Training language models without backpropagation

Technical summary · Yuren Hao, University of Illinois Urbana-Champaign · August 2026 · yurenh2@illinois.edu

What we do

We train standard transformer language models from scratch with Equilibrium Propagation (EP), a learning rule in which the training signal comes from the system relaxing twice and reading local differences, with no backward pass anywhere. The trained model is an ordinary transformer and runs ordinary forward inference. Every design choice is constrained to operations a physical substrate can perform, because the point of the exercise is a learning rule that analog and other physics based hardware can run natively.

Why this question is open. Backpropagation requires a global backward pass that a physical system cannot perform on itself, so an analog accelerator must either digitize every intermediate state or carry a second adjoint copy of the hardware. Both destroy the energy advantage that motivates the substrate. Learning rules that a substrate can run natively do exist, and the published evidence for them stops at small vision models. Whether they hold at language model scale is the gap this project addresses.

Results to date (all against matched backprop twins)

Twin discipline: identical architecture, tokenizer, data order, optimizer, step budget, and evaluation. The only difference is the training rule. Two seeds per arm at the smaller sizes, three backprop seeds at 72M. Corpus is FineWeb-Edu with a 32k vocabulary.

Model	EP (val CE)	Backprop twin	Perplexity difference
11M	4.194	4.197	-0.4% (EP ahead)
18M	3.945	3.933	$+1.2\%$
36M	3.649	3.648	$+0.1\%$
72M	3.400	3.397	$+0.4\%$
135M	rerun in progress	3.469	level at matched progress

All values are tail window means over the last tenth of training, which is the statistic we quote everywhere. Every difference above is smaller than or comparable to the spread between backprop seeds, which is 0.003 in cross entropy at 72M, so EP and backpropagation are statistically indistinguishable at these sizes. The 135M entry compares both arms at the same point in an unfinished run and carries one seed per arm, so it is provisional. To our knowledge the 135M model is the largest language model trained from scratch with no backpropagation at any layer.

Training cost is 2.7 times backpropagation in wall clock, measured on the same GPU with the same model and batch and with all other load suspended, and 3 times in analytic FLOPs. We quote the measured ratio against backprop rather than a relaxation-iteration count, because iteration counts in this literature measure different operations: the closest comparable work updates layers asynchronously, so one of its iterations moves information about one layer, while one of ours transmits through the whole stack. That work reports 6.7 to 12 times backprop in wall clock.

A result from this week, and why it matters for how we work

Between 72M and 135M we had been measuring a quality loss that grew with model width, resisted every algorithmic remedy we tried, and looked exactly like a scaling limit of the learning rule. A per layer probe traced it instead to arithmetic in our own simulator. The training signal is a small displacement added to a much larger activation; in single precision the components of that displacement below machine epsilon are destroyed by the addition, and the readout, which recovered the displacement by subtracting the activation back out, returned the damaged copy without any error signal. Wider models put more layers below that threshold, which is why the loss looked like a scaling wall.

Reading the training signal from the stored displacement directly, never as the difference of two large numbers, removes the effect at no computational cost. Verification was pre registered: a double precision control, a test that shifted the affected layers exactly as predicted when the nudge amplitude was scaled by eight in each direction, a five width probe showing the damage appear precisely where the displacement ratio crosses the precision threshold, complete closure of the measured loss on a controlled instrument, and a null test confirming the fix changes nothing at the sizes that were never below the threshold.

We report this here because it is representative of how the project is run and because it is likely to matter to others. Any implementation of this family of learning rules that recovers the contrast by differencing two large states will hit

the same floor, and the symptoms mimic an algorithmic limitation closely enough that it would be easy to publish the wrong conclusion.

What compute would buy

1. **Validation at 300M** on 6B tokens with a matched backprop control. Report within three weeks of access. Roughly 250 H100 hours by measured throughput.
2. **A matched ladder from 150M to 600M**, two seeds per arm, with the hardware relevant measurements (quantization, injected device noise, operating windows) run at every size rather than as a separate track. Roughly 2,750 H100 hours.
3. **A 1B stage**, 2,000 to 3,000 H100 hours, conditioned on agreed loss and stability gates from the first two.

What gets released

Checkpoints at every size for both the EP models and their backprop twins, full training logs, training and measurement code, the probe that produced the precision result above, and the negative results. The intent is a reference artifact for the physical learning and neuromorphic communities: the empirical answer to whether hardware compatible learning scales, in a form other groups can build on without spending the compute again. Publication is open by default.

People

Yuren Hao (UIUC) leads the project, advised by ChengXiang Zhai on the language modeling side and by Rainer Engelken on learning dynamics. Collaborators include Alexi Gladstone (UIUC, energy based models), Xiang Wan (Stanford), and Zeyi Liu (UIUC). The hardware measurement program is being designed in consultation with experimentalists in physical learning.