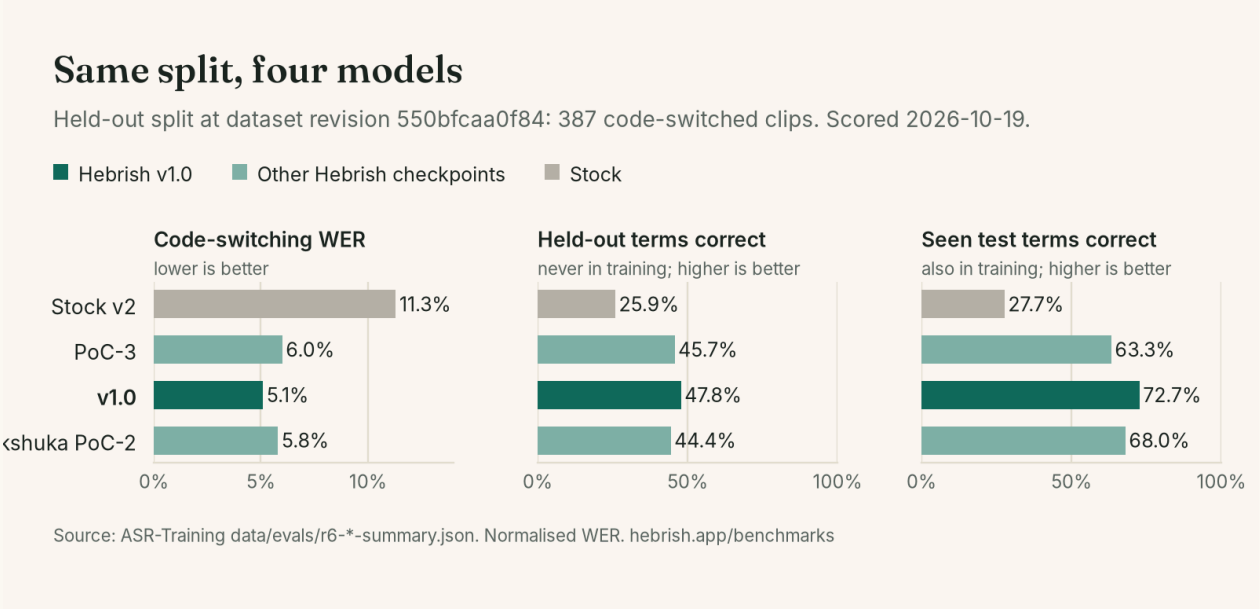
מודל (מבחן ייחוס ציבורי, תת-הסט הנקי)	זיהוי מונחים	מונחים לא מוכרים	WER	WER בבקרת אנגלית
Hebrish Sabich v1.0	**73.8%**	**68.4%**	**3.27%**	**1.99%**
Hebrish Jachnun v1.0	**73.3%**	**70.1%**	**3.32%**	**2.17%**
Hebrish PoC-3	67.4%	63.2%	4.0%	2.5%
Hebrish PoC-2	65.7%	60.7%	4.2%	2.4%
Hebrish PoC-1	62.2%	61.5%	4.6%	2.5%
Hebrish Shakshuka PoC-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%

על 172 מופעים, הפרש של 2 עד 3 נקודות בזיהוי הוא בערך 4 עד 5 פגיעות, בתוך טווח הרעש. בין Sabich ל-Jachnun ההפרש הוא פגיעה אחת מתוך 172 בזיהוי המונחים, ושתי פגיעות מתוך 117 בזיהוי המונחים הלא מוכרים.



איור 3. שיעור שגיאות מילים (WER) במבחן הייחוס הציבורי, תת-הסט הנקי, כל אחד-עשר המודלים שנמדדו.

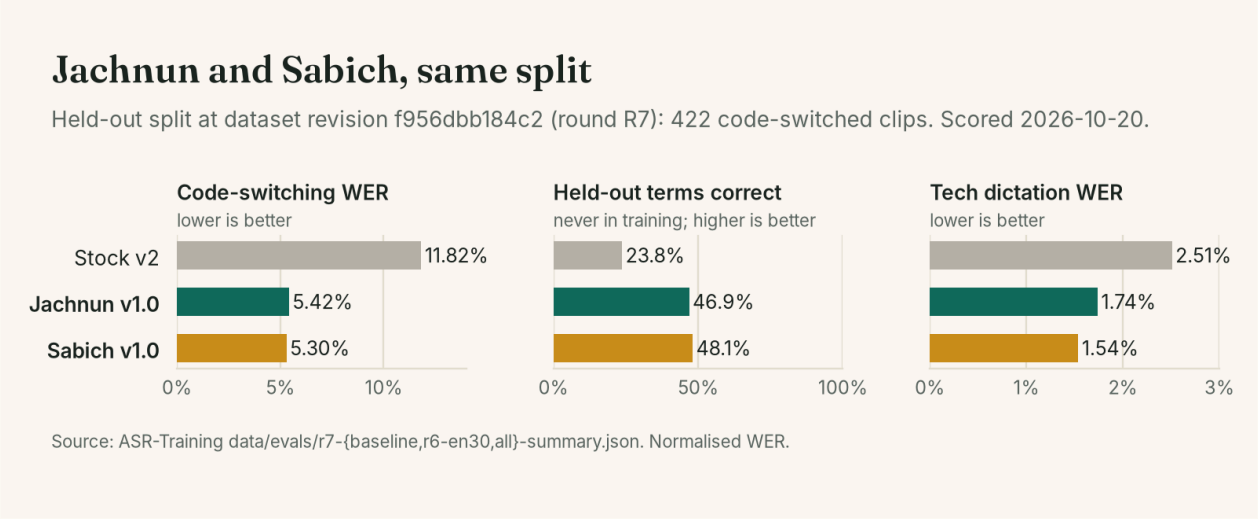
Jachnun על סט המבחן המוחזק (`550bfcaa`)



איור 4. v2 המקורי, PoC-3, Jachnun v1.0 ו-Shakshuka PoC-2 על אותו סט מבחן מוחזק.

הקלטות ארוכות	LibriSpeech	אנגלית הדובר	מונחים מוכרים (278)	מונחים מוחזקים (293)	CS-ב WER	550bfcaa
9.6%	1.2%	3.1%	27.7%	25.9%	11.3%	Parakeet v2 המקורי
5.8%	1.0%	2.3%	63.3%	45.7%	6.0%	Hebrish PoC-3
5.0%	**1.2%**	**1.7%**	**72.7%**	**47.8%**	**5.1%**	**Jachnun v1.0**
5.8%	2.0%	3.3%	68.0%	44.4%	5.8%	Shakshuka PoC-2 (לא מקוון)

Jachnun ו-Sabich על אותם קטעים (f956dbb1, סבב R7)

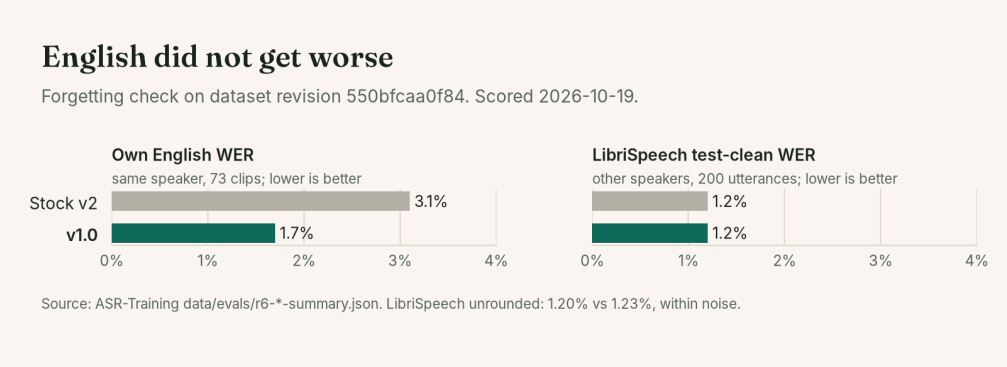


איור 5. v2 המקורי, Jachnun v1.0 ו-Sabich v1.0 על סט המבחן המוחזק של R7. חדש בנוסח הזה.

הקלטות ארוכות	LibriSpeech	טכנולוגיה	בלי עברית	אנגלית הדובר	מונחים מוחזקים	CS-ב WER	f956dbb1
10.34%	1.23%	2.51%	1.79%	2.81%	23.77%	11.82%	Parakeet v2 המקורי
5.84%	**1.20%**	**1.74%**	**1.23%**	**1.56%**	**46.91%**	**5.42%**	**Jachnun v1.0**
5.77%	**1.30%**	**1.54%**	**1.23%**	**1.62%**	**48.15%**	**5.30%**	**Sabich v1.0**

מונחים מוחזקים: שיעור מופעי המונחים מרשימת המוחזקים שנכתבו נכון (גבוה יותר עדיף). כל שאר העמודות הן WER (נמוך יותר עדיף). "בלי עברית" הוא תת-הסט של 57 קטעים מתוך האנגלית של הדובר שאין בהם מילים בעברית. הגרסה הזו חדשה וגדולה יותר מ-550bfcaa, ולכן המספרים של Jachnun כאן שונים מאלה שבטבלה הקודמת.

האנגלית לא נפגעה

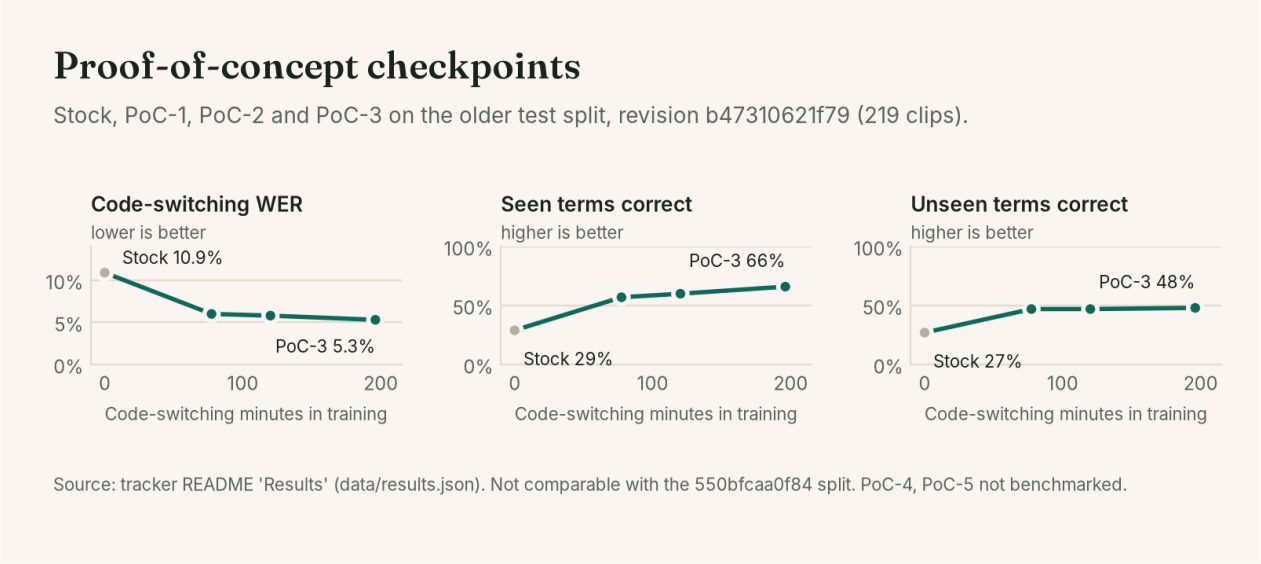


איור 6. בדיקת שכחה עבור Jachnun. "אנגלית הדובר" היא אותו דובר; LibriSpeech הוא דוברים אחרים.

ה-WER של Jachnun 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, הוא השיג 1.20% לעומת 1.23% במודל המקורי: הפרש ברמת הרעש, וזו בדיוק הנקודה; הוא עבר את סף השכחה של הפרויקט. אצל Sabich, האנגלית של הדובר (1.62%) והאנגלית בלי עברית (1.23%) נשארות הרבה מתחת למודל המקורי (2.81%, 1.79%). ב-LibriSpeech השיג Sabich 1.30% לעומת 1.23% במודל המקורי: כישלון צר בסף השכחה, בערך 3 מילים ב-200 קטעים, שדניאל רוזהיל קיבל כרעש.

מה התוצאות אומרות

- **מילים בעברית.** שני המודלים מכפילים יותר מפי שניים את שיעור זיהוי המונחים של Parakeet המקורי במבחן הייחוס הציבורי (73.3% ו-73.8%) לעומת 32.6%) ומובילים על פני כל המודלים שנמדדו, כולל Whisper Large V3 (45.9%) וכוונון ה-Whisper הקודם (54.1%). זיהוי המונחים הלא מוכרים של Jachnun (70.1%) קרוב לזיהוי הכולל שלו, מה שמרמז שהשיפור אינו רק שינון של מונחי האימון.
- **WER בהחלפת קוד** ירד מ-11.3% (מקורי) ל-5.1% בסט המוחזק `550bfcaa`, ומ-9.5% ל-3.3% במבחן הייחוס הציבורי.
- **Sabich לעומת Jachnun.** הנתונים הנוספים של טכנולוגיה ומשימות לא עלו דבר בהחלפת קוד (WER של 5.30% לעומת 5.42%, זיהוי מונחים מוחזקים של 48.15% לעומת 46.91%, ושוויון במבחן הייחוס הציבורי) והביאו שיפור מתון בהכתבה טכנולוגית (1.54% לעומת 1.74%). לכן Sabich הפך ב-2026-10-21 למודל המומלץ באפליקציות. ההפרשים קטנים, ומקורם בסבב מבחנים אחד על קול אחד.
- **לעומת PoC-3**, Jachnun 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.



איור 7. נקודות הביקורת של שלב הוכחת ההיתכנות על סט המבחן הישן (גרסה `b47310621f79`, 219 קטעים). אינו בר השוואה לאיורים 4 ו-5.

9. בעבודה: Shakshuka, קו הסטרימינג

> **לא חלק ממהדורה 1.** Shakshuka עדיין בבנייה. נקודות הביקורת שלו הן הוכחות היתכנות, הוא אינו באפליקציות Hebrish, והדיוק שלו בסטרימינג לא נמדד.

Shakshuka מיישם את ההקלטות והמתכון של Jachnun על מודל הסטרימינג של NVIDIA שמנהל מטמון הקשר (cache-aware), `nvidia/nemotron-speech-streaming-en-0.6b`, כך שהטקסט יכול להופיע בזמן שהדובר עדיין

מדבר. קיימות שתי נקודות ביקורת, שתיהן ב-`Hebrish/Nemotron-Speech-Streaming-EN-0.6B-English-Hebrew-Codeswitching` (בבקשת גישה; DOI `10.57967/hf/11107`):

נקודת ביקורת	ריצה	שוחרר	נתונים
Shakshuka PoC-1	`r5-nemotron-en30`	2026-10-18	אומן לצד PoC-5; בהיסטוריה של המאגר
Shakshuka PoC-2	`r6-nemotron-en30`	2026-10-19	הנתונים והמתכון של Jachnun v1.0 אפוק מיטבי 18, WER אימות 6.3% (גולמי), 45.8 דקות

בשהוא נמדד **במצב לא מקוון** (הקשר מלא), Shakshuka PoC-2 מפגר אחרי Jachnun בכל מדד: זיהוי מונחים של 62.2% ו-WER של 4.7% במבחן הייחוס הציבורי, ו-WER של 5.8% בהחלפת קוד בסט המוחזק `550bfcaa`. הוא לא כתב נכון אף אחד מ-8 אסימוני המספרים במבחן הזה. אין לו פסק דין לגבי שבחה, כי Nemotron המקורי לא נוקד מחדש על הגרסה הזו. הצעד הבא: אימון מחדש ראשון, Shakshuka v1.0, והערכה במצב סטרימינג, שבשבילו הקו הזה קיים.

10. מגבלות

- **קול של דובר אחד.** כל קטע, באימון ובמבחן, הוא של אותו דובר: יליד אירלנד, דובר אנגלית, גר בירושלים. רק הריענון ו-test-clean של LibriSpeech הם דוברים אחרים. התוצאות מדרגות מודלים עבור הדובר הזה, והן 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. הסטים המוחזקים גדולים יותר, 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 (*JSON*, *grep*, *Proxmox*, ...), 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".
- **השיפור על פני PoC-3 בסט המוחזק בתוך טווח הרעש** (פרק 8). PoC-4 ו-PoC-5 מעולם לא נמדדו בבנצ'מרק.
- **הייצואים המכווצים** נבדקו על חמישה או עשרה קטעים כל אחד, ולא נמדדו בבנצ'מרק.
- **לא נמדד:** קיימת גרסה של Sabich עם 30 אפוקים, `r7-all-e30` (WER אימות 4.17%), אבל היא לא נוקדה על אף אחד מסטי המבחן.
- **האיות הקנוני הוא מוסכמה.** מודל שכותב *tsaharon* במקום *tzaharon* אינו טועה במובן מוחלט; הבנצ'מרק מדווח לצד זה גם זיהוי מקל (lenient recall).

11. חישוב

שלב	משאבי חישוב
ריצת אימון (Jachnun) `r6-en30`	41.0 דקות על L40S אחד (Modal)
ריצת אימון (Sabich) `r7-all`	52.1 דקות על L40S אחד (Modal)
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

צוואר הבקבוק הוא ההקלטה, לא החישוב. אימון מחדש ומדידה של מודל בגודל הזה נמשכים בערך שעה של זמן GPU, ולכן הפרויקט מאמן מחדש על הנתונים המצטברים בכל פעם שהוקלט מספיק אודיו חדש.

12. קישורים

- אתר, מודלים ובנצ'מרקים: hebrish.app/models · hebrish.app/benchmarks · מערכי נתונים: hebrish.app/datasets
- Jachnun v1.0: Hebrish/Jachnun · Sabich v1.0: Hebrish/Sabich (בבקשת גישה); כל הגרסאות והייצואים ב-huggingface.co/Hebrish
- Shakshuka: Hebrish/Nemotron-Speech-Streaming-EN-0.6B-English-Hebrew-Codeswitching
- תיעוד שלב הוכחת ההיתכנות: danielrosehill/Parakeet-TDT-0.6B-v2-English-Hebrew-Codeswitching (PoC-1 עד PoC-5)
- טבלת דירוג: English-Hebrew-Codeswitching-ASR-Leaderboard · נתוני הבנצ'מרק: `Hebrish/English-Hebrew-Codeswitching-ASR-Benchmark`
- המסמך הזה (עברית): hebrish.app/whitepaper/hebrish-release-1-whitepaper-he.pdf · המקור באנגלית: hebrish.app/whitepaper/hebrish-release-1-whitepaper.pdf

---

דניאל רוזהיל. הנוסח הזה, מ-2026-10-21, ממסגר מחדש את נייר העמדה הקודם "Hebrish v1.0" (2026-10-19, עם נספח על Sabich v1.0 מ-2026-10-20) כמהדורת מודלים 1 שמכסה את שני המודלים, ומוסיף את פרק 9 על Shakshuka ואת איור 5. אף תוצאה לא השתנתה. תרגום לעברית: 2026-10-21; במקרה של סתירה, הנוסח האנגלי קובע. הטקסט מורשה ברישיון **CC-BY-4.0**. המודלים יורשים את רישיון CC-BY-4.0 מ-`nvidia/parakeet-tdt-0.6b-v2`; LibriSpeech-PC, ששימש לריענון, מורשה ב-CC-BY-4.0 הבסיס של Shakshuka מורשה ברישיון NVIDIA Open Model License. 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.