

Backprop-free training of transformers — results and a staged hardware program

One-page brief for hardware-side collaborators · rev. 2026-07-12 · Yuren Hao (UIUC)

The idea in three sentences

We train **standard multi-layer transformers** with **Equilibrium Propagation** on a layered energy: training consists of two relaxation phases and a **local** contrast update per weight — no backpropagation anywhere — and inference is an ordinary forward pass. On GPU this now works at language-model scale with essentially no quality gap to backprop. The hardware program starts with the cheapest object that can validate the physical learning rule — a **clockless twin-network analog tile (~\$300)** — and climbs rung by rung to an in-memory-compute transformer block.

GPU-scale results (2026-07, measured)

- **A 12-layer, 42.7M-parameter transformer LM trained for a full epoch (59k steps, 361M tokens) with no backpropagation in the training loop; it generates coherent text.** To our knowledge the first transformer language model trained fully this way.
- **Gap to a tuned, same-architecture backprop control: 0.05 nats** (at 4k steps: statistically indistinguishable, 3 seeds/arm). Prior backprop-free attempts at scale all report qualitative gaps.
- **EP step = 3.2× backprop FLOPs** (measured); mixed-precision training validated; the two known EP-specific instabilities are mechanistically diagnosed and closed (an estimator-SNR floor with a β -schedule law; a relaxation-contractivity crossing eliminated by norm placement).
- **Every trained operation chosen analog-implementable:** crossbar MVM, divisive normalization, fixed I/Q rotations (position code), translinear gated MLP, subthreshold-exponential softmax, two-phase relaxation for the learning rule.

Measured fault tolerances (fault injection at the trained model)

fault	free	marginal	dead
weight precision	8-bit (Δ CE +0.004)	6-bit (+0.05)	4-bit
forward state noise	1%	—	—
error-channel (nudge) noise	10% relative	30%	—
divider mismatch / gate gain / phase error	3% / 10% / 0.03 rad	10% / — / 0.1 rad	—

Under every non-fatal fault the learning signal tracks the *faulted* network (gradient cosine ≈ 0.97 invariant): **the rule co-adapts to the device**. The only hard spec is ~ 7 -bit effective weights.

The hardware ladder (each rung publishable alone)

1. **One-edge metrology tile (\$70–130):** twin MOSFET edge, shared weight capacitor, exact (difference-of-squares) and sign-only local update channels, OTA current nudge — no processor, converter, clock, or sampled memory in the learning loop.

2. **8-edge twin network (\$170–300)**: nonlinear regression; EP current-nudge vs Coupled-Learning voltage-clamp on one board; exact-vs-sign update comparison; measured bias-vs-nudge-magnitude curve (the same β -SNR law we measured in simulation).
3. **32-edge network (\$450–900)**: replication-class nonlinear tasks, robustness study.
4. **Reciprocal attention microcell (+\$100–250)**: two tokens, one head, energy-based attention.
5. **CIM transformer block (partner phase)**: analog MVM + in-situ two-phase EP weight update — the piece no shipping analog-AI chip has (all are inference-only or on-chip-backprop).
6. **North star: a few-M-parameter TinyStories LM trained on analog hardware.**

What we bring / what we ask

Bring: the trained models and recipe, the estimator theory (β -SNR law, stability walls), the measured tolerance ledger, SPICE-first costed build plan, and parts funding (rungs 1–3 are <\$1k). **Ask (rungs 1–3):** bench access, analog-design mentorship, and/or a student who enjoys discrete analog — six-week plan, instruments = a scope and a DMM. **Ask (rung 5):** a CIM/mixed-signal partnership where the substrate expertise is yours and the learning rule is ours.

(Detail: CLOCKLESS_ANALOG_MVP_PLAN.md — full BOM, schedule, acceptance criteria, claim limits; COMPONENT_HW_MAP.md — per-operation analog mapping + tolerance status.)