EP transformer program → analog board: call sheet

supplements the July 12 brief and build plan · Yuren Hao (UIUC) · August 2026

Program state since the July brief (measured)

- **72M-parameter transformer LM sealed:** EP final CE 3.364 (2 seeds) vs. a tuned backprop twin at 3.288 (3 seeds, ±0.006); gap +0.076 nats (≈8% perplexity). The symmetric two-phase estimator narrows this to +0.044 (≈4.5%) at 1.67× cost. Zero instability events in both EP seeds. To our knowledge the largest LM trained with no backpropagation anywhere in the loop.
- **135M trained end to end** (2.7B tokens), also with zero instability events. It surfaced a width-dependent estimator loss localized to the upper layers — which a per-layer probe has since traced to the simulator's own *readout floor*: at larger widths the nudge displacement falls below single-precision machine epsilon relative to the activations and is rounded away in the contrast read.
- **The fix is free, and it is the software version of differential readout:** read the contrast from the stored displacement itself, never as the difference of two large signals. Verified against a double-precision control and a pre-registered amplitude-shift test; the loss closes 100% at zero cost. At the smaller rungs of our scaling ladder, EP now matches or beats the matched backprop twin (2 seeds per arm).
- **Random-sign nudging** (nudge polarity flipped randomly per batch) is the stability winner at both scales: it removes the systematic single-sided bias at zero extra cost. On a board it is a polarity switch.

The four primitive measurements (recipe-independent)

measurement	simulation anchor (measured)	board question
1. contrast-update fidelity (exact difference-of-squares channel)	update tracks the *faulted* network at cos ≈ 0.97 under 8-bit weights, 1% forward noise, 10% nudge noise	does the same co-adaptation law hold in physics
2. nudge operating window and its floor	ceiling tracks loop gain and sinks over training (~ $1/\sigma^2$ of the output map); error channel is amplitude-indifferent across 20× (multiplicative), so the floor is set by *additive* readout noise	measure the physical floor; track window drift over training
3. sign-only vs. exact vs. random-sign updates	sign works; random sign removes the bias at scale; exact channel is the reference	continuous three-way comparison on one board
4. EP current nudge vs. CL voltage clamp	the distinction formalized by McGinnis, Li, and Mori	same board, same task, same metrology

Where we most want your judgment

1. Is this the right primitive-level list? What would you add or cut, from your imperfection-characterization experience?
2. **Reading the contrast:** our width-scaling loss was ultimately a readout-floor artifact, and the lesson is that the contrast must be measured differentially, never as a difference of two large signals. How do your update cells read the contrast, and what sets the smallest resolvable contrast (the physical epsilon) on the board?
3. **Additive vs. multiplicative separation:** in simulation we sweep nudge amplitude at a fixed task. What is the right bench protocol for the same separation on the board?
4. **Settling as a gate:** in simulation the relaxation residual predicts every blowup and gates the weight update. What settling observable did your boards expose, and did you gate on it?
5. **Twin replicas vs. a time-multiplexed single network** at 8–32 edges: matching cost against memory cost, your call.

We are not committing the board before this conversation; rungs 1–3 remain under $1k in parts, funded on our side. Results are intended for open publication.