BY MENDY / 1 OCT 2026

Likkutei Sichos OCR: every score

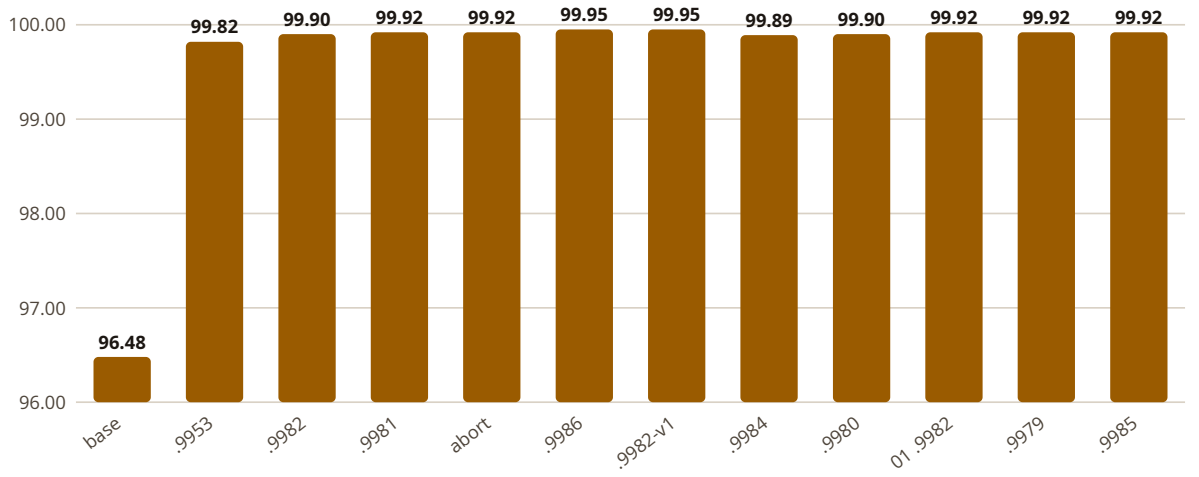
Every accuracy score from the PP-OCRv6 training run on Likkutei Sichos vol 39, with charts.

99.91% letters — focus model, kept	**99.68%** words — 4,500 vol 39 lines
97.7% lines perfect	**179** mistakes on 4,500 lines (was 357)

Every score from the PP-OCRv6 run on Likkutei Sichos, all on vol 39 lines the model never trained on. Updated 2026-10-01.

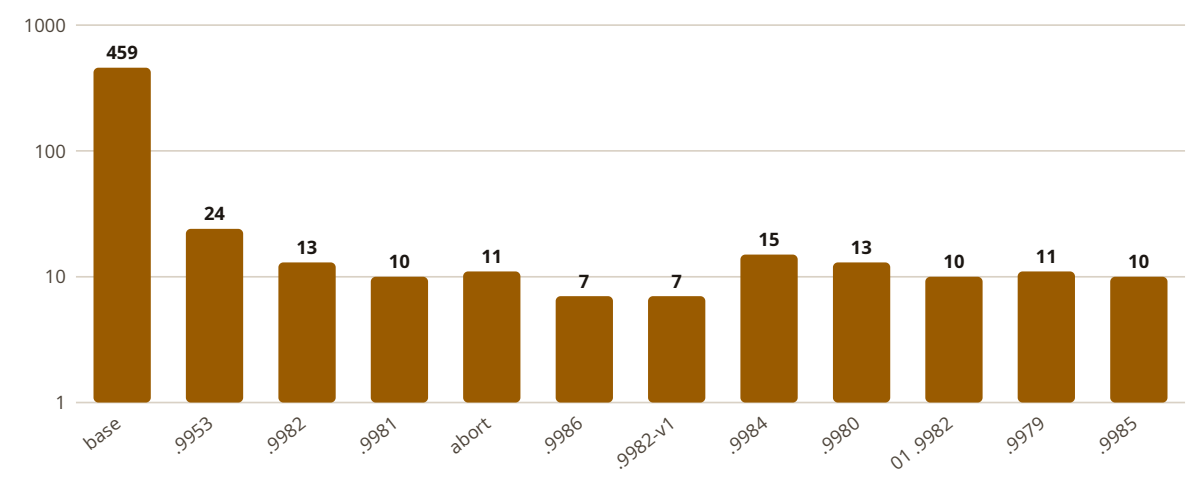
ⓘ **Checked labels** = the 300 lines with every label fixed by the relabeller and 12 more fixed by eye against the scan. These are the honest scores. **Pod labels** = what the pod scored itself against live — a few label errors still in them, so those numbers run a little low.

300 checked lines — accuracy across saves



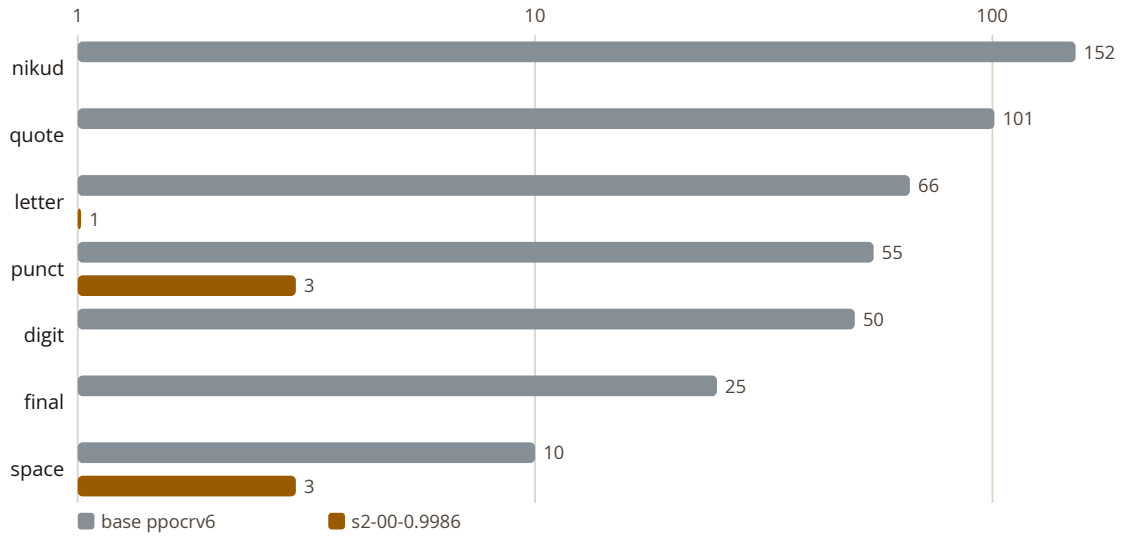
Letters, checked labels. Axis starts at 96% — every save of the run sits between 99.89% and 99.95%; the gaps between saves are 1–6 mistakes in 300 lines, i.e. noise. The 4,500-line final test decides which save is kept.

Mistakes per save (300 lines, log scale)



Base PP-OCRv6 started at 459 mistakes. Best saves: 7.

Mistakes by kind — base vs best save



Log scale. ניקוד and quote marks were the base model's whole game; the trained model is down to spaces, punctuation, and one or two letter look-alikes.

Headline scores (checked labels)

MODEL OR SAVE	SAVED (UTC)	LETTERS	WORDS	LINES PERFECT	BODY	NOTES
base ppocrv6 (start)	before training	96.48%	86.07%	44.3%	96.41%	96.95%
s2-00-0.9953 (first save)	early run	99.82%	99.24%	94.7%	99.8%	99.94%
s2-00-0.9982	08:13	99.9%	99.64%	97.7%	99.89%	100.0%
s2-00-0.9981	08:23	99.92%	99.64%	97.7%	99.93%	99.88%
s2-abort	08:33	99.92%	99.64%	97.3%	99.93%	99.82%
s2-00-0.9986	**08:57**	**99.95%**	**99.68%**	**98.0%**	**99.94%**	**100.0%**
s2-00-0.9982-v1	09:07	99.95%	99.72%	98.0%	99.94%	100.0%
s2-00-0.9984	09:16	99.89%	99.52%	96.7%	99.89%	99.82%
s2-00-0.9980	09:26	99.9%	99.56%	97.0%	99.92%	99.77%
s2-01-0.9982	09:36	99.92%	99.6%	97.0%	99.95%	99.77%
s2-01-0.9979	09:45	99.92%	99.52%	97.3%	99.92%	99.88%

MODEL OR SAVE	SAVED (UTC)	LETTERS	WORDS	LINES PERFECT	BODY	NOTES
s2-00-0.9985	10:14	99.92%	99.68%	98.0%	99.91%	100.0%

Same saves on the pod labels (as reported live)

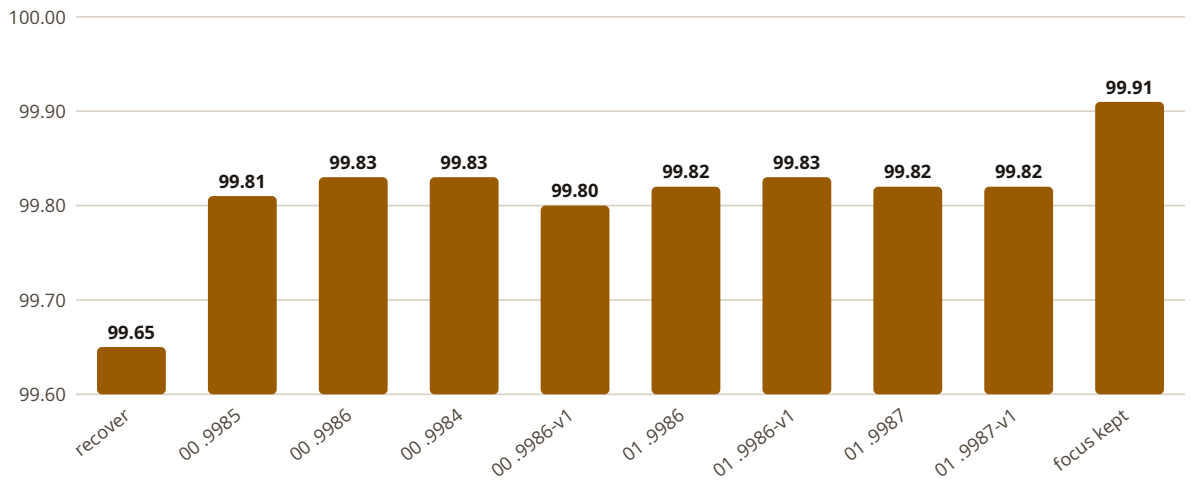
MODEL OR SAVE	LETTERS	WORDS	LINES PERFECT	MISTAKES
base ppocrv6 (start)	97.45%	91.43%	51.0%	332
s2-00-0.9953 (first save)	99.45%	98.84%	91.3%	72
s2-00-0.9982	99.82%	99.6%	97.3%	23
s2-00-0.9981	99.85%	99.56%	97.3%	20
s2-abort	99.85%	99.6%	97.3%	19
s2-00-0.9986	99.86%	99.68%	98.0%	18
s2-00-0.9982-v1	99.85%	99.56%	97.0%	20
s2-00-0.9984	99.83%	99.4%	96.0%	22
s2-00-0.9980	99.83%	99.48%	96.3%	22
s2-01-0.9982	99.85%	99.44%	96.3%	20
s2-01-0.9979	99.83%	99.4%	96.0%	22
s2-00-0.9985	99.85%	99.56%	97.3%	20

Mistakes by kind and top confusions (checked labels)

MODEL OR SAVE	BY KIND	TOP CONFUSIONS
base ppocrv6 (start)	nikud 152, quote 101, letter 66, punct 55, digit 50, final 25, space 10	־ 61, „ " 59, ֽ 48, (,29 ֿ 9 7 ל ,10 ח ת ,11 . , ,27
s2-00-0.9953 (first save)	punct 7, space 5, letter 3, nikud 3, quote 2, final 2, digit 2	5, . 4, — ,2 ־ ,2 י 1 . ,1 ן ,1 " ,1
s2-00-0.9982	punct 5, space 3, letter 3, quote 1, digit 1	. 3, 3, ,1 (,1 ן ז ,1 ו 1 י ,1 * ,1 ,
s2-00-0.9981	punct 4, space 3, letter 2, digit 1	3, . 2, ,1 (,1 כ ב ,1 " י 1 5 ,1 *
s2-abort	punct 4, letter 3, space 2, digit 2	. 2, 2, ,1 2 ,1 ח ת ,1 " י 1 י ,1 * ,1 (
s2-00-0.9986	**space 3, punct 3, letter 1**	**2, 1, — 1,) 1, * 1, 1 מ**
s2-00-0.9982-v1	punct 3, space 2, letter 2	1, ,1 . ,1 — ,1 ל ,1 ן ז 1 * ,1
s2-00-0.9984	punct 4, space 4, letter 3, nikud 2, quote 1, digit 1	4, ־ ,1 — ,2 . ,3 כ ב 1 ·, ,1 " ,1 * ,1
s2-00-0.9980	punct 4, space 4, letter 3, final 1, digit 1	4, . 2, ,1 ח ת ,1 כ ב ,1 ו ן 1 * ,1 ן ז ,1 —
s2-01-0.9982	letter 4, punct 2, digit 2, nikud 1, space 1	,1 2 ,1 ן ז ,1 ד ־ ,1 — ,3 כ ב 1 5 ,1 ,1 *
s2-01-0.9979	space 4, punct 4, letter 2, digit 1	4, . 1,) 1, ,1 ד ר ,1 ן ז 1 5 ,1 * ,1 —
s2-00-0.9985	punct 3, space 3, letter 2, nikud 1, digit 1	3, . 2, ,1 * ,1 ־ ,1 " י 1 1 3 ,1 כ

= nothing (a dropped or extra mark), = space. What is still really misread: a dropped line-end dash, dots in ". . .", a space, and one or two letter look-alikes (כ/ב, ת/ח, ן/ז).

Final comparison: 4,500 vol 39 lines (dev + test)



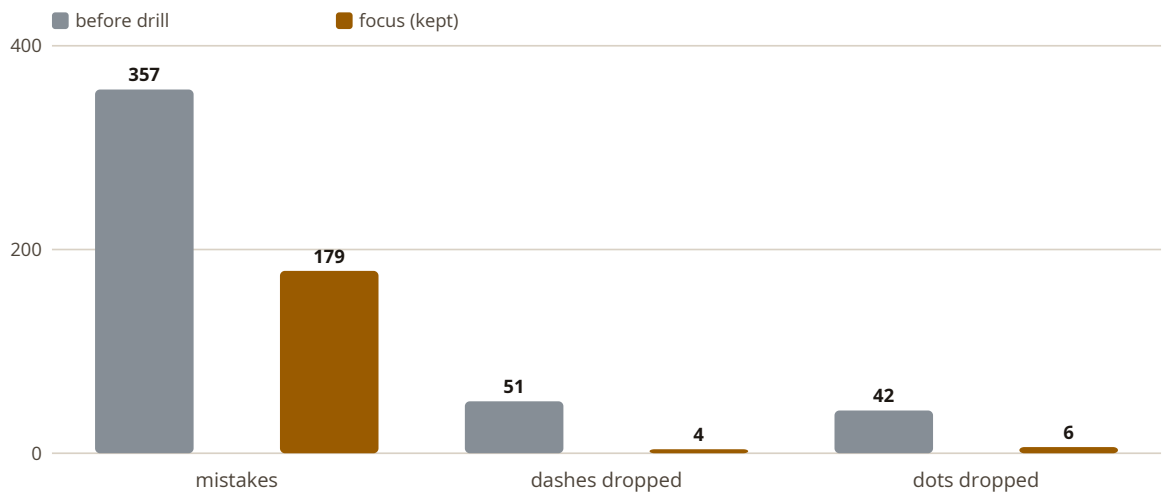
Letters on 4,500 lines, full float32 on the GPU. Labels: typed text after the relabeller and the ink fix, plus 35 line-edge dashes and ". . ." the focus model read that the labels still lacked (9 of 9 checked by eye were printed on the scan).

MODEL	LETTERS	WORDS	LINES PERFECT	BODY	NOTES	MISTAKES	H
recover start (10:01)	99.65	98.38	88.9	99.70	99.33	699	74
00-0.9985 (10:13)	99.81	99.41	96.2	99.82	99.79	372	24
00-0.9986 (10:23)	99.83	99.37	96.0	99.85	99.73	338	19
00-0.9984 (10:32)	99.83	99.36	95.9	99.84	99.74	346	14
00-0.9986-v1 (10:42)	99.80	99.30	95.6	99.82	99.70	399	18
01-0.9986 (10:52)	99.82	99.36	95.8	99.85	99.65	366	20
01-0.9986-v1 (11:02)	99.83	99.40	96.0	99.85	99.68	347	17
01-0.9987 (11:11)	99.82	99.38	96.0	99.84	99.69	361	17
01-0.9987-v1	99.82	99.41	96.1	99.84	99.69	357	18

MODEL	LETTERS	WORDS	LINES PERFECT	BODY	NOTES	MISTAKES	HA
(11:21, main best)							
focus (12:17), kept	**99.91**	**99.68**	**97.7**	**99.93**	**99.81**	**179**	**18**

"Hard" = כ/ב, ת/ח, ו/ן/ז mix-ups. Dash −/+ and dots −/+ = dropped / added.

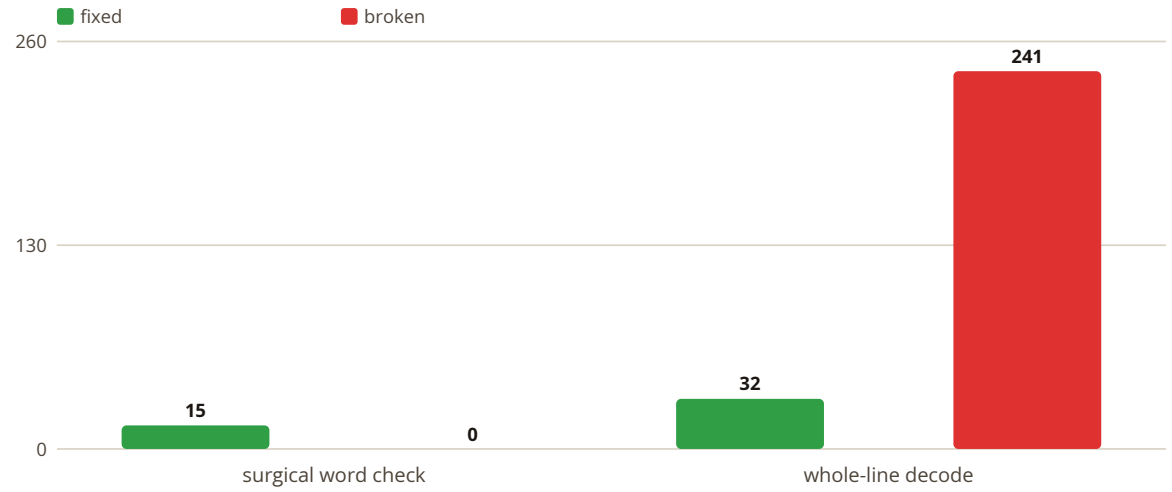
The focus drill, before and after



The drill halved the mistakes (357 179), almost all from marks: dropped dashes 51 4, dropped dots 42 6, with no invented dashes. The hard letters did not move (18 18) — the fake-letter drill did not teach them; they need a different approach.

Word check on the focus model

After reading, each word is checked against a vocabulary and a 5-gram word model built from the whole Chabad Library crawl (67,623 pages, 33.6M words, LS vol 39 left out). A look-alike letter (כ/ב, ת/ח, ו/ן/ז, ר/ד, ח/ה, ס/ם…) is swapped only where the word read is unknown and the swap gives a known word, or where both are words and the phrase with the swap is at least 30× likelier — and only where the network itself gave the letter weight. Nothing else is touched.



Hard-letter lines fixed (green) and broken (red). The surgical word check: 15 of 16 fixed, 0 broken. The old whole-line decoder: 32 fixed, 241 broken.

	LETTERS	WORDS	LINES PERFECT	MISTAKES	HARD-LETTER LINES
focus model, as read	99.87	99.55	97.0	179	16
+ word check (small vocabulary, vols 30–38)	99.87	99.59	97.2		11 fixed, 0 broken
+ word check (library vocabulary + phrase model)	**99.88**	**99.60**	**97.3**		**14 of 16 fixed, 0 broken** (17 lines fixed in all)
whole-line word-model decode (old beam), for comparison	99.74	98.91	92.3		32 fixed, 241 broken
+ word check, phrase margin 30×	**99.88**	**99.60**	**97.3**		**15 of 16 fixed, 0 broken (18 lines fixed in all)**

ⓘ The phrase margin was swept on the 4,500 lines: at 100×, 17 lines fixed; at 30× and even 10×, 18 fixed and none broken — so 30× is the default. The line it gained is the Yiddish איז read as אין (39_0215_b29): the library phrase model finds איז 37× likelier there, which 100× had refused. An LS-only phrase model (vols 30–38, 977k words) was too thin to tell them apart (0.16 log10), so the library model stays. Left: שכ׳ read as שב׳ (39_0102_n02) — an abbreviation; keeping the mark in the word key still finds it only 4× likelier, so this one needs the letter judge or a stronger context model.

Calibration — are the confidences honest?

s2-01-0.9979 on the 300 checked lines (`code/calib.py`).

LETTER CONFIDENCE	LETTERS	WRONG	REAL ERROR RATE	MODEL EXPECTS
below 0.5	1	1	100%	75%
0.5–0.95	37	0	0%	~15%
0.95–0.99	70	1	1.4%	1.8%
0.99–0.999	331	0	0%	0.35%
0.999 and up	12,604	1	0.008%	0.003%

ⓘ The letters it reads are honest. In the top bin it is wrong a little more often than it expects (0.008% vs 0.003%) — but that is one letter in 12,604. Most mistakes (13 of 16) are marks it left out, and a left-out mark has no confidence. Three lines lost a line-start dash "— " at confidence 0.996–1.0, so confidence cannot find them. Checking the 20 least sure lines (6.7%) finds 6 of the 9 wrong lines; the 3 it misses are exactly those dashes. 300 lines hold only 3 wrong letters — too few to settle it; repeat on the 4,500-line test.

Notes

- Without the 35 added marks (labels after the ink fix only), the focus model still wins: 99.87% letters and 97.0% lines perfect, against 99.84% and 96.3% for the main best. Against the raw typed text it scores lower (99.82%), because the typed text lacks the marks it now reads.
- The ink fix misses a mark when the reading model happened to read it (the ink is then "explained" but the label still lacks it). The next relabel should also take an edge mark the model reads when the ink check confirms ink there.
- Cross-check (10:45 UTC): s2-01-0.9979 read again in full on a CPU in full precision gives 99.88% letters, 99.52% words, 97.0% lines perfect, against 99.92 / 99.52 / 97.3 rebuilt from its pod report. That reading ran on a CPU in full precision and the pod's ran on the GPU in bf16, so a few lines come out a little differently. Treat the checked-label scores above as accurate to about ±0.04% letters.

- Per-save wrong-line reports: `errors/<save>.txt` next to the scores file.
- Earlier stages, for history: base ppocrv6 on the 4,500-line vol 39 test set — 96.45% letters, 86.39% words. Stage 1 best save — 99.14% letters on its dev set (old, unfixed labels).