Tiny Transformer: Predictions for Eight Runs

Claude and MJC

2026-11-24

Eight training runs of the tiny-transformer track, each with thresholds fixed in advance. The probabilities are Claude's. The architecture tt-2 and the checklist it must satisfy are in the literature review; the question behind most of these runs is whether an *induction* circuit (attend to the byte after an earlier occurrence of the current byte, and copy it) forms and persists when the training data is enwik9 text.

What is measured

bpc cross-entropy in bits per byte on enwik9 bytes $[10^7, 1.1 \cdot 10^7)$, in non-overlapping 128-byte windows. Context restarts in every window.

mass on 1000 windows (400 for removal), each a random printable segment of $R \in [24, 64]$ bytes repeated to 128: the largest, over heads, mean attention from a position after the first copy to the byte following the earlier occurrence of its byte. R varies so that a fixed offset cannot score.

copy loss bits per byte on the same windows, at positions after the first copy. Inside the first copy the segment is random and costs about 6.6.

removal the copy loss when a layer-0 component is replaced by its window mean: a head as a whole, or one writer's contribution to layer 1's Q, K and V inputs only.

copy gain bpc saved on held-out text by mixing in the byte after the longest in-window match.

Reference values

Run	Training	bpc	mass	copy loss	copy gain
tt-2-ep1	enwik9$[0, 10^7)$, 1 epoch	2.4352	< 0.02	9.74	0.069
tt-2	the same, 8 epochs	2.2183	< 0.01	9.71	0.057
diag-synth100	repeated segments only, 2 epochs' steps	7.62	0.81	0.02	0.554

tt-2: 2 layers, 4 heads, width 64, ReLU MLP 256, RMSNorm, RoPE, 115,008 parameters, context 128, batch 32. One epoch is 2441 steps. The copy loss above 8 on the enwik9 runs reflects the random bytes being outside their training data; the mass is the relevant column. All are single seeds, and the seed-to-seed spread of bpc is unmeasured.

Known when these were written

A one-layer model on repeated segments ended at copy loss 5.44 (synth-tt1), with bpc 6.38 on held-out text. A second seed of diag-synth100 (synth-s2) was at step 1750 of 4882 and its copy loss had fallen to 0.10; its interventions had not been run. None of the other runs had started.

Predictions

	Run	Predicted outcome	P
1	synth-s2: diag-synth100, seed 2	one layer-0 head removal raises copy loss above 5	0.75
2		removing the layer-0 MLP's contribution to Q, K and V costs more copy loss than removing any one layer-0 head's	0.60
3	survive: initialize from diag-synth100, 1 epoch of enwik9	mass ≥ 0.3 at the end	0.65
4		bpc below tt-2-ep1's 2.4352	0.75
5	warm-ep1: 1000 steps of repeated segments, then enwik9, 1 epoch	mass ≥ 0.3 at the end	0.55
6		bpc above tt-2-ep1's 2.4352	0.60
7	warm-ep8: the same, 8 epochs	mass ≥ 0.3 at the end	0.55
8		bpc below tt-2's 2.2183	0.60
9	mix10-ep8: 10% of windows repeated segments, 8 epochs	mass ≥ 0.3	0.85
10		copy loss below 1	0.80
11		bpc below tt-2's 2.2183	0.45
12		copy gain below tt-2's 0.057	0.75
13	tt-1: one layer, enwik9, 8 epochs	bpc more than 0.05 above tt-2's	0.90
14		mass < 0.05 in every head	0.95
15	tt-2-swiglu: SwiGLU MLP of width 176	bpc at least 0.01 below tt-2's	0.60
16		mass < 0.1	0.85
17	tt-2-nonorm: no RMSNorm	bpc above tt-2's	0.65
18		bpc above 5 (training failed)	0.10
19		mass < 0.1	0.88

Reading of 3, 5 and 7: copying is worth little to a model that already predicts text well (0.057 bpc for tt-2), and much more to one that does not (0.554 for diag-synth100), so an existing circuit has an incentive to survive early natural training; whether it lasts 8 epochs is nearly a coin flip. For 6 and 8, the warm-up replaces 1000 natural steps: 41% of the one-epoch run, 5% of the eight-epoch run.

Scoring

Each line is true or false from the finished run's `eval.json`, `copyval.txt` and `ablate.txt`. A run that fails to complete voids its lines, except that a diverged run counts for 18. The score is the mean of $(P - \text{outcome})^2$; always saying one half gives 0.25.