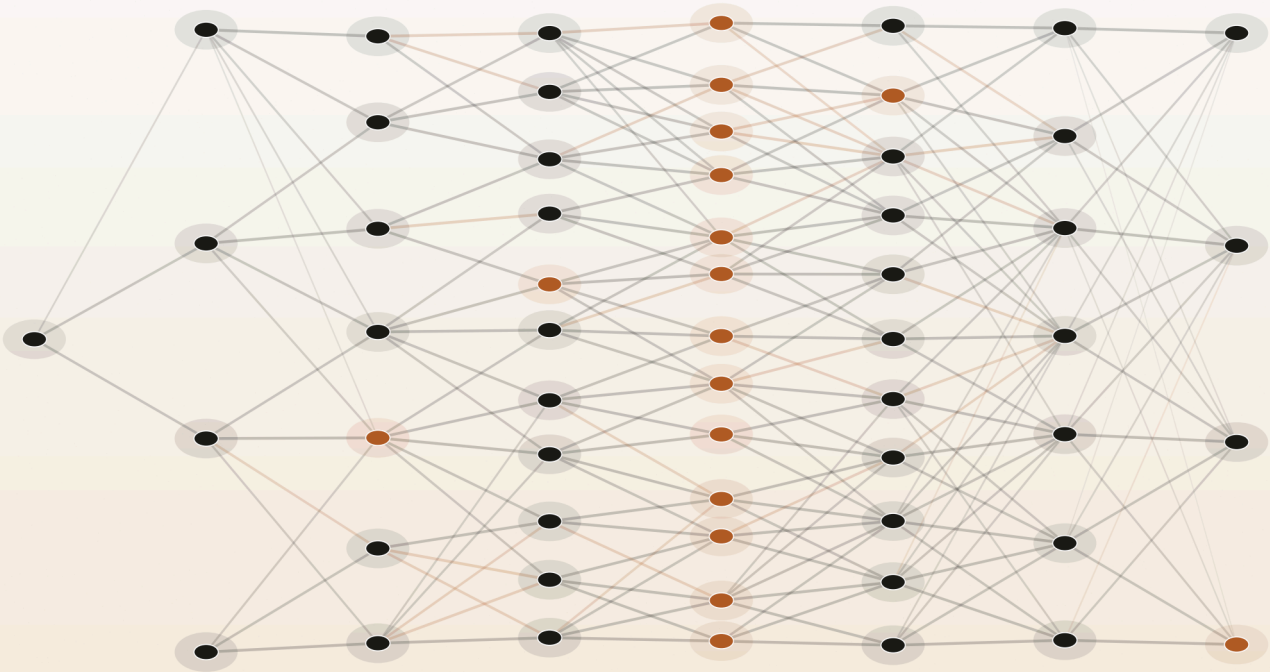
מדריך שדה מעשי

בנה לעצמך *בינה מלאכותית*, ב־2026

מבט כן על שני המסלולים האמיתיים — מאפס מוחלט בברזל החשוף, ומיזוג רב־מודלים — עם מחירים, מאזני שיקולים, ומורה גיטרה דובר עברית כדוגמה רצה.



מאת עידן

04 · 05 · 2026

תוכן עניינים

חלק ראשון

א

שני המסלולים, בכנות

> *"להעמיד פנים שאתם על מסלול ה־bare-metal כשבעצם אתם עושים מיזוג — זאת הטעות הנפוצה ביותר בשדה כולו."*
>
> מתוך פרק 2

פרק 1

מה זה בעצם «לבנות בינה מלאכותית משלך» ב־2026

שני מסלולים אמיתיים. אף אחד מהם אינו שגוי. רוב הבונים האמיתיים בוחרים בשני, ומעמידים פנים שהם הולכים בראשון.

אם תיכנסו לחדר מלא באנשים שאומרים שהם «בנו לעצמם בינה מלאכותית» ב־2026, תגלו שתי סוגי עבודה שונים מאוד שמסתתרים תחת אותו ביטוי.

הסוג הראשון הוא מסלול הברזל החשוף. אתם כותבים את לולאת ה־autograd בעצמכם, ממשים את מנגנון ה־attention מאפס, אוספים מאגר נתונים, צופים בעקומת ה־loss יורדת לאורך שעות או שבועות של זמן GPU, ובסוף יש לכם מודל שכל משקל בו מובן לכם משום שאתם הנחתם אותו שם. זהו מסלול ה*ריבונות והלמידה*. זהו גם המסלול שעוצר במידת ה«צעצוע», שכן אימון מודל בן 7 מיליארד פרמטרים מאתחול אקראי עולה בין חמישים אלף לחמש מאות אלף דולר אמריקאי במחשוב — לא דבר שעושים מהשולחן שלכם בתל אביב.

הסוג השני הוא מיזוג רב־מודלים. אתם לוקחים את המודלים פתוחי־המשקלים שמעבדות החזית שחררו ברבעון האחרון — Llama 3.3, Qwen 3, DeepSeek-V3, Whisper, Stable Diffusion, MediaPipe — וממזגים אותם, מכוונים אותם, משרשרים אותם, חותכים ומשלבים אותם למשהו שמתאים לתרחיש שמחבריהם המקוריים מעולם לא דמיינו. זהו מסלול ה*מעשיות והשילוח*. זה מה שכמעט כל מוצר בינה מלאכותית עצמאי ומוצלח ב־2026 הוא בפועל, מתחת לשיווק.

> *המשפט הכן: ב־2026, «בניתי לעצמי בינה מלאכותית» משמעו כמעט תמיד «הרכבתי וכוונתי ערימה של משקלים פתוחים לבעיה הספציפית שלי.» זה לא דבר פחות. זה דבר אחר ולגיטימי באותה מידה.*

§ למה המדריך הזה קיים

האינטרנט מלא בשני סוגי כשל בנושא הזה. הראשון הוא מסלול־מדריך־היוטיוב: «בנה טרנספורמר ב־200 שורות PyTorch!» — נכון טכנית, חסר תועלת לכל דבר שתרצו לפרוס. השני הוא מסלול־היוקר: «פשוט תקרא ל־API של OpenAI» — שזה בסדר עד שאתם צריכים עברית, או עבודה לא מקוונת, או הנתונים שלכם, או שאתם לא רוצים לשלם ארבעים סנט לכל שיחה ארוכה לנצח.

המדריך הזה בוחר באמצע הכן. הוא מתייחס לשני המסלולים כשני אחים — אחד ללמוד לעומק ואחד לשלח באופן שימושי — ועובר את אותה דוגמה, מורה גיטרה דובר עברית מבוסס בינה מלאכותית, בשניהם. הוא מדביק מחירים אמיתיים מ־2026 הנחתו בישראל על החומרה. הוא עושה את חישוב שלוש השנים של בעלות על תחנת

עבודה לעומת שכירת GPU בענן. וכשצריך, הוא מודה שהתשובה הפופולרית שגויה.

§ הדוגמה הרצה: מורה גיטרה דובר עברית

לאורך המדריך נחזור לאותו פרויקט: בינה מלאכותית שצופה במתחיל מנגן בגיטרה דרך מצלמת אינטרנט, מאזינה דרך מיקרופון, ונותנת משוב מדובר בעברית. במפורש, היא חייבת:

- להבין שאלות בעברית מדוברת מצד התלמיד («למה האקורד הזה רוטט?»)
- לזהות את מיקום היד על הסולם (פרטבורד) מתוך הווידאו
- לשמוע את מה שהתלמיד מנגן ולזהות גובהי צליל ותזמון
- להפעיל היגיון פדגוגי — «אתה מנגן נמוך במיתר ה־G כי האצבע שלך רחוקה מדי מהסריג»
- להשיב בעברית טבעית, עם טרמינולוגיה שמתחיל יכול לעקוב אחריה

זהו פרויקט אמיתי, ספציפי, וברי־השגה לאדם אחד ב־2026. יש לו גם את התכונה ששום מודל מוכן מהמדף לא יכול לעשות את כל מה שדרוש. חייבים לשלב. זה מה שהופך אותו לדוגמה ההוראתית הנכונה.

§ מה תמצאו לפניכם

פרק 2 פורש את שני המסלולים זה לצד זה, עם מספרי המאמץ והעלות הכנים. פרקים 3 ו־4 לוקחים את מורה הגיטרה לאורך כל אחד מהם בצורה ממשית. פרק 5 מדביק מחירי 2026 אמיתיים על שלוש שכבות חומרה ביתיות. פרק 6 משווה אותן להשכרת GPU בענן ומחשב את נקודת האיזון. פרק 7 מכסה את הדברים שהאנשים שוכחים — חשמל, פחת, הזמן שלכם. פרק 8 עוסק במשמעות האמיתית של הנתונים שלכם בכל אפשרות. פרק 9 הוא עמוד אחד של ערכת התחלה מומלצת למישהו במצב המדויק שלכם.

> לא נהיה מתחסדים. כשמסלול אחד עדיף בבירור, נאמר זאת. כשהתשובה הפופולרית שגויה, נאמר גם את זה.

פרק 2

שני המסלולים, זה לצד זה

ברזל חשוף מאפס מוחלט, או למזג את מה שכבר קיים. הטבלה הכנה.

לפני שנלך עם מורה הגיטרה דרך כל מסלול, נתבונן במה כל מסלול הוא בפועל, מה הוא עולה, ומה הוא נותן בחזרה. המספרים בפרק הזה אינם משאלות — הם משקפים את מה שאדם אחד, שעובד מהשולחן בישראל ב־2026, יכול בפועל לעשות.

מסלול א׳ — ברזל חשוף מאפס

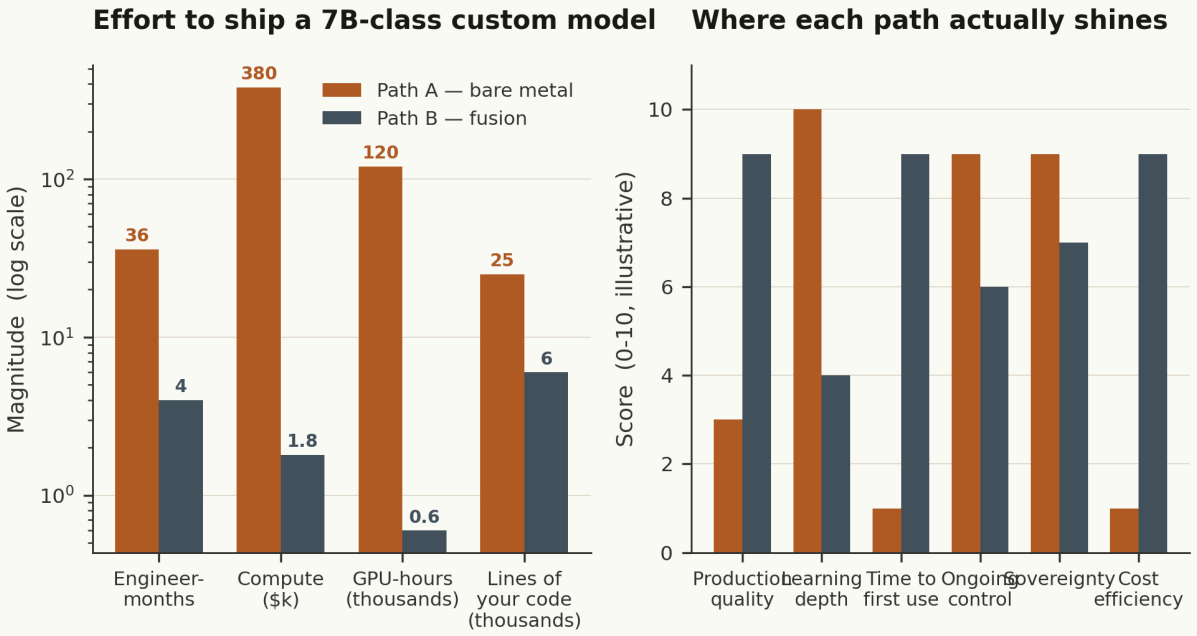
- אתם כותבים את ארכיטקטורת המודל ב־PyTorch או JAX
- אתם מרכיבים או מאמנים מראש על מאגר משלכם
- אתם בעלים של כל משקל וכל החלטת עיצוב
- תקרה ביתית מציאותית: מודלים בני ~100M–1B פרמטרים, או כיוונון של גדולים יותר
- הכי טוב ל: למידה עמוקה, ארכיטקטורות מותאמות אישית, ריבונות על מודל נישתי
- הכי גרוע ל: לשלח מוצר שימושי השנה

מסלול ב׳ — מיזוג רב־מודלים

- אתם לוקחים מודלים פתוחי־משקלים מ־Hugging Face
- אתם מכוננים את החלקים שזקוקים להתמחות תחומית (LoRA / QLoRA)
- אתם משרשרים אותם בקוד דבק משלכם — התזמור הוא שלכם
- תקרה ביתית מציאותית: עוזרי עברית בדרגת ייצור, צינורות רב־מודאליים
- הכי טוב ל: שילוח בשבועות לא בשנים, באיכות בדרגת חזית
- הכי גרוע ל: לטעון ש«בניתם את המודל» — לא בניתם, הרכבתם

§ ההשוואה הכנה

תמונה אחת שווה כאן פסקה. הגרף משמאל הוא בסקאלה לוגריתמית — אימון מקדים בברזל חשוף הוא בערך שני סדרי גודל יותר מאמץ ממיזוג עבור אותה יכולת מוגמרת. הגרף מימין מראה למה כל מסלול *טוב* בפועל. הם אינם יתרים; כל אחד מנצח בממדים אחרים.



איור 2.1 משמאל: הבדל סדר גודל במשאבים נדרשים לשילוח מודל מותאם בדרגת 7B מאפס לעומת ממשקלים פתוחים. מימין: היכן כל מסלול באמת מנצח. מספרים להמחשה לפרויקט פיתוח של אדם יחיד; עלויות אימון מקדים מבוססות על נתוני מחשוב שפורסמו עבור Llama-3 ו־DeepSeek-V3, מקנה־מידה ל־7B.

§ מה זה באמת «מאפס» ב־2026

זהו המקום לפוצץ פנטזיה מסוימת. אימון מקדים של מודל שפה כללי שימושי בן 7 מיליארד פרמטרים מאתחול אקראי דורש כ־1.5 טריליון טוקנים של נתוני אימון ו־200,000 שעות־H100 של מחשוב. במחירי הענן הנוכחיים זה בין 250,000 ל־500,000 דולר. רק חשמל בבית עבור אותו זמן ריצה היה מגמד את חשבון החומרה שלכם. **לא תאמנו 7B באיכות חזית מאפס בבית ב־2026. אף אחד לא יעשה זאת.**

אז כשהמדריך הזה אומר «מסלול א' — ברזל חשוף מאפס», הוא מתכוון לאחד משלושה דברים כנים, בסדר יורד של תכיפות:

1. **מודלי צעצוע להבנה.** טרנספורמר ברמת תווים בן 10 מיליון פרמטרים שאומן על קורפוס ציבורי, בשעתיים על ה־4070 שלכם. זהב חינוכי. לא מוצר.
2. **ארכיטקטורות מותאמות קטנות למשימות צרות.** מקודד בן 50M פרמטרים לניתוח סימוני אקורדים בעברית. מתאמן בלילה על 4090. שימושי כרכיב בתוך ערימת מיזוג גדולה יותר.
3. **כיוונון ככד של משקלים פתוחים כאילו היה מאפס.** קחו את Llama-3.3-8B, הריצו אימון מקדים מתמשך על 50 GB של טקסט עברי והקורפוס שלכם, ואז כוונו להוראות. כמה ימים על קופסת שכבה 2. רוב מה שאנשים ב־2026 מתכוונים אליו ב«אימנתי מודל משלי» הוא זה.

◆ **האמת הלא־רומנטית**

פריט 3 לעיל הוא מה שמשולח. פריט 1 הוא מה שעליו כותבים בלוג. להעמיד פנים שפריט 1 הוא פריט 3 היא הטעות הנפוצה ביותר בכל התחום הזה. בונים כנים יודעים מה הם עושים היום ולא מבלבלים בין השניים.

§ אז מה לבחור?

התשובה הנכונה כמעט לכולם שקוראים את זה היא **שניהם, ברצף**. הקדישו שבועיים למסלול א' — בנו טרנספורמר קטן מאפס במסורת ה־nanoGPT של קרפתי או מהספר של סבסטיאן רשקה — לא בגלל שהמודל המתקבל שימושי, אלא כי שום דבר אחר נותן את אותה אינטואיציה למה שקורה בתוך המשקלים הפתוחים שתבלו את השנתיים הבאות בכיוונון שלהם. ואז עברו למסלול ב' לכל מה שאתם באמת רוצים לפרוס.

אם תתעלמו ממסלול א' לחלוטין, תמצאו את עצמכם מבצעים מתכוני כיוונון שאינכם מבינים והניפוי שלכם יהיה אומלל. אם תחיו על מסלול א' לנצח, תבזבזו חמש שנים על מה שהיה אמור להיות פרויקט בן שישה שבועות. הגבול ביניהם הוא המקום הכי שימושי לחיות בו.

פרק 3

מסלול א' — מורה הגיטרה, מאפס

אם הייתם הולכים בכביש הברזל החשוף מקצה לקצה, איך זה היה נראה? תרשים כן של החודשים שלפניכם.

בעמיד פנים, לאורך הפרק הזה, שאתם עקשנים. אתם רוצים לבנות את מורה הגיטרה דובר העברית עם תלות מינימלית במודלים מאומנים מראש של אחרים. איך נראית תוכנית הפרויקט בפועל?

§ מה הייתם צריכים לבנות, ביד

חמש תת־מערכות אמיתיות, כל אחת תת־פרויקט בפני עצמה ארוך־חודשים:

1. **מזהה דיבור עברי.** מודל אקוסטי ומודל שפה מותאמים שאומנו על מאות שעות הדיבור הציבורי בעברית הקיימות (קורפוס ivrit.ai, Mozilla Common Voice he, הקלטות הקול שלכם). זה לבד פרויקט מחקר של חצי שנה לאדם יחיד, והתוצאה תהיה גרועה במידה ניכרת מ־Whisper-large-v3 מהמדף עם כיוונון ivrit.ai שתוכלו להוריד הערב.
2. **מודל ראייה שמוצא אצבעות על סולם הגיטרה.** תתייגו אלפי פריימים של וידאו גיטרה, תאמנו מודל קונבולוציה או טרנספורמר ראייה מאפס, ותבנו את לוגיקת קואורדינטות הסולם מקואורדינטות פיקסל גולמיות. MediaPipe Hands עושה את חלק נקודות־הציון של היד בחינם; אתם תמציאו אותו מחדש.
3. **מזהה גובה צליל / אקורדים.** רשת קונבולוציה קטנה על ספקטרוגרמות CQT. זה אכן בר־ביצוע מאפס בכמה שבועות — התחום בוגר מספיק כדי שיש מדריכים אקדמיים טובים.
4. **מודל חשיבה פדגוגי.** מודל שפה קטן דובר עברית שיודע מספיק על פדגוגיה גיטרית כדי לתת משוב שימושי. מאפס: לא בר־ביצוע. (ראו פרק 2 על עלויות אימון מקדים.) אפילו כיוונון של מודל בן 1B מאתחול אקראי על קורפוס מאוצר ידנית לא ישתווה ל־LoRA של 30 דקות מעל Llama-3.3.
5. **מסנתז דיבור עברי (טקסט־לדיבור).** בערך אותה צורה של פרויקט כמו מזהה הדיבור — כבול לקורפוס, חצי שנה מינימום, לא ינצח את Coqui XTTS-v2 עם כיוונון לעברית.

§ סדר־הגודל הכן

אם אתם מהנדסי ML מוכשרים שעובדים לבד, במשרה מלאה, עם גישה לקופסת שכבה 2 ביתית (RTX 4090, 64 GB RAM): **שלוש עד ארבע שנות עבודה, והתוצאה תהיה גרועה מבחינה מדידה בכל ציר ביחס למה שתוכלו לשלח דרך מסלול ב' בעשרה סופי שבוע.** החומרה לא תהיה הצוואר של הבקבוק שלכם. הנתונים יהיו.

! זה לא פסימיות

מעבדות חזית מוציאות תקציבים בני תשע ספרות ומעסיקות מאות אנשים כדי לייצר את המשקלים הפתוחים שאתם יכולים להוריד בחינם. הסיבה שמסלול א' מפסיד אינה שאתם לא מספיק חכמים. היא שהמשקלים הפתוחים כבר ספגו כל כך הרבה ממחשוב העולם, שזה לא־רציונלי לחזור על העבודה ההיא.

§ אז למה שמישהו יבחר במסלול א'?

שלוש סיבות לגיטימיות:

ללמוד לעומק.

מימוש attention מכפלי מטריצות גולמיות, צפייה בעקומת ה־loss יורדת על ה־GPT הצעצועי שלכם, ניפוי LoRA שמומש מאפס — כך אתם בונים את האינטואיציה שמאפשרת לכם לנפות את עבודת מסלול ב'. כל מהנדס בינה מלאכותית רציני צריך לעשות זאת פעם אחת.

לחקור רעיון אמיתי וחדש.

אם יש לכם ארכיטקטורה חדשנית (מנגנון attention שונה, קידוד מיקום שונה, חלופה לא־טרנספורמרית), הדרך היחידה לבדוק אותה היא מאפס. משקלים פתוחים אופים בחירות ארכיטקטוניות מסוימות.

לריבונות על מודל זעיר וצר.

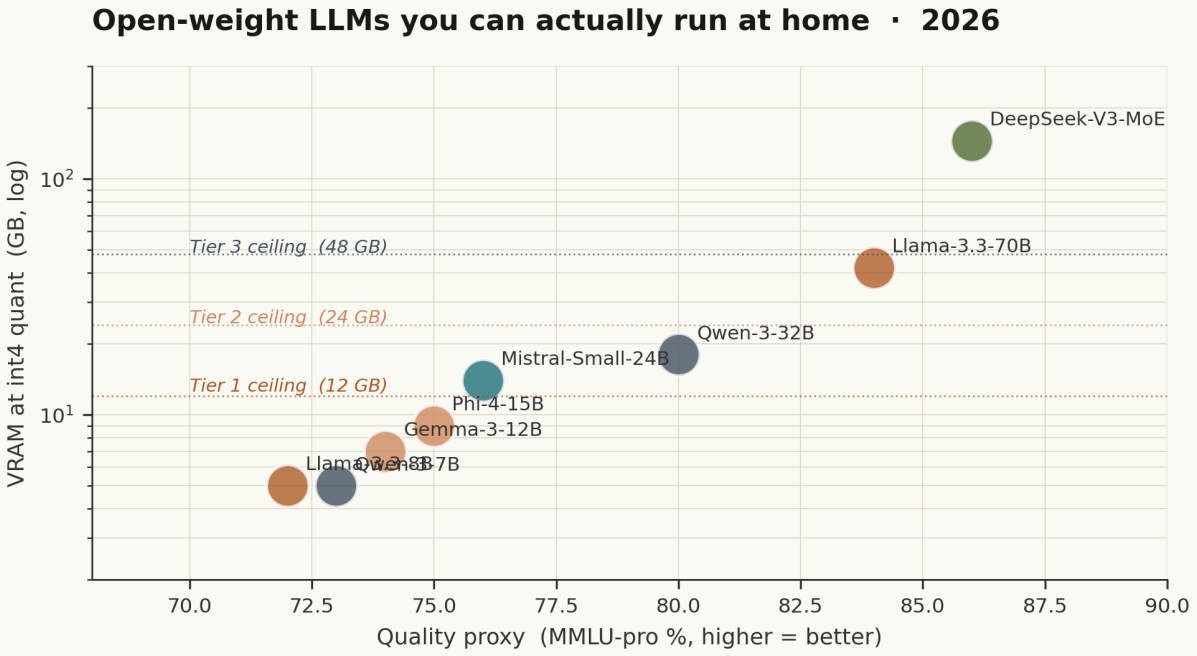
אם אתם רוצים מודל שמזהה שמות מקומות בעברית מתוך עדויות שואה, או מסווג סגנונות ציטוט תלמודיים, אף אחד לא ישחרר משקלים מאומנים מראש לכך. מודל קטן מאפס על הנתונים שלכם הוא התשובה הנכונה.

§ שבוע הלמידה המינימלי של מסלול א'

הנה מה שאנחנו ממליצים על «עשו את מסלול א' פעם אחת» כטקס מעבר. הקצו שבעה ימים. קחו את *nanoGPT* של קרפתי או את *Build a Large Language Model From Scratch* של רשקה כעמוד השדרה. בסוף השבוע יהיה לכם:

- טרנספורמר ברמת תווים בן 10 מיליון פרמטרים שאומן על התנ"ך (נחלת הכלל, 1.2 MB), שמייצר שטויות עבריות תקינות מבחינה תחבירית
- הבנה עובדת של: attention, זרמי שאריות, נורמליזציית שכבה, גרף ה־autograd, הלולאה הפנימית של האופטימייזר, פיצול אימון/אימות, צבירת גרדיאנטים, דיוק מעורב
- אותה אינטואיציה שתגרום לכם, ביום השמיני כשתעברו למסלול ב', להבין מיד למה דרגת LoRA של 16 שונה מדרגה 64

שבוע ההשקעה הזה הוא הלמידה בעלת המנוף הגבוה ביותר בכל התחום. הוא אינו תחליף למסלול ב'. הוא הבסיס שגורם למסלול ב' לא להרגיש כמו קסם.



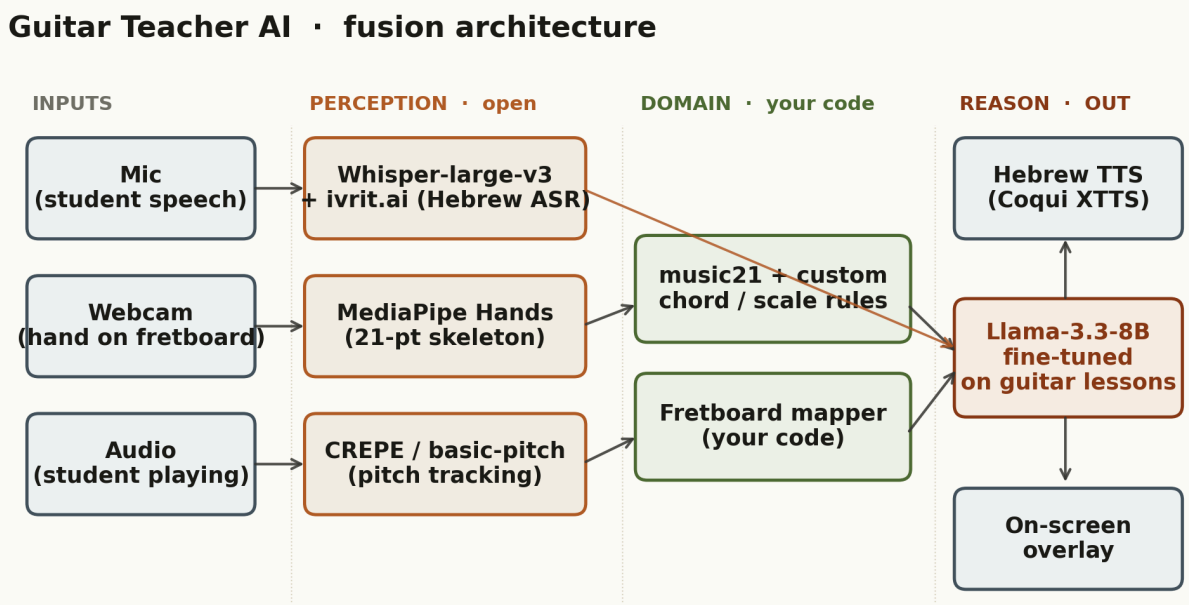
איור 3.1 הסיבה שמסלול א' מפסיד בעבודת ייצור: כל נקודה כאן מייצגת עשרות מיליוני דולרים של מחשוב שאתם הייתם מוציאים מחדש. מדד איכות הוא קירוב לאחוז MMLU-pro; VRAM הוא בקוונטיזציית int4. מקורות: כרטיסי מודל רשמיים ב־Hugging Face, צילומי לוח lmsys.org רבעון ראשון 2026.

פרק 4

מסלול ב' — מורה הגיטרה, במיזוג

איך באמת נראה לשלח את אותו פרויקט על ידי שילוב של משקלים פתוחים. רכיבים מוחשיים, ערכות כלים אמיתיות, לוח זמנים של עשרה סופי שבוע.

עכשיו הגרסה הכנה. הנה מה שמפתח עצמאי בתל אביב היה בונה בפועל בחודשיים הקרובים כדי לשלח גרסה אחת של מורה הגיטרה דובר העברית. כל רכיב שמוזכר כאן הוא אמיתי, חינמי, וברי־הורדה היום.



איור 4.1 ערימת המיזוג של מורה הגיטרה. שלושה זרמי קלט (מיקרופון, מצלמה, אודיו) זורמים דרך שלושה מודלי תפיסה פתוחי־משקלים, ואז דרך קוד התחום שלכם (החלק היחיד שאתם באמת חייבים לכתוב), ואז לתוך Llama-3.3 מכוונן לחשיבה פדגוגית, וחזרה החוצה כדיבור עברי וכשכבת־על על המסך.

§ הרכיבים, בשמותיהם

בעבודה דרך התרשים מקלט לפלט:

רכיבים פתוחי־משקלים למורה הגיטרה דובר העברית, 2026

תת־מערכת	מודל פתוח־משקלים	VRAM (INT8)	איפה זה חי
זיהוי דיבור בעברית	Whisper-large-v3 + ivrit.ai LoRA	3 GB	HF: ivrit-ai/whisper-large-v3-tuned
זיהוי נקודות יד	MediaPipe Hands (TFLite)	CPU בלבד	pip mediapipe
מעקב גובה צליל ואקורדים	basic-pitch (Spotify) + CREPE	2 GB	pip basic-pitch
מנוע תורת המוזיקה	music21 (קוד שלכם מעל)	CPU בלבד	pip music21 + ~600 שורות
חשיבה פדגוגית	Llama-3.3-8B-Instruct + LoRA	8 GB	מקומי דרך ollama או llama.cpp
סינתזת דיבור בעברית	Coqui XTTS-v2 (כיוונון עברית)	4 GB	מקומי; שיבוט קול אופציונלי

סך VRAM אם הכל טעון בו־זמנית: בערך 17 GB ב־int8 — בנוחות בתוך RTX 4090 בודד (24 GB) או אפילו 4080 Super (16 GB) אם מחליפים מודלים. על 4070 (12 GB) תריצו את ה־LLM ב־int4 ותפרקו את Whisper בין אמירות.

§ מה אתם בעצם כותבים בעצמכם

זה החלק שספקני מסלול ב' מזלזלים בו. כן, המשקלים הכבדים הגיעו ממקום אחר. אבל המערכת היא עדיין *שלכם*, ואלה החלקים שאתם מבלים עליהם זמן:

1. **ממפה הסולם.** קחו את 21 נקודות הציון של היד מ־MediaPipe, כיילו אל מול תמונת סולם חד־פעמית, וייצרו `{string: 1-6, fret: 1-19, finger: index/middle/ring/pinky}` לכל פריים. כ־400 שורות פייתון.
2. **גשר תורת המוזיקה.** שלבו את פלט האקורד/גובה הצליל מ־basic-pitch עם מפת הסולם כדי לייצר אירוע JSON מובנה *«התלמיד מנסה לנגן Em בסריג השביעי כשהזרת על המיתר הלא נכון»*. כ־600 שורות עם music21.
3. **תבנית הפרומפט הפדגוגי.** הפכו את אירוע ה־JSON הזה לשאילתה בעברית עם פרומפט מערכת ל־LLM: *«אתה מורה גיטרה סבלני. התלמיד ניסה כעת <אירוע>. תנוחת היד שלו הייתה <קואורדינטות>. השב בעברית ידידותית, שני משפטים, עם טיפ ספציפי אחד.»* כ־150 שורות פלוס YAML של סגנונות הוראה.
4. **לולאת התזמור.** וידאו ב־20 פריימים בשנייה, אודיו במקטעים של 100 מילישנייה, debounce, ניתוב, רינדור שכבת־על, תור TTS. כ־800 שורות asyncio.
5. **כיוונון ה־LoRA.** כוונו את Llama-3.3-8B על 5,000 דיאלוגים מורה־תלמיד בעברית מאוצרים ידנית מתוך הקלטות השיעורים שלכם. שש שעות על 4090 בעזרת Unsloth, 200 MB של משקלי LoRA בסוף. זה מה שגורם למודל להישמע כמו מורה גיטרה דובר עברית ולא כמו צ'אטבוט גנרי.

סך כל הקוד שאתם כותבים: בערך 2,000 שורות, פלוס מאגר כיוונון בן 5,000 דוגמאות שבילתם עליו שני סופי שבוע בהכנה. התוצאה היא שלכם, רצה לחלוטין על החומרה שלכם, לעולם לא שולחת בית לאף אחד, ותחרותית עם כל מה שמעבדת חזית הייתה יכולה לשלח באותה עלות.

§ ערכות כלים שכדאי להכיר ב־2026

Unsloth

הספרייה המהירה ביותר לכיוונון LoRA / QLoRA ב־GPU צרכניים ב־2026. כפי 2 מהירה יותר מ־peft של Hugging Face עבור אותו מתכון, וכוללת ברירות מחדל טובות ל־Llama, Qwen, Gemma, Mistral.

Axolotl

מסגרת כיוונון כבדה יותר, מונעת YAML, יותר ניתנת להגדרה מ־Unsloth. השתמשו בה כשתגדלו מעבר לברירות המחדל של Unsloth — בדרך כלל בערך כשתתחילו כיוונונים בדיוק מלא על יותר מ־GPU אחד.

llama.cpp / Ollama

שתי הדרכים שבהן רוב האנשים בפועל מריצים LLM פתוחי־משקלים מקומית ב־2026. Ollama הוא עטיפה ידידותית; llama.cpp הוא המנוע. שניהם מריצים Llama-3.3, Qwen-3, Gemma-3 בהתקנה של שורה אחת.

vLLM

שרת הסקה גבוה־תפוקה לכשמשתמש אחד הופך לעשרה. לא להוביסט יחיד; שימושי כשמורה הגיטרה מקבל עשרים תלמידים בו־זמנית.

Hugging Face transformers + peft

עדיין השפה הנפוצה לכל טעינת מודל. תיגעו בזה לא משנה באיזו ערכת כלים גבוהה יותר תשתמשו.

רגיל asyncio או LangGraph

לתזמור. בכנות, asyncio רגיל פלוס קומץ dataclasses מוקלדים מנצח את רוב «מסגרות הסוכן» לפרויקט בגודל הזה. הושיטו יד ל־LangGraph רק כשמכונת המצבים יש בה יותר מעשרה צמתים.

§ לוח הזמנים הכן

בעבודה לבד, ערבים וסופי שבוע, על קופסת שכבה 2, בבניית הפרויקט הזה בלבד:

10	~₪200	2,000	0
סופי שבוע להדגמת v1 עובדת	סך הוצאה, רובה חשמל	שורות קוד משלכם	בייטים שנשלחו לענן כלשהו

זה אינו ניסוי מחשבתי. זה מה שבונים בפועל משלחים ב־2026. עבודת מעבדות החזית הביאה לדמוקרטיזציה כל כך גדולה, שצוואר הבקבוק זז כלפי מעלה — *מהאם המודל יכול להתקיים להאם תוכלו למצוא שימוש שווה לעמל*. מורה הגיטרה הוא שימוש כזה.

חלק שני

ב

החומרה שתרכשו בפועל

> *"רוב ההובאיסטים מעריכים את הניצולת שלהם ביתר של פי 5–10. עקבו אחר דקות ה־GPU בפועל למשך חודש לפני שמחויבים לשכבה 3."*

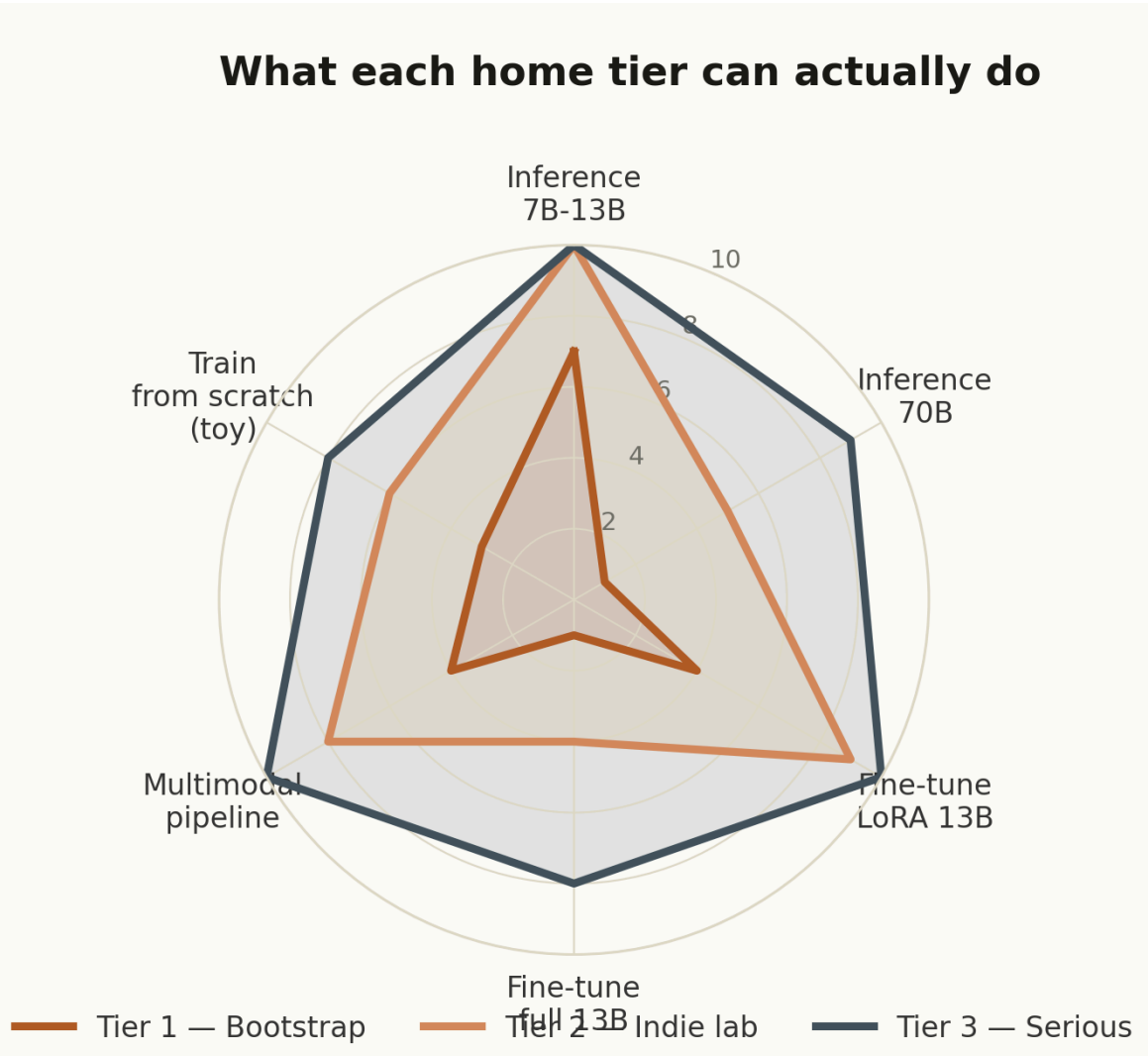
מתוך פרק 6

פרק 5

שלוש שכבות מעבדה ביתית, עם מחירים אמיתיים בישראל

בוטסטראפ, מעבדה עצמאית, רציני. מה כל אחת יכולה לעשות, מה כל אחת עולה הנחתה בישראל, מה כל אחת לא יכולה.

כל המחירים הם של 2026, בשקלים, הנחתו בישראל — אחרי מע"מ 17%, מכס, ושילוח. נדגמו מ־KSP, Bug, Plonter, Ivory ושוק Amazon IL בסוף אפריל 2026 (ציטוט בשורה אחת מתחת לכל טבלה). לחלקים שיובאו אישית מארה"ב, הוספנו ~12% מעל מחיר ארה"ב, התעריף האמפירי של 2025–2026.



איור 5.1 מה כל שכבה באמת יכולה לעשות, מדורגת 0–10. שימו לב ששכבה 1 מוכשרת בהסקת מודלים קטנים אבל חסרת תועלת לכיוונון רציני; שכבה 2 היא הנקודה המתוקה למפתחים בודדים; שכבה 3 מתחילה להיראות כמו מעבדה קטנה.

§ שכבה 1 — בוטסטראפ (3₪–8 אלף הוצאה נוספת)

כבר יש לכם מחשב שולחני — אולי קופסת גיימינג בת שנתיים. אתם מוסיפים (או יש לכם) GPU עכשווי בן 12–16 GB. זו כל הוצאת ההון לבינה מלאכותית. רוב הקוראים כבר נמצאים כאן בלי להבין.

שכבה 1 — תצורת בוטסטראפ, מחירי 2026 הנחתים בישראל

רכיב	בחירה	מחיר ישראל (₪)	הערות
GPU	RTX 4070 Super 12 GB	2,950–3,400	שכבת ביניים ב־2026
GPU (חלופה)	RTX 4070 Ti Super 16 GB	3,800–4,400	16 GB עושים הבדל אמיתי ל־13B ב־int4
שדרוג זיכרון	64 GB DDR5-5600	700–900	אם הבילד הקיים שלכם היה רק 32 GB
שדרוג NVMe	2 TB Gen 4 NVMe	450–600	למשקלי המודלים; תמלאו אותו
סך הוצאה נוספת		**4,100–5,300**	מניח שהשאר של ה־PC קיים

מקורות: ksp.co.il, bug.co.il, plonter.co.il רישומים שנדגמו ב־28 באפריל 2026. המחירים כוללים מע"מ 17%.

מה שכבה 1 בפועל מריצה

- Llama-3.3-8B ב־int8: חלק (~25 טוקנים/שנייה ב־4070 Super)
- כיוונוני Qwen-3-7B ב־int4: חלק
- Llama-3.3-70B: שכחו מזה. אולי ב־int4 עם פירוק חלקי, איטי בכאב
- כיוונון LoRA של מודלים בדרגת 7B: כן, בריצות לילה של 6–10 שעות
- Whisper-large-v3 + Stable Diffusion 1.5 + LLM במקביל: צפוף אבל אפשרי אם מחליפים
- אימון מקדים מאפס: רק מודלי צעצוע (פחות מ־100M פרמטרים)

◆ **פסק דין כן על שכבה 1**

זה מספיק באמת לפרויקט מורה הגיטרה מקצה לקצה. זה גם מספיק ל־80% מפרויקטי הבינה המלאכותית העצמאיים ב־2026. הסיבה לשדרג אינה היכולת כשלעצמה — היא מהירות איטרציה. כיוונון שלוקח 8 שעות על 4070 Super לוקח 90 דקות על 4090, וזה מצטבר על פני מאה ניסויים.

§ שכבה 2 — מעבדה עצמאית (16₪–32 אלף הכל כלול)

תחנת עבודה ייעודית לבינה מלאכותית, לא גיימינג־שעושה־גם־בינה. החלטת ליבה ב־2026: 4090 מול 5090 — ל־5090 יש 32 GB GDDR7 ופי 1.7 תפוקה, אבל הוא עולה כמעט פי 2 הנחת בישראל.

שכבה 2 — תצורת מעבדה עצמאית, מחירי 2026 הנחתים בישראל

רכיב	בחירה	מחיר ישראל (₪)	הערות
GPU (ראשי)	RTX 4090 24 GB	7,800–9,200	מלאי מ־KSP, Plonter 2026
GPU (חלופה פרימיום)	RTX 5090 32 GB	12,500–15,000	זמינות בישראל התייצבה ברבעון 1/2026
מעבד	AMD Ryzen 9 9900X (12c)	1,650–1,900	או Intel Core Ultra 9 285K
לוח אם	X870 (PCIe 5.0)	1,150–1,500	שטח לכרטיס שני אחר כך
זיכרון	128 GB DDR5-5600	1,800–2,200	חובה לכל מיזוג מודלים או עבודה בדיוק מלא
NVMe	2× 4 TB Gen 4 NVMe	1,400–1,800	אחד למערכת+מודלים, אחד למאגרים
ספק כוח	1000 W Platinum	650–850	שטח לכרטיס שני
מארז + קירור	Fractal Torrent / AIO 360 mm	1,400–1,700	זרימת אוויר חשובה יותר מאסתטיקה
סך (בילד 4090)		**15,850–19,150**	טווח שמרני
סך (בילד 5090)		**20,550–24,950**	טווח פרימיום

מקורות: ksp.co.il, plonter.co.il, bug.co.il, ivory.co.il רישומים שנדגמו ב־28 באפריל 2026; הוצלבו עם סקירת השקה TechSpot 2025 והיסטוריית המחירים של zap.co.il. כולל מע"מ 17%.

מה שכבה 2 בפועל מריצה

- כל מודל 7–13B ב־fp16, חלק
- Llama-3.3-70B ב־int4: ~12–18 טוק/ש' על 4090, ~25 על 5090. שמיש.
- כיוונוני QLoRA של 13B בפחות משעתיים
- כיוונון LoRA בדיוק מלא של 7B בפחות משעה
- צינורות רב־מודאליים אמיתיים: Whisper + LLM + TTS + ראייה כולם טעונים במקביל עם מקום בנוח
- אימון מקדים: עדיין רק צעצועים, אבל גדולים יותר (300–500M פרמטרים)

◆ **פסק דין כן על שכבה 2**

זוהי השכבה הנכונה ל«אני לוקח את הבינה המלאכותית ברצינות כמלאכה ורוצה מכונה אחת שלא תיצור צוואר בקבוק לשלוש השנים הבאות.» בילד ה־4090 הוא הבחירה הרציונלית; ה־5090 הוא לאנשים שזמנם שווה את המהירות השולית. אף אחד מהם לא יאמן מראש משהו באיכות חזית, ושניהם יריצו כל LLM פתוח־משקלים מלבד דרגת ה־405B.

§ שכבה 3 — רצינית (70₪–200 אלף)

מעבר לטריטוריית ההובי. השכבה שבה תאמנו מראש מודלים קטנים משמעותיים, תכוונו 70B ביום, תריצו צינורות רב־מודאליים ללקוחות משלמים. שני מסלולים שונים באמת: שני GPU צרכניים (זול־ככה־ככה, קירור מסובך, ללא NVLink או מקבילות מודלים קשה יותר), או כרטיס תחנת עבודה / מרכז נתונים יחיד (יקר, נכנס נקי, נתמך היטב על ידי כל ערכות הכלים).

שכבה 3 — תצורות רציניות, מחירי 2026 הנחתים בישרא

רכיב	בילד א': כפול 4090	בילד ב': RTX 6000 ADA / A6000	בילד ג': H100 PCIE (משומש)
GPU	2× RTX 4090 (~₪16k)	1× RTX 6000 Ada 48 GB (~₪28k)	1× H100 80 GB משומש (~₪95k)
סך VRAM	48 GB (2×24)	48 GB	80 GB
NVLink	אין (PCIe בלבד)	אין רלוונטיות	אין רלוונטיות
לוח תחנת עבודה	~₪7,000	~₪7,000	~₪9,000
זיכרון 256 GB ECC DDR5	~₪7,500	~₪7,500	~₪7,500
סך 8 TB) NVMe)	~₪3,000	~₪3,000	~₪3,000
ספק כוח 1600 W	~₪1,400	~₪1,200	~₪1,400
מארז שרת + קירור	~₪3,500	~₪2,500	~₪4,000
אומדן סך	**~₪38–45 אלף**	**~₪49–55 אלף**	**~₪120–135 אלף**

מחירי בילד א' ובילד ב': ksp.co.il, plonter.co.il ואינטגרטורים ישראליים שנדגמו 28–29 באפריל 2026. בילד ג': H100 80 GB PCIe אינו במלאי בישראל ללקוחות קצה; הנתון מייצג ייבוא אישי ממוכרי שרתים אמריקאים (כגון ServerMonkey, Bargain Hardware) כולל תוספת ~12% להנחתה בישראל. *[טווח 2026, מבוסס על מגמות 2025–2026 — שוק ה־H100 המשומש תנודתי; אמתו תוך 30 יום מהקנייה.]*

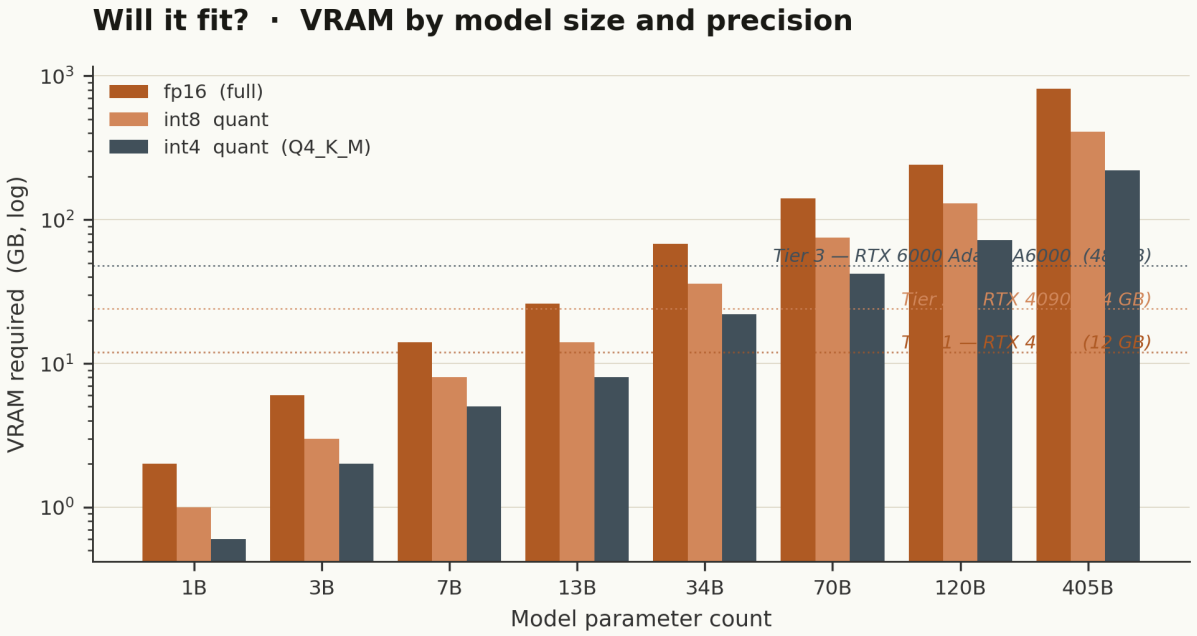
מה שכבה 3 בפועל מריצה

- Llama-3.3-70B ב־fp16: נוח (בילד ב'/ג'); כואב אך אפשרי (בילד א')
- כיוונון בדיוק מלא של 70B ב־1–3 ימים

- אימון מקדים של מודלים בני 1–3 מיליארד פרמטרים מאפס ב־1–4 שבועות — זרימת מחקר אמיתית
- צינורות ייצור רב־מודאליים ל־5–50 משתמשים במקביל
- החזקת כל ה־DeepSeek-V3 MoE ב־int4 על שני 4090 עם פריקה (בילד א')

! הריאליות הלא־סקסית של שכבה 3

שאיבת חשמל מתמשכת של 1.6 קוט"ש לשעות במהלך כיוונונים. ייתכן שהמעגל החשמלי בבית שלכם זקוק לתשומת לב — מעגל ישראלי בודד של 16 אמפר מטפל ב־3.5~ קוט"ש סך הכל, והמזגן + תחנת העבודה + המסך יכולים להרוות אותו. חום: קופסת שכבה 3 בחדר סגור מעלה את הטמפרטורה ב־4–6 מעלות בקיץ. רעש: מעל 50 dB תחת עומס — לא מכונת שולחן, מכונת ארון. שום דבר מזה אינו עוצר־עסקה, אבל תכננו לזה לפני שתוציאו ₪50 אלף.



איור 5.2 הגורם המכריע ל«יתאים?». כל גודל מודל בעל שלוש עמודות (fp16, int8, int4). קווים אופקיים מסמנים את תקרות ה־VRAM של שלוש השכבות. 70B ב־int4 זה בקושי 24 GB — לכן ה־4090 הוא הנקודה המתוקה. מודל 405B מחוץ לתחום של כל בילד GPU בודד ביתי.

פרק 6

הענן, בכנות

שישה ספקים, תעריפים אמיתיים לשעה, נקודת איזון בשעות לחודש לכל שכבה ביתית. החישוב שמכריע «לקנות או לשכור.»

דה אינטואיציה ש«ענן יקר» או «לחברות גדולות» שגויה משני הצדדים ב־2026. למשתמשים בניצולת נמוכה, הענן זול דרמטית מבעלות. למשתמשים בניצולת גבוהה, בעלות משתלמת מהר. השאלה היא איפה נקודת המעבר עבורכם.

§ תפריט השכרת ה־GPU של 2026

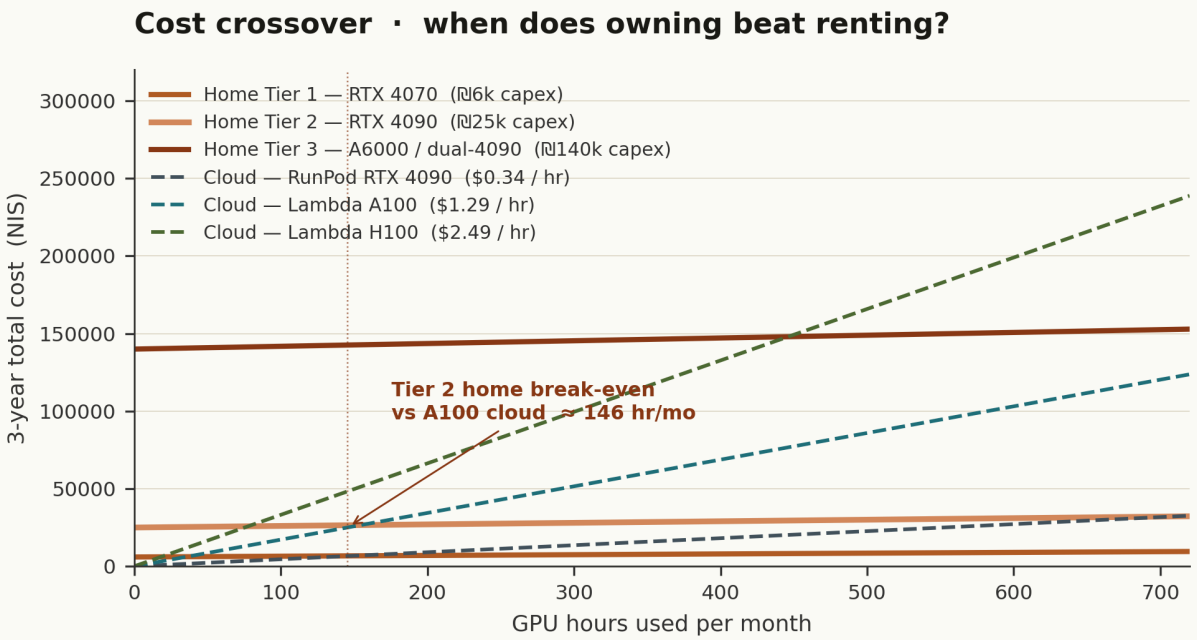
תעריפי GPU בענן, נדגמו בסוף אפריל 2026

ספק	GPU	VRAM	$ / שעה	₪ / שעה	הערות
RunPod (community)	RTX 4090	24 GB	$0.34	₪1.26	הזול ביותר בשכבה זו; סגנון spot, יכול להיקטע
RunPod (secure)	RTX 4090	24 GB	$0.69	₪2.55	מופע יציב
RunPod	A100 80 GB	80 GB	$1.64	₪6.07	סוס עבודה לכיוונוני 70B
Lambda Labs	A100 80 GB	80 GB	$1.29	₪4.77	לפי דרישה; A100 הזול האמין
Lambda Labs	H100 80 GB	80 GB	$2.49	₪9.21	למתי שצריך באמת מהר
Lambda Labs	8×H100 cluster	640 GB	$22.40	₪82.88	רב־צמתים; שעתי
Vast.ai	RTX 4090 (peer)	24 GB	$0.28–0.50	₪1.04–1.85	משתנה; בדקו דירוג מוכר
Together AI	Llama-3.3-70B	n/a	$0.88 / מיליון טוק	₪3.26 / מיליון טוק	לפי טוקן, לא לפי שעה
Anthropic API	Claude Opus 4.7	n/a	$15 in / $75 out	₪55 / ₪278	חזית; אין לזה מקבילה ב־self-host
GCP Vertex AI	A100 40 GB	40 GB	$3.67	₪13.58	פרימיום היפרסקיילר
AWS SageMaker	A100 40 GB (p4d)	40 GB	$4.10	₪15.17	פרימיום היפרסקיילר

מקורות: דפי תמחור של runpod.io, lambdalabs.com, vast.ai, together.ai, anthropic.com, cloud.google.com, aws.amazon.com, נגישו 28 באפריל 2026. המרה ב־1 USD = 3.7 NIS. תעריפי היפרסקיילרים אינם כוללים תעבורה יוצאת, אחסון ותקורת תזמור.

§ נקודת איזון — הגרף שמכריע

השאלה השימושית ביותר היא: *בכמה שעות-GPU לחודש שכבה ביתית מחזירה את עצמה לעומת השכרת GPU מקביל בענן?* להלן התשובה להשוואה הנפוצה ביותר — קופסת שכבה 2 ביתית (25₪ אלף הוצאת הון) לעומת Lambda A100 ($1.29/שעה).



איור 6.1 עלות מצטברת ל־3 שנים כפונקציה של שעות GPU בחודש. הקו המקווקו מסמן את נקודת האיזון של שכבה 2 ביתית מול ענן A100. מתחת ל־~155 שעות/חודש (כ־5 שעות/יום), הענן זול יותר. מעליו, הבעלות מנצחת — והפער מתרחב מהר.

§ טבלת נקודת האיזון להדפיס ולשמור

שעות GPU לחודש לאיזון — אם תשתמשו יותר, הבעלות זולה יותר (אופק 3 שנים)

שכבה ביתית	מול RUNPOD 4090 ($0.34/ש)	מול LAMBDA A100 ($1.29/ש)	מול LAMBDA H100 ($2.49/ש)
שכבה 1 (6₪ אף הכל)	133 שעות/חודש	36 שעות/חודש	19 שעות/חודש
שכבה 2 (25₪ אף הכל)	555 שעות/חודש	~155 שעות/חודש	79 שעות/חודש
שכבה 3 (140₪ אף הכל)	3,100 שעות/חודש	~840 שעות/חודש	440 שעות/חודש

אופק 3 שנים, חשמל בתעריף מגורים ישראלי 0.62₪/קוט"ש, GPU מושך TDP ממוצע מדורג תחת עומס. תעריפי ענן נגעלו ברמות אפריל 2026 (סביר שירדו, מצמצמים את היתרון הביתי). 24×30 = 720 שעות/חודש זו התקרה המוחלטת.

איך לקרוא את זה

אם תשתמשו ב־GPU יותר מ־5 שעות ביום בממוצע, בילד ביתי שכבה 2 משתלם מול Lambda A100 תוך שלוש שנים. אם תשתמשו פחות משעה ביום, אל תקנו — שכרו. מספרי שכבה 3 נראים מאיימים, ובצדק: H100 בודד בארון בבית מנצח את Lambda H100 רק בכ־14 שעות שימוש מתמשך ביום, מה שזה בערך מה שסטארטאפ קטן שמריץ מוצר משלם צריך.

> ◆ **העובדה התפעולית הכי שימושית**
>
> רוב ההובייסטים מעריכים יתר על המידה את הניצולת שלהם פי 5–10. הם מדמיינים את עצמם «מאמנים את כל סוף השבוע» אבל בפועל משתמשים ב־4 GPU שעות בשבת ואז כלום במשך שבועיים. **עקבו אחרי דקות ה־GPU האמיתיות שלכם במשך חודש לפני שאתם מתחייבים לשכבה 3.** `nvidia-smi --query-gpu=utilization.gpu --format=csv -l 60 >> util.log` — השורה האחת הזו, שרצה ברציפות, תאמר לכם אם אתם משתמש של 30 שעות בחודש או 300 שעות בחודש.

§ ההיברידי שכמעט כולם מסיימים בו

מעט בונים אמיתיים הם ענן־בלבד או בית־בלבד. הדפוס הכן של 2026 הוא:

- **הסקה וניסויים קטנים בבית** (שכבה 1 או 2). תמיד פעיל, עלות שולית נמוכה, הנתונים שלכם נשארים מקומיים.
- **ריצות אימון גדולות בענן** (RunPod, Lambda). הפעילו אשכול 8×A100 לריצה של 12 שעות, שלמו 200$, סגרו. הקופסה הביתית שלכם הייתה לוקחת שלושה שבועות לאותה עבודה.
- **קריאות API חזית לדברים שרק החזית יכולה** (Anthropic, OpenAI). כשצריך חשיבה בדרגת GPT למשימה חד־פעמית, לשלם 15$ על מיליון טוקני קלט עוקף חודש של ניסיון להתאים את זה מקומית.

ההיברידי אינו פשרה — זוהי הקצאה אופטימלית של הכסף והזמן שלכם. הטעות היא מחויבות דתית לצד אחד.

פרק 7

העלויות הסמויות של בית

חשמל, פחת, הזמן שלכם, מצבי הכשל שאף אחד לא מזהיר אותכם עליהם. התמונה המלאה של שלוש שנים.

GPU ביתי זול מענן» נכון בדרכים צרות ומטעה בדרכים רחבות יותר. הוצאת ההון היא הכותרת; שלוש עלויות אחרות נצמדות אליכם לשלוש שנים, ורוב ההשוואות המקוונות מתעלמות מהן.

§ חשמל, בישראל

תעריף מגורים של חברת החשמל, ינואר 2026: **₪0.62/קוט"ש** רגיל, ~₪0.42 בלילה. קופסת שכבה 2 בשימוש כבד מושכת ~600 W תחת עומס (450 W GPU + 150 W שאר), במנוחה ~80 W. אימון 4 שעות/יום בממוצע:

שכבה 2, 4 שעות/יום עומס	שכבה 3, 8 שעות/יום עומס	שכבה 1, 2 שעות/יום עומס
~₪900/שנה	**~₪1,800/שנה**	**~₪260/שנה**

אף אחד מהמספרים האלה לא ישנה את ההחלטה שלכם בפני עצמו. אבל לאורך שלוש שנים, חשמל שכבה 3 הוא ₪5,500 — בערך עלות של עוד NVMe או חודשיים השכרת ענן. וזה משולם בחשבונות חודשיים שאי אפשר לבטל.

§ פחת, בכנות

שווי השוק של GPU בשנה השלישית לבעלות, מבוסס על דפוסי שוק יד שנייה 2024–2026:

שווי שאריתי ב־36 חודשים — כרטיסי NVIDIA אחרונים

כרטיס	מחיר השקה בישראל	מחיר יד שנייה בשנה 3	פחת אפקטיבי
RTX 3090 (2020)	~₪7,000 השקה	~₪2,800 ב־2023	60%
RTX 4090 (2022)	~₪10,500 השקה	~₪5,500 ב־2025	48%
RTX 4070 (2023)	~₪3,400 השקה	~₪2,000 ב־2026	41%

מקורות: ארכיוני היסטוריית מחירים של zap.co.il, רישומי yad2.co.il / facebook-marketplace IL שנדגמו על פני 2023–2026. מגמות בקירוב.

הפרשנות הכנה: 4090 שאתם קונים היום ב־₪8 אלף שווה כ־₪4 אלף ב־2029. אותם ₪4 אלף הפרש זו עלות ה־GPU האמיתית שלכם לאורך שלוש שנים — לא המדבקה המלאה. השוואות ענן שמשתמשות במדבקה המלאה כ«הוצאת הון ביתית» מנפחות את עלות הבית.

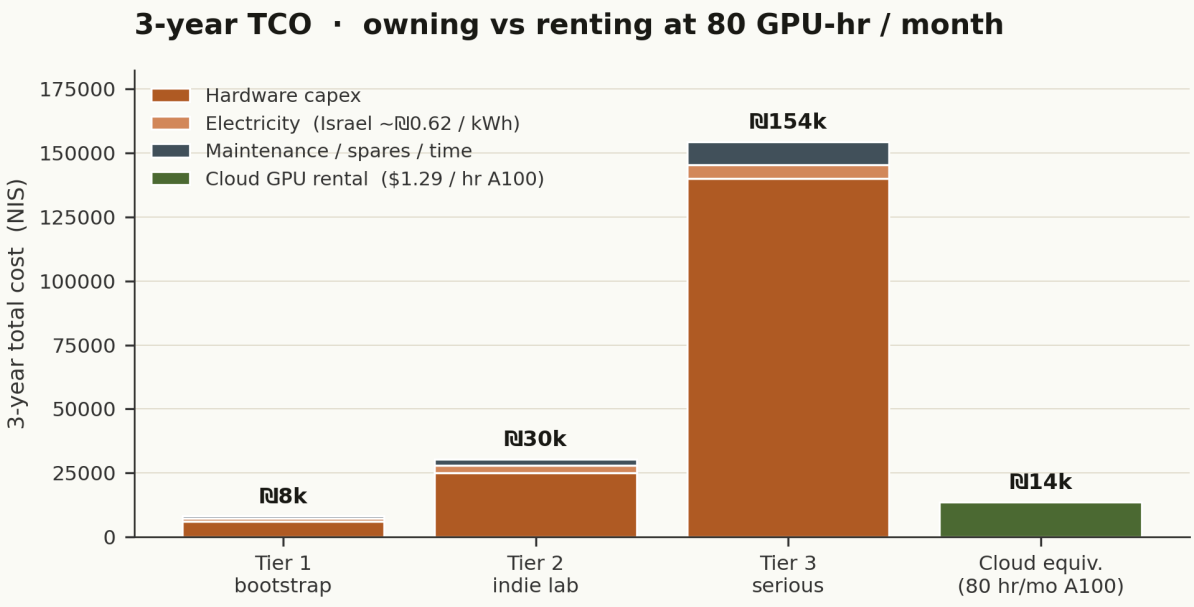
§ הזמן שלכם — העלות שכולם שוכחים

בעלות על תחנת עבודה עולה זמן. ספציפית:

- **בנייה ראשונית:** 4–8 שעות הרכבה + מערכת הפעלה + מנהלי התקן, פלוס סוף שבוע של נסיעת ערכת הכלים / CUDA / Python כיוונון
- **תחזוקה שוטפת:** עדכוני מנהלי התקן ששוברים דברים (~3 אירועים/שנה, ~2 שעות כל אחד), ניקוי תרמי (1 שעה / 6 חודשים), כשל רכיב מזדמן (~1 אירוע ל־18 חודש, חצי יום אבחון והחלפה)
- **ניהול חשמל:** החלטה אם להשאיר דלוק בלילה, הגדרת שינה, ההכרה הבלתי נמנעת ב־3 לפנות בוקר ששכחתם להפעיל ריצת אימון

אם אתם מעריכים את זמנכם ב־₪150/שעה (נתון שמרני למפתח תוכנה ישראלי ב־2026), עלות הזמן של בעלות על שכבה 2 לאורך שלוש שנים היא בערך **₪3,500** — בערך כמו עוד GPU. משתמשי ענן לא משלמים את זה; הם מחליפים אותו בסט שונה של חיכוכים (המתנה למופעים, פירוק משאבים ששכחתם, חשבונות תעבורה־יוצאת מפתיעים).

§ גרף ה־TCO המלא של 3 שנים



איור 7.1 סך עלות בעלות ל־3 שנים, כל הרכיבים כלולים, לתרחיש שימוש מתון (80 שעות GPU/חודש, ~2.7 שעות/יום ממוצע). בניצולת זו, שכבה 1 ביתית היא המנצחת הברורה; שכבה 3 עודפת; הענן תחרותי רק כי התעריף שמונח הוא A100 הזול ($1.29/ש'). החליפו את קו הענן להיפרסקיילר ($3.67/ש' GCP) והעמודה הופכת ל־₪32 אף.

§ מה ציבור הענן ממעיט בו

עכשיו הצד השני. למשתמשי ענן יש עלויות סמויות משלהם שאוונגליסטים של בית רק לעיתים נדירות מזכירים:

חיובי תעבורה יוצאת.
היפרסקיילרים גובים 0.30₪–0.50 לג"ב להזיז נתונים *מחוץ* לענן שלהם. אם אתם מאמנים מודל ורוצים את המשקלים על המחשב הנייד שלכם — זה 15 ג"ב תעבורה לכיוונון 7B, או כ־7₪. זעיר לאירוע; גדול בקנה מידה, והחשבון מגיע חודש לאחר מכן.

זמן הפעלה.
הפעלה קרה של מופע RunPod חדש לוקחת 3–6 דקות; AWS p4d טרי, 8–12 דקות כולל משיכת AMI. לעשרה ניסויים ביום, זה שעה מהחיים שלכם בהסתכלות על «Provisioning...».

משאבים שנשכחו.
לכל משתמש ענן יש לפחות סיפור מלחמה אחד על מופע GPU שנשכח שרץ סוף שבוע ב־5$/שעה. זה 240$. החלופה הביתית היא חשבון החשמל שלכם מתלונן בעדינות.

נעילה דרך גרביטציית נתונים.
ברגע שיש לכם 200 ג"ב של נתוני אימון, נקודות עצירה של מודל ולוגים יושבים ב־S3, החלפת ספקים כואבת באמת. לא בחרתם בחופשיות ביום 365 גם אם בחרתם בחופשיות ביום 1.

תאימות וריבונות נתונים.
אם הפרויקט שלכם כולל נתונים אישיים של אזרחים ישראלים, חוק הגנת הפרטיות הישראלי המקביל ל־GDPR, או משהו שאי אפשר לאחסן חוקית מחוץ לישראל, חלקים גדולים מתפריט הענן מחוץ לתחום. (ראו פרק 8.)

§ תשובת שלוש השנים

למפתח בודד שעושה את הסגנון של פרויקט מורה הגיטרה — בעיקר הסקה, כיוונון מזדמן, 50–100~ שעות GPU/חודש — הדירוג הכן הוא:

1. **בית שכבה 1** (10₪~ אלף הכל לאורך 3 שנים, חשמל כלול). העלות הכוללת הטובה ביותר. ריבון לחלוטין.
2. **בית שכבה 2** (31₪~ אלף לאורך 3 שנים). מהירות איטרציה טובה יותר; משתלם אם תעשו ניסויי כיוונון רבים.
3. **ענן בלבד** (14₪~ אלף ב־80 שעות/חודש על RunPod 4090). זול באופן מפתיע אם השימוש שלכם נמוך ופרצי; מפסיד בריבונות נתונים.
4. **היברידי** (14₪~ אלף בית שכבה 1 + 3₪~ אלף/שנה פרצי ענן סלקטיביים). התשובה הכנה כמעט לכולם.
5. **בית שכבה 3**. רק אם אתם מריצים עומסי מוצר משלם או עושים מחקר אמיתי בבית.

אם תזכרו דבר אחד: עקבו אחרי ניצולת ה־GPU האמיתית שלכם במשך חודש לפני שתוציאו משהו מעל שכבה 1. נקודת הנתונים האחת הזו שווה יותר מכל בלוג בנצ'מרק שתקראו.

חלק שלישי

ג

מה אתם באמת מחזיקים בסוף

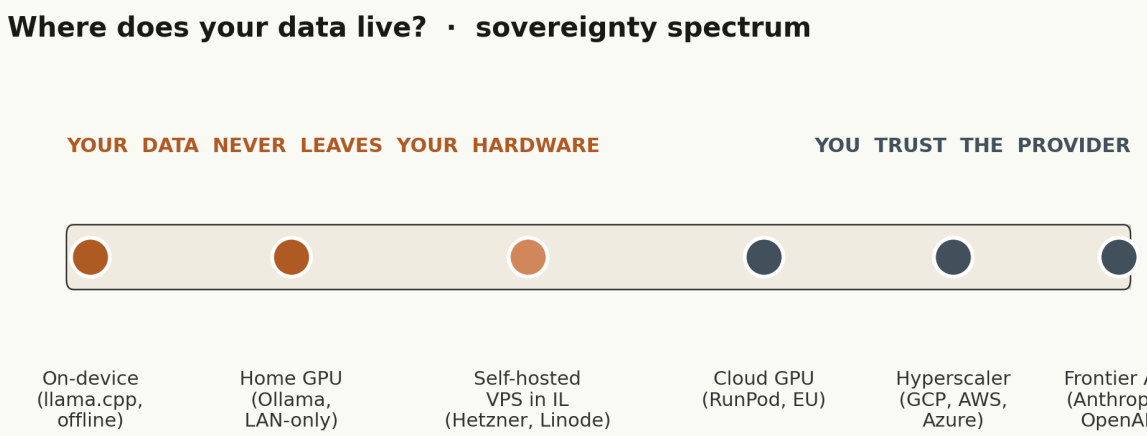
> *"העלות האמיתית של הענן איננה החשבון — היא מה שהחשבון מונע מכם לעשות."*
>
> מתוך פרק 8

פרק 8

מה באמת המשמעות של «הנתונים שלכם»

ריבונות, פרטיות, נעילה. המסגור הכן — לא «הענן הוא רוע» אלא «הנה מה שכל אופציה עולה לכם במונחי נתונים.»

דרך הזה קיים כי השיח סביב פרטיות בינה מלאכותית הוא לא־רגיל בגרועותו. מחנה אחד אומר «הענן בסדר, אין לכם מה להסתיר.» המחנה השני אומר «כל שימוש בענן הוא קולוניאליזם נתונים.» שניהם מפספסים את השאלה האמיתית, שהיא משעממת ושימושית יותר: *מה ספציפית קורה לנתונים שלכם בכל אופציה, וזה משהו שאכפת לכם ממנו?*



איור 8.1 הספקטרום הכן. כל נקודה מייצגת אופציה אמיתית להרצת עומס בינה מלאכותית ב־2026. תזוזה ימינה (כלפי מימין בגרסה זו: שמאלה במקור) נותנת איכות טובה יותר והקמה קלה יותר; תזוזה לצד הריבוני נותנת יותר שליטה היכן הנתונים יושבים.

§ שש האופציות הכנות

1. על־מכשיר, לא־מקוון לחלוטין (למשל llama.cpp במחשב נייד ללא רשת)
הנתונים שלכם לא עוזבים את המכונה שעליה הקלדתם. המודל לא רואה את האינטרנט במהלך הסקה. גם אם אתם טועים בהגדרת אבטחה, אין משטח התקפה. **תמורה:** מוגבלים למה שמתאים לזיכרון של המחשב הנייד, ואחראים לכל עדכון.

2. GPU ביתי, LAN בלבד (Ollama רץ על תחנת העבודה שלכם, נגיש ממכשיריכם האחרים דרך WiFi)
אותה ריבונות כמו אופציה 1, עם חומרה טובה יותר. המודל מקומי; הפרומפטים עוברים את ה־WiFi הביתי שלכם אבל אף פעם לא את המודם. **תמורה:** עליכם לתחזק את הקופסה.

3. אחסון עצמי על VPS בישראל (ל־Hetzner Falkenstein יש אופציות בתחום שיפוט ישראלי; Linode POP בתל אביב)

הנתונים שלכם על מכונה ששכרתם בלעדית, בתחום שיפוט שאתם סומכים עליו, על חומרת ספק. זמן פעילות טוב יותר מבית; פחות ריבונות (המארח יכול עקרונית לקרוא דיסק אם נדרש בצו). **תמורה:** 200₪–800/חודש ל־VPS GPU שמיש, לעיתים פחות GPU לשקל מבית.

4. השכרת GPU בענן (RunPod, Lambda, Vast)

אתם מפעילים קופסת לינוקס עם GPU, עושים את העבודה שלכם, סוגרים. הנתונים שלכם יושבים על תשתית משותפת; מדיניות הספק אומרת שלא יסתכלו אבל הם טכנית יכולים. משקלי המודל שאתם מעלים נמצאים על הדיסק של מישהו אחר למשך הזמן. **תמורה:** עלות מצוינת לשעת GPU, חשיפת נתונים אמיתית אבל מוגבלת.

5. בינה מלאכותית מנוהלת של היפרסקיילר (GCP Vertex, AWS SageMaker, Azure ML)

אותו דבר כמו אופציה 4 פלוס שילוב עמוק עם שאר המערכת האקולוגית של ההיפרסקיילר ו־SLA חזק יותר — וגם נעילה חזקה יותר. הנתונים שלכם נשלטים על ידי תנאי ההיפרסקיילר, שהם נרחבים ומשתנים. **תמורה:** פי 3–5 יותר יקר לשעת GPU, כלים טובים יותר בהרבה, נעילת ספק גדולה.

6. API חזית (Anthropic, OpenAI)

אתם שולחים את הפרומפט שלכם, מקבלים השלמה. מעבדת החזית רואה כל פרומפט. יש להן מדיניות כתובה (Anthropic ו־OpenAI שתיהן כברירת מחדל לא מאמנות על נתוני API, אבל מדיניות היא מדיניות, לא חוק). **תמורה:** אתם מקבלים את המודלים הטובים בעולם, אתם מקבלים חשבון לטוקן, ואתם סומכים על חברה אמריקאית עם הפרומפטים שלכם.

§ איפה הנתונים של עידו חיים, ספציפית

המדריך הזה הוא לבונה שאכפת לו מדברים ספציפיים — תמונות משפחה של אחותו המנוחה, הקלטות קול של בני משפחה, תוכן עברי שנוצר על ידי אנשים שיש לו את הסכמתם אבל מעולם לא הסכימו להיות במאגר האימון של OpenAI, טקסטים דתיים בהערות אישיות. לנתונים בעלי אופי כזה, התשובה הנכונה היא לעיתים נדירות «האופציה הזולה ביותר.» היא קרובה יותר ל«האופציה שבה התשובה ל*'מי עוד יכול עקרונית לראות את זה?'* היא הקצרה ביותר.»

> ◆ **היוריסטיקת הריבונות הכנה**
>
> אם הנתונים בכונן שלכם הם משהו שלא הייתם רוצים לקרוא בדלף מאגר ארגוני בעוד חמש שנים, עשו את העבודה מקומית. אם הנתונים הם משהו שהייתם בשמחה מפרסמים בבלוג ציבורי מחר, אופציית הענן הזולה בסדר. רוב הפרויקטים יש להם *שני* סוגי נתונים — עשו את החלקים הרגישים מקומית, השאר בענן.

§ דפוס הריבונות ההיברידי

למורה הגיטרה, הפיצול הטבעי הוא:

- **מקומי בלבד:** הקלטות קול הסטודנט, וידאו, וכל ביומטריה ממצלמת הרשת. אלה אף פעם לא עוזבות את המכונה הביתית. Whisper ו־MediaPipe שניהם רצים מקומית.
- **מקומי בלבד:** מאגר הכיוונון של דיאלוגי מורה־תלמיד שלכם. זה הקניין הרוחני שלכם; העלאה לענן תהיה טיפשית.

- **ענן בסדר:** הורדת מודל הבסיס Llama-3.3-8B הראשונית (זה ציבורי, אתם רק מורידים). עדכוני ספריית Unsloth. חבילות פייתון.
- **ענן בסדר:** מטריקות מצטברות ואנונימיות («השיעור הזה לקח N שניות, M טעויות אקורד התגלו»). שימושי לשיפור מוצר, לא מזהה אישית.
- **פרץ ענן, לפעמים:** אם אי פעם תעשו כיוונון בדיוק מלא בקנה מידה 70B, תשכרו אשכול A100 ליום. את המאגר שאתם מעלים יש לחטא תחילה. המשקלים המאומנים שאתם מורידים אז שלכם לנצח — הענן רק החזיק אותם בתעבורה.

§ איך נראית נעילה בפועל

העלות האמיתית של הענן אינה החשבון — היא מה שהחשבון מונע מכם לעשות.

אם תבנו את מורה הגיטרה כולו על ה־API של Anthropic, הפרויקט מת ביום ש־Anthropic מוציאה משימוש את המודל שעליו בניתם, או משנה את התמחור, או מגבילה את התרחיש. (כל שלושת אלה קרו למישהו שאתם מכירים ב־24 החודשים האחרונים.) אם תבנו אותו על Llama-3.3-8B מקומי + 200 מ"ב LoRA שאימנתם בעצמכם, הפרויקט שלכם חי כל עוד הכונן הקשיח שלכם חי. את ההבדל הזה קשה להמיר למחיר, אבל הוא אמיתי וצריך לשקול אותו בהחלטה.

הפוך, אם תסרבו לענן באדיקות דתית, תבלו שלושה חודשים בבניית TTS עברי מאפס כאשר כיוונון XTTS-v2 מריצת RunPod של יום אחד היה עושה את זה אחר הצהריים. בפנקס הכן יש עלויות בשני הצדדים.

פרק 9

ערכת התחלה מומלצת, למישהו במצבכם

תשובה דעתנית אחת. החומרה, התוכנה, עשרת סופי השבוע הראשונים של עבודה.

צה דחוסה לקורא שקרא את שמונת הפרקים הקודמים ועכשיו רוצה שיגידו לו מה לעשות. הפרק הזה דעתני. יש תשובות אחרות בנות קיום. זאת זו שהיינו נותנים לחבר.

§ אם יש לכם PC גיימינג עכשווי: הישארו שם

הוסיפו GPU בן 16 GB (RTX 4070 Ti Super, ₪3.8–4.4 אלף) ו־64 GB זיכרון אם אין לכם. סך הוצאה 4₪–5 אלף. זה מספיק לכל פרויקט מורה הגיטרה, ומספיק כמעט לכל עבודת בינה מלאכותית עצמאית ב־2026. אל תקנו עוד חומרה לפני שתשתמשו במה שיש לכם.

§ אם אין לכם כלום ואתם רוצים להתחיל נקי: בנו שכבה 2

בילד RTX 4090 בקצה השמרני של הטווח: ~16₪ אלף. הוסיפו חשבון Lambda Labs ($0 להירשם, אתם משלמים רק על מה שמשתמשים) לפרצי H100 מזדמנים על עבודות שלא מתאימות. השילוב הזה מטפל ב־95% ממה שבונה עצמאי רציני יעשה לאורך שלוש השנים הבאות.

§ עשרת סופי השבוע הראשונים — סילבוס

תוכנית לימודים דעתנית של עשרה סופי שבוע

סוף שבוע	מה לעשות	מה יהיה לכם בסוף
1	התקנת Python 3.12, PyTorch 2.6, CUDA 12.6, Ollama. הריצו Llama-3.3-8B מקומית. דברו איתו בעברית.	LLM מקומי עובד. סביבה עובדת.
2	nanoGPT של קרפתי או רשקה פרקים 1–3. אמנו טרנספורמר 10M פרמטרים על תנ"ך.	LLM צעצוע משלכם. האינטואיציה שחשובה.
3	Whisper-large-v3 + ivrit.ai, תמללו פודקאסט עברי ארוך. השיגו דירוג טעויות מילים כן.	מודול תפיסה אמיתי. ביטחון בצינור האודיו.
4	מדריך MediaPipe Hands. כיילו א מול הגיטרה שלכם. הוציאו JSON אצבע־על־סריג.	מודול הראייה של מורה הגיטרה.
5	basic-pitch + music21. זהו אקורדים מהקלטות שלכם.	מודול ניתוח האודיו.
6	חברו 3–5 יחד. הדגמה מקצה לקצה: מצלמת אינטרנט פנימה, אירוע JSON החוצה.	עמוד השדרה של התזמור.
7	אצרו ידנית 200 דיאלוגי מורה־תלמיד בעברית. עצבו לכיוונון.	מאגר האימון.
8	כיוונון Unsloth LoRA של Llama-3.3-8B על המאגר. שני אפוקים, בלילה.	LLM בטעם מורה גיטרה.
9	Coqui XTTS-v2 עם עברית. חברו לפלט ה־LLM.	תגובות עבריות מדוברות.
10	ליטוש, טיפול בשגיאות, חלון Tk GUI פשוט. הראו לתלמיד אמיתי אחד.	גרסה 1, רצה לחלוטין על המכונה שלכם.

§ מה לקרוא לאורך הדרך

- **סבסטיאן רשקה, *Build a Large Language Model From Scratch*** — הטקסט הבודד הטוב ביותר לאינטואיציית מסלול א'. קראו את שלושת הפרקים הראשונים בסוף שבוע 2.
- **סדרת היוטיוב *neural networks: zero to hero* של אנדריי קרפתי** — חינם, עמוקה, עדיין סטנדרט הזהב.
- **קורס ה־NLP של Hugging Face** — הגשר המעשי למסלול ב'. חינם.
- **ה־README של Unsloth ב־GitHub** — מתכון הכיוונון שתשתמשו בו בסוף שבוע 8. מתעדכן תכופות; בדקו גרסה מול ה־Llama שמותקן לכם.
- **הבלוג של ליליאן וונג** (lilianweng.github.io) — לכשתרצו להבין את מה שזה עתה השתמשתם בו.

§ פסקה כנה אחרונה

החיזוי הבודד הגדול ביותר אם תשלחו את מורה הגיטרה אינו איזה GPU קניתם. הוא אם תעשו את סוף שבוע 2 לפני סוף שבוע 8. רוב האנשים מדלגים על סוף שבוע 2 (מסלול א'), הולכים ישר לסוף שבוע 8 (כיוונון LoRA), לא מבינים למה ה־loss שלהם מתפוצץ, מתייאשים. השבוע של עבודת מודל הצעצוע הוא מה שהופך את שאר הסילבוס לבר־ביצוע. אל תדלגו עליו.

מעבר לכך — הפרויקט שלכם. החומרה נקנתה. מעבדות החזית עשו את החלק היקר. עשרה סופי שבוע משבת זו, יוכל להיות לכם משהו שרץ על השולחן שלכם, שלפני עשר שנים היה דורש מעבדת מחקר.

— עידו, תל אביב, 2026

ביבליוגרפיה

קריאה נוספת

סבסטיאן רשקה · *Build a Large Language Model From Scratch*
Manning, 2024 — הספר היחיד הטוב ביותר לבניית אינטואיציה למסלול א.

אנדריי קרפת'י · *Neural Networks · Zero to Hero (YouTube)*
חינם, עמוק, עדיין סטנדרט הזהב. youtube.com/karpathy

Hugging Face · *NLP Course*
הגשר המעשי למסלול ב. huggingface.co/learn/nlp-course

צוות Unsloth · *Unsloth GitHub README*
מתכון הכיוונון בפועל שתשתמשו בו בסוף שבוע 8. github.com/unslothai/unsloth

Lilian Weng · *lilianweng.github.io*
לקריאה כשרוצים להבין מה הרגע השתמשתם בו.

פרויקט ivrit.ai · *ivrit-ai/whisper-large-v3-tuned*
LoRA לזיהוי דיבור בעברית ב־Hugging Face. huggingface.co/ivrit-ai

Anthropic · *תיעוד מודלי Claude*
הפניה לעמודת ה־API של החזית בהשוואה. docs.anthropic.com

קולופון

המהדורה השנייה של *בנו בעצמכם AI 2026* עוצבה והוקלדה בתל אביב, מאי 2026. הגוף מוקלד ב־**Frank Ruhl Libre** מאת מיכל סהר ופונט־פינדר, עיצוב מודרני של כותב המאה ה־19. הכותרות מוקלדות ב־**Heebo** מאת עודד עזר, וקטעי קוד ב־**JetBrains Mono**. המהדורה האנגלית משתמשת ב־EB Garamond ו־Inter.

אמנות העטיפה היא חתך מסוגנן של רשת נוירונים, מרונדר תוכנתית באמצעות matplotlib (ראו `scripts/cover.py`). כל האיורים בטקסט הם matplotlib ב־300 DPI (ראו `scripts/figs_premium.py`). הרכבת העמוד נעשית ב־HTML/CSS גולמי ומרונדרת ל־PDF דרך מנוע ההדפסה של Chromium ללא ראש.

גרסה 2 · מאי 2026 · עידו, עם Claude Opus 4.7. לזכרה האהוב של עלמה ז״ל.