LATENT TEST-TIME COMPUTATION WITH ENERGY TRANSFORMER COMPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy transformers (ET) offer an intriguing perspective on computation in looped transformers: with only minimal architectural differences from a standard transformer, their forward pass iterates a preconditioned gradient step on a learned energy. Could iterating ET produce qualitatively different computation from iterating a regular GPT? We train **9 hybrid ET/GPT architectures** at 400M, 800M, and 1B on 126B tokens, varying attention, FFN, and update layout. LogitLens reveals two regimes: *distributed refinement*, where energy-attention blocks improve the prediction every iteration (correct-token rank $110 \rightarrow 2.3$), and *concentrated construction*, where standard attention defers prediction to the final block. Energy attention is necessary but not sufficient: pairing it with SwiGLU under the parallel layout overshoots ($\mathbf{2.9}\times$ the hidden-state norm) and collapses ($+\mathbf{555}\%$ PPL at T=3); either the energy FFN or the consecutive layout bounds updates and restores robustness. A convergence-rate halting criterion saves **19%** measured FLOPs at $+0.5\%$ PPL for EGPT-Dual; standard-attention models tolerate only modest halting and lose quality under aggressive iteration reduction. At full iterations all models reach similar quality, though EGPT-Dual lags on GSM8K at 1B (12.0 vs. RecGPT 18.3), a gap absent at 400M/800M. We propose that the difference stems from the value tying $\boldsymbol{W}_V = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ of energy attention, which confines each block's writes to its own read space. In trained models these write operators align *bimodally* with the LM-head's singular spectrum, with excess **+0.13 to +0.24** in the top-32 (dominant readout directions) and **+0.34 to +0.43** in the bottom-32 (near-null tail), while standard attention stays close to random ($\leq +0.09$) at every scale. Write operators are also shared across blocks far above chance (0.035–0.12 for EGPT-Dual vs. ≤ 0.002 for RecGPT). The trained models are energy-structured rather than energy-minimising.

1 INTRODUCTION

Energy-based modeling casts desirable datapoints as low-energy, high log-likelihood points that can be generated via explicit optimization. The Energy Transformer (ET) (Hoover et al., 2024) showed that the core attention and feed-forward network (FFN) of a transformer block can be constructed as a gradient descent step of a scalar energy. The forward pass through ET is a recurrent, multi-step, gradient-based optimization on a fixed energy landscape. Dehmamy et al. (2026) extended ET to language modeling and showed that such energy-based GPT (EGPT) models can match or exceed standard transformers in quality.

Superficially, EGPT looks like a recurrent (looped) GPT with extra weight tying ($\boldsymbol{W}_V = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$). Looping shared weights is a powerful axis for compute: looped LMs scale to billions of parameters while matching larger standard transformers (Zhu et al., 2025); concurrently, Prairie et al. (2026) establish scaling laws and a spectral stability framework, and Blayney et al. (2026) show looped models self-organise into "stages of inference." Energy-based attention (Krotov & Hopfield, 2016; Ramsauer et al., 2021; Hoover et al., 2024) reads attention as a step of energy descent. A broader review of adaptive-compute methods and the recurrent-transformer line is in App. T. Most prior works use *group looping*, where a block-group repeats as a unit (Figure 2, left). We study a different regime: *per-block recurrence* (Figure 2, right), where each block iterates independently before passing to the next. This is a harder stability requirement, since each block must converge its own representation without assistance from later blocks in the next recurrence pass.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

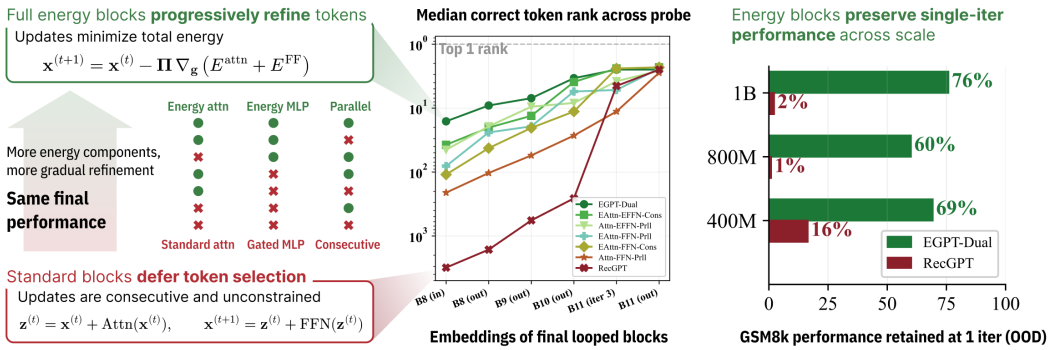


Figure 1: **Energy-Transformer components produce clean, predictable token dynamics, even when run for fewer iterations than seen during training.** *(Left & Middle)* A LogitLens probe decodes hidden embeddings into the final vocabulary, showing the correct token prediction moving steadily toward the top rank 1 prediction across all recurrent blocks (B8–B11). Note the log-scale y-axis, where standard architectures defer correct token prediction (rank $\sim$250 $\rightarrow$ rank 2) to final recurrent blocks, leaving intermediate states much less predictive. Richer energy structure with more energy components yields smoother trajectories. *(Right)* This behaviour holds across model scales (400M, 800M, 1B params) under out-of-distribution iteration reduction: at one loop iteration, EGPT-Dual preserves 60–76% of its training iteration GSM8K performance from 400M to 1B, while RecGPT preserves only 1–16%. See Sections 3.3 and 4.1for experimental details.

However, little is known about how energy structure affects the behavior of looped transformers. To our knowledge, no prior work has studied how energy-based components interact with looping, or whether they produce qualitatively different computation from standard looped designs. This leads to our central questions:

Are there fundamental differences between how looped transformers and energy transformers compute? How does the introduction of energy modules affect refinement of representations as the computation progresses? Do energy components affect the usefulness of intermediate states?

The answer determines whether a recurrent model can trade iterations for quality at inference time, and whether energy structure is a useful inductive bias for recurrence. Each energy component has both a clear theoretical motivation *and* a standard counterpart (energy attention vs. standard attention, energy FFN vs. standard SwiGLU, consecutive vs. parallel application of these components inside the transformer block), enabling a controlled setting that isolates how each choice shapes the computation.

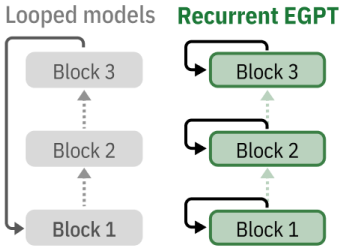


Figure 2: **Group looping vs. per-block recurrence.** Traditional looped models loop a block-group as a unit. Our recurrent EGPT iterates each block independently.

We train 9 architectures from scratch at 400M, 800M, and 1B on 126B tokens with shared data and hyperparameters, then probe them with per-iteration LogitLens (Belrose et al., 2023), which projects intermediate hidden states through the language model head to decode what the model "believes" at each iteration. This distinguishes our work from prior studies of energy transformers on vision (Hoover et al., 2024) and small-scale language modeling (Dehmamy et al., 2026).

Our goal is not to outperform recurrent or standard baselines at full iteration count; at full iterations the architectures we study reach similar quality (Table 2). Instead, we ask how energy components change the way recurrent blocks compute, and what this implies for the usefulness of intermediate states and for running fewer iterations at inference. Our main findings:

1. **Not all recurrent transformers compute the same way.** Energy-attention blocks show *distributed refinement*: they improve the prediction at every block, so the correct token is already near the top (rank $\approx$3) when entering the final block (Figure 1). Standard-attention blocks show *concentrated construction*, deferring the prediction to the final iterations.

Energy attention is necessary for this but not sufficient: with SwiGLU in the parallel layout the update overshoots (2.9× the hidden-state norm), and either the energy FFN or the consecutive layout keeps updates below 0.8× and restores robustness.

2. **Energy and distributed refinement enable free compute savings over concentrated construction.** Energy attention models have convergence-rate criterion that saves 19% of measured FLOPs at 400M scale, and 12.8% savings at 1B scale, with no quality loss. Standard-attention models can be halted only modestly; under aggressive halting they lose significant GSM8K performance (Figure 1, right). All findings replicate across 400M, 800M, and 1B scales.

3. **Bimodal LM-head alignment is consistent with value tying as the mechanism.** Value tying $\boldsymbol{W}_V = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ confines each energy-attention block's writes to its own read space. In trained models these write operators align with both ends of the LM-head spectrum (excess $+\mathbf{0.13}$ to $+\mathbf{0.24}$ in the top-32 directions and $+\mathbf{0.34}$ to $+\mathbf{0.43}$ in the bottom-32) and are shared across blocks far above chance, at all three scales. Standard attention shows neither signature.

Energy attention constrains representations to remain prediction-aligned across iterations, while the energy FFN or the consecutive layout bounds update magnitudes and prevents the overshoot that would otherwise destroy this alignment. For practitioners: use energy attention with the energy FFN or the consecutive layout when intermediate states must be usable or iterations may be cut at inference. Standard recurrent architectures are a fine choice when full-iteration quality is the only objective.

2 Energy GPT and recurrent transformer blocks

This section defines the three energy-inspired components we study and the hybrid architecture that serves as our testbed (Table 1). A *recurrent* block applies its weights T times in its forward pass. We compare two update rules:

Consecutive (Cons). Attention and FFN are applied as two residual substeps, as in standard GPT and RecGPT baselines: $\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{Attn}(\mathrm{RN}_1(\boldsymbol{h}))$; $\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{FFN}(\mathrm{RN}_2(\boldsymbol{h}))$.

Parallel / energy-style (Prll). Attention and FFN are computed from the same state, combined, and passed through a learned preconditioner. For the per-token energy $E_A = E_A^{\mathrm{attn}} + \lambda E^{\mathrm{ff}}$, the single-projection update is $\boldsymbol{h} \leftarrow \boldsymbol{h} - \boldsymbol{\Pi}(\mathrm{Attn}(\boldsymbol{h}) + \lambda \mathrm{FFN}(\boldsymbol{h}))$, which is one *preconditioned, query-role partial-gradient* step on E_A (there is no single global scalar energy; each token has its own E_A, and the gradient is taken only through the query role, equivalently stop-gradient on the key role, to preserve autoregressive causality). This coincides with energy descent only when the symmetric part $\boldsymbol{\Pi}^s \succeq 0$, which we do not impose in training (see section 5). Allowing distinct $\boldsymbol{\Pi}_a, \boldsymbol{\Pi}_f$ for the two channels generalizes beyond gradient flow into Helmholtz form (section D); we call this variant EGPT-Dual.

Table 1: **Naming conventions for recurrent block updates**. EAttn = energy attention ($\boldsymbol{V}{=}\boldsymbol{W}_Q^\top \boldsymbol{K}$); Attn = standard softmax. EFFN = energy FFN; FFN = standard SwiGLU. Π = learned $d{\times}d$ preconditioner.

Model	Attention	FFN	Layout	Update rule
GPT / RecGPT	Standard	SwiGLU	Consec.	$\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{Atn}$; $\boldsymbol{h} \leftarrow \boldsymbol{h}{+}\mathrm{FF}$ ($T{=}1/T{>}1$)
EGPT-Dual	EAttn	Energy	Parallel	$\boldsymbol{h} \leftarrow \boldsymbol{h} - (\boldsymbol{\Pi}_a \mathrm{EAtn} + \lambda \boldsymbol{\Pi}_f \mathrm{EFF})$
Attn-FFN-Prll	Standard	SwiGLU	Parallel	$\boldsymbol{h} \leftarrow \boldsymbol{h} - \boldsymbol{\Pi}(\mathrm{Attn} + \lambda \cdot \mathrm{FF})$
EAttn-FFN-Cons	EAttn	SwiGLU	Consec.	$\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{EAtn}$; $\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{FF}$
EAttn-FFN-Prll	EAttn	SwiGLU	Parallel	$\boldsymbol{h} \leftarrow \boldsymbol{h} - \boldsymbol{\Pi}(\mathrm{EAtn} + \lambda \cdot \mathrm{FF})$
Attn-EFFN-Prll	Standard	Energy	Parallel	$\boldsymbol{h} \leftarrow \boldsymbol{h} - \boldsymbol{\Pi}(\mathrm{Attn} + \lambda \cdot \mathrm{EFF})$
EAttn-EFFN-Cons	EAttn	Energy	Consec.	$\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{EAtn}$; $\boldsymbol{h} \leftarrow \boldsymbol{h} + \mathrm{EFF}$

2.1 COMPONENTS

For the standard GPT baseline we use standard attention with RMSNorm and SwiGLU (Shazeer, 2020) as the FFN: $\mathrm{SwiGLU}(\boldsymbol{g}) = \boldsymbol{W}_o[\mathrm{swish}(\boldsymbol{W}_g\boldsymbol{g}) \odot \boldsymbol{W}_u\boldsymbol{g}]$ with $\boldsymbol{g} = \mathrm{RN}(\boldsymbol{h})$ (RMSNorm). We compare against the two energy variants below.

Energy FFN. Differentiating the scalar energy $E^{\mathrm{FF}}(\boldsymbol{h}) = -\boldsymbol{g}^\top \boldsymbol{W}_2\, \sigma(\boldsymbol{W}_1\boldsymbol{g})$ gives

$$\mathrm{EFFN}(\boldsymbol{h}) \;=\; -\boldsymbol{\Pi}\nabla_{\boldsymbol{g}} E^{FF}(\boldsymbol{h}) \;=\; \underbrace{\boldsymbol{\Pi}\boldsymbol{W}_2\, \sigma(\boldsymbol{W}_1\boldsymbol{g})}_{\text{Standard MLP}} \;+\; \underbrace{\boldsymbol{\Pi}\boldsymbol{W}_1^\top\big[\sigma'(\boldsymbol{W}_1\boldsymbol{g}) \odot \boldsymbol{W}_2^\top\boldsymbol{g}\big]}_{\text{SwiGLU}}, \tag{1}$$

a standard MLP *plus* a gated branch with SwiGLU-like form (σ' as the gate; weight tying $\boldsymbol{W}_u = \boldsymbol{W}_2^\top$, $\boldsymbol{W}_o = \boldsymbol{\Pi}\boldsymbol{W}_1^\top$). EFFN realises SwiGLU-style gating with $\approx 2\,d\,d_{\mathrm{ff}} + d^2$ parameters versus SwiGLU's $\approx 3\,d\,d_{\mathrm{ff}}$. Full gate-substitution analysis: Appendix C.

Energy attention. Hoover et al. (2024) (adapted to GPT in Dehmamy et al. (2026)) derive an attention variant as the gradient of an energy. With $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_h \times d}$, the per-token attention energy is

$$E_A^{\mathrm{attn}}(\boldsymbol{h}) = -\log \sum_{B<A} \exp\Big[\boldsymbol{Q}_A\boldsymbol{K}_B^\top/\sqrt{d_h}\Big], \quad \boldsymbol{g} = \mathrm{RN}(\boldsymbol{h}),\ \boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{g},\ \boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{g}. \tag{2}$$

Its gradient *through the query role only* (stop-gradient on the key role, preserving autoregressive causality) yields the EGPT attention update

$$\mathrm{EAttn}(\boldsymbol{h})_A = -\boldsymbol{\Pi}\nabla_{\boldsymbol{g}_A} E_A^{\mathrm{attn}}(\boldsymbol{h}) = \frac{\boldsymbol{\Pi}}{\sqrt{d_h}} \sum_{B<A} \mathrm{softmax}\Big(\boldsymbol{Q}_A\boldsymbol{K}_B^\top/\sqrt{d_h}\Big)\,\boldsymbol{V}_B, \quad \boldsymbol{V} = \boldsymbol{W}_Q^\top\boldsymbol{K}, \tag{3}$$

which ties $\boldsymbol{W}_V = \boldsymbol{W}_Q^\top\boldsymbol{W}_K$ (eliminating an independent $\boldsymbol{W}_V$) and replaces the value projection by the preconditioner $\boldsymbol{\Pi}$. The sum runs strictly over $B < A$ (causal); the first token (A=0) has an empty sum and receives no EAttn update.

Interpretation. The attention energy has a natural probabilistic reading. By the polarization identity it is, up to a per-query quadratic, the negative log-likelihood of the query under a Gaussian mixture centred on the observed past keys, so each iteration improves the fit of the current token to its context. It also connects energy attention to Dense Associative Memories, in which the query is a state attracted to the keys as memories (Krotov & Hopfield, 2016; Krotov et al., 2025). Following Hoover et al. (2024) and Dehmamy et al. (2026), gradients are taken with respect to $\boldsymbol{g} = \mathrm{RN}(\boldsymbol{h})$. Derivations and the gradient convention are in App. A.

The components above can be freely combined with their standard counterparts (standard attention, SwiGLU, consecutive layout), yielding a factorial design space that we explore in Section 3.

3 TWO COMPUTATION MODES IN RECURRENT TRANSFORMERS

If recurrence is a preconditioned, partial-gradient iteration on an energy function, one might expect that all iterations contribute equally to refining the prediction. But does this hold for standard attention recurrence too, where no energy function exists? And does it matter for practical robustness? We address these questions by training 9 hybrid architectures that systematically vary the recurrent block components, evaluating their quality, testing their robustness under iteration reduction, and then looking inside with LogitLens to reveal the mechanism.

3.1 ARCHITECTURE VARIANTS

A 4-block study at 200M (Appendix H) shows that iteration helps (RecGPT 26.6 vs. GPT 31.5 PPL) but pure EGPT underperforms (34.7 PPL), energy constraints are too restrictive when every layer must obey them. We therefore use a *hybrid* architecture: 8 standard GPT blocks (B0–B7) provide general feature extraction, followed by 4 recurrent blocks (B8–B11) where we vary the energy-based components. At 400M/800M, blocks B8–B11 use iterations $[3, 4, 6, 9]$ with iteration

Table 2: Quality at full training iterations across scales. Avg% = mean of 11 benchmarks (HellaSwag, ARC-E/C, PIQA, BoolQ, COPA, SciQ, WinoGrande, RACE, OpenBookQA, LAMBADA; App. J); acc_norm where available. GSM8K^{h}:8-shot, ####-delimited format. B8–B11 iterations: $[3, 4, 6, 9]$ with dropout $\pm[1, 2, 2, 4]$ at 400M/800M; $[2, 3, 4, 6]$ with dropout $\pm[1, 1, 2, 2]$ at 1B.

Scale	Model	Params	PPL ↓	Avg% ↑	MMLU% ↑	GSM8K^{h}%
400M	EGPT-Dual	**417M**	22.28	52.2	29.8	6.5
	RecGPT	443M	22.37	**53.3**	29.8	**6.7**
	GPT	443M	**22.05**	52.4	29.8	5.9
800M	EGPT-Dual	**745M**	19.00	55.4	34.8	10.5
	RecGPT	796M	19.56	55.0	**35.1**	10.5
	GPT	796M	**18.82**	**56.4**	**35.1**	**10.6**
1B	EGPT-Dual	**944M**	17.79	56.5	35.3	12.0
	RecGPT	1010M	18.12	56.7	37.2	**18.3**
	GPT	1010M	**17.59**	**57.9**	**37.4**	15.2

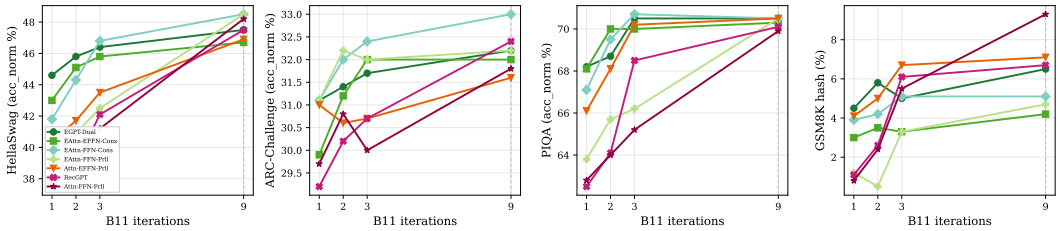


Figure 3: Benchmark under iteration reduction (400M). T=6–9 are ID, T=1–3 are OOD. The robust energy-attention models (greens) degrade gracefully even at T=1; standard-attention models (reds) collapse progressively.

dropout $\pm[1, 2, 2, 4]$; at 1B we use $[2, 3, 4, 6]$ with dropout $\pm[1, 1, 2, 2]$ (full details in Appendix I). The heterogeneous schedule allocates more iterations to later blocks, but the robustness partition is not an artifact of it: the partition reappears under a uniform schedule of six iterations per block (Appendix Q). By sharing the same standard prefix and varying only the recurrent blocks, we isolate each component's effect. The prefix is trained jointly with the recurrent blocks, and energy recurrent blocks make even the shared prefix more prediction-aligned: the correct-token rank entering B8 is 16 for EGPT-Dual versus 3127 for RecGPT (Table 17). Table 1 defines the architectures as a factorial over attention × FFN × layout. EGPT-Dual is the full Energy Transformer with dual preconditioner; RecGPT is a standard GPT iterated with shared weights; the remaining variants mix components.

3.2 Quality and Robustness to Fewer Iterations

All models are pretrained from scratch on 126B tokens (Appendix I). At full iterations all three primary architectures reach similar quality (Table 2), though EGPT-Dual lags on GSM8K at 1B (12.0 vs. RecGPT 18.3; a trade-off we discuss in §6); the differentiator is what happens when iterations are reduced.

Throughout this paper, "robustness" means robustness to reduced test-time iterations, not a broader notion such as robustness to distribution shift or adversarial inputs. We vary B11's iteration count from T=9 (training) down to T=3 (OOD: below the iteration-dropout minimum of T=5). A clear partition emerges. Reducing an earlier block instead, or all blocks together, preserves the same ordering, though the gaps are smaller when only B10 is reduced (Appendix R). Three energy-attention models degrade gracefully: **EGPT-Dual** (+6.1%), **EAttn-EFFN-Cons** (+3.9%), and **EAttn-FFN-Cons** (+9.8%). Two models degrade substantially (Attn-EFFN-Prll +122%, RecGPT +136%), and two collapse (EAttn-FFN-Prll +555%, Attn-FFN-Prll +1736%). The gap extends to downstream benchmarks (Figure 3; full PPL table in Appendix M, all-benchmark table in Appendix N). All robust models use energy attention, but EAttn-FFN-Prll also uses energy attention yet collapses at +555%. Attention alone is not the answer.

Table 3: Rank trajectory inside B11 and update magnitude in B9 (400M). Robust models refine predictions with bounded update magnitudes. EAttn-FFN-Prll *overshoots*: its rank rises at t=1–2, driven by a B9 update of 2.93× the hidden-state norm. Attn-FFN-Prll's rank also rises at t=1, but its early updates are tiny (0.005×) and it makes a large late correction instead. RecGPT starts too high (concentrated construction). Values > 1.0 for $\|\Delta \boldsymbol{h}\|/\|\boldsymbol{h}\|$ indicate the update exceeds the hidden state norm.

Model	Mode	Input	t=1	t=2	t=3	t=5	t=9	Top-1 at t=3	B9.1 $\|\Delta\|$	B9.2 $\|\Delta\|$	Robust?
EGPT-Dual	Refinement	3.4	3.1	3.3	**2.5**	2.4	2.5	**82.4%**	0.55	0.47	**Yes**
EAttn-EFFN-Cons	Refinement	3.9	3.1	2.6	**2.4**	2.3	2.3	**86.4%**	0.34	0.17	**Yes**
EAttn-FFN-Cons	Refinement	11.3	6.2	3.3	**2.4**	2.3	2.3	**84.2%**	0.70	0.40	**Yes**
EAttn-FFN-Prll	Overshoot	5.5	14.2	18.1	5.2	2.5	2.4	51.9%	0.63	**2.93**	No
Attn-FFN-Prll	Overshoot	26.9	33.0	21.5	11.3	2.8	2.8	43.3%	0.005	0.006	No
RecGPT	Late constr.	**258**	93	29	4.5	2.4	2.5	52.5%	–	–	No

3.3 LogitLens reveals two computation modes

We apply LogitLens (Belrose et al., 2023) at every (block, iteration), projecting hidden states through the final RMSNorm and LM head and measuring median correct-token rank, entropy, top-1 match with the final prediction, $D_{\mathrm{KL}}(p^{(t)} \| p^{(T)})$, and convergence rate $\|\boldsymbol{h}^{(t)} - \boldsymbol{h}^{(t-1)}\|/\|\boldsymbol{h}^{(t-1)}\|$ on 400M models. Per-block rank trajectories (Appendix Table 17, Figure 7) split sharply: all energy-attention models enter B11 with low rank (EGPT-Dual 3.4, EAttn-EFFN-Cons 3.9, EAttn-FFN-Prll 5.5, EAttn-FFN-Cons 11), while softmax models enter at rank 8.3 to 258. The lowest softmax entry rank belongs to Attn-EFFN-Prll, which has the energy FFN: energy components pull intermediate states toward the prediction even without energy attention, and this model is also the least-collapsing softmax variant (+122%). Entropy on entering B11 tells the same story: EGPT-Dual and EAttn-FFN-Cons enter at 2.1 and 3.4 nats, while RecGPT enters at 6.1 nats and sharpens only inside B11. We call these two regimes **distributed refinement** (each block improves the prediction; intermediate states are useful) and **concentrated construction** (intermediate blocks transform in a non-output-aligned space; the final block performs the critical conversion).

Three convergence behaviours in B11. Per-iteration rank trajectories *within* B11, together with update magnitudes $\|\Delta \boldsymbol{h}\|/\|\boldsymbol{h}\|$, identify three convergence behaviours (Table 3). **Refinement** (EGPT-Dual, EAttn-EFFN-Cons, EAttn-FFN-Cons): enter B11 at rank 3–11, reach 82–86% top-1 by t=3 with update magnitudes below 0.8×, so stopping early loses little. **Overshoot** (EAttn-FFN-Prll, Attn-FFN-Prll): rank *increases* at t=1−2 before recovering, and the two models fail differently. EAttn-FFN-Prll overshoots early, through an update of 2.93× the hidden-state norm in B9 (the subtraction $\boldsymbol{h} - \Pi(\mathbf{g})$ overshoots by nearly 3×). Attn-FFN-Prll barely moves in early blocks (0.005×) and instead makes a large late correction inside B11 (1.1× at t=5, Table 21). **Late construction** (RecGPT): enters B11 at rank 258 and improves monotonically (93→29→4.5), but t=3 reaches only 52.5% top-1; the prediction is half-formed and needs t=4–5 to converge. This explains the robustness partition of §3.2: in distributed-refinement models intermediate states are already useful, so stopping early loses little, while concentrated-construction models have not yet formed the prediction. Section 4 decomposes which component interactions drive the partition. Detailed rank-trajectory and KL-convergence figures are in Appendix L.

4 Factorial analysis and practical consequences

Section 3 traced the robustness partition to two computation modes, but attention alone does not explain it: EAttn-FFN-Prll collapses despite using energy attention. Here we decompose which component *interactions* determine the outcome, and show that this understanding translates to practical iteration scheduling.

4.1 Factorial decomposition

We organize our models as a factorial over three binary factors: attention (energy vs. standard), FFN (energy vs. SwiGLU), and layout (parallel vs. consecutive). Figure 4 shows the T=3 PPL degradation for each combination.

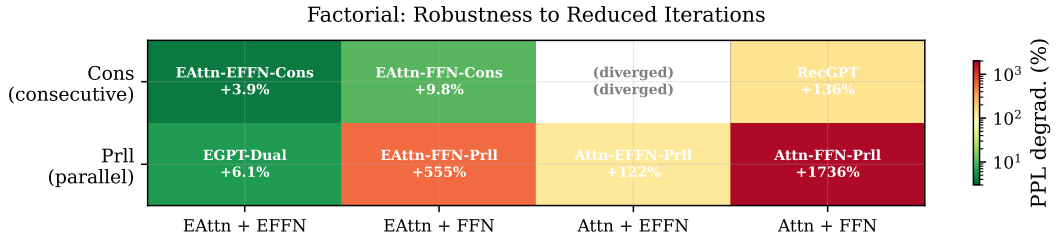


Figure 4: Factorial analysis: T=3 PPL degradation (%) for each combination of attention, FFN, and layout. Green = robust (<15%); red = collapse. Energy attention is necessary (the standard-attention columns on the right are never robust); within energy attention, either the consecutive layout or the energy FFN restores robustness.

Four findings emerge from Figure 4:

1. **Energy attention is necessary.** No model with standard attention is robust, regardless of FFN or update rule. The best standard attention model (Attn-EFFN-Prll) still degrades by **122%**. The remaining standard-attention cell (energy FFN, consecutive layout) has no result: its training loss diverged midway through training.

2. **Energy attention is not sufficient.** EAttn-FFN-Prll collapses at **+555%**, while EAttn-FFN-Cons is robust (**+9.8%**). The two differ *only* in layout (parallel with a learned preconditioner vs. consecutive).

3. **Two paths to robustness.** Within energy attention, robustness is achieved by either (a) switching to a consecutive layout (EAttn-FFN-Cons: **+9.8%**), which avoids the projection entirely; or (b) using the energy FFN (EGPT-Dual: **+6.1%**, EAttn-EFFN-Cons: **+3.9%**), which stabilizes regardless of the update rule.

4. **Energy FFN reduces collapse even with standard attention.** Attn-EFFN-Prll (softmax + energy FFN) degrades by 122%, **14× less** than Attn-FFN-Prll (softmax + SwiGLU) at 1736%, consistent with its low B11 entry rank (8.3, §3.3).

Practical iteration scheduling. A convergence-rate criterion $\rho_t = \|\boldsymbol{h}^{(t)} - \boldsymbol{h}^{(t-1)}\|/\|\boldsymbol{h}^{(t-1)}\| < 0.1$, calibrated on 100 C4 validation samples, saves **19%** of measured FLOPs on EGPT-Dual at 400M (**+0.5%** PPL), **17%** at 800M (**+0.2%** PPL), and **12.8%** at 1B (all benchmarks within $\pm$0.3pp of baseline). Under identical reduced schedules, the contrast with standard attention is stark. With every recurrent block capped at four iterations, RecGPT loses **11.8%** PPL at 400M and **17.0%** at 800M, while EAttn-FFN-Cons loses only **1.1%** (400M) and EGPT-Dual only **0.9%** (800M). At 1B under the same reduced schedule, RecGPT drops **3.4pp** on GSM8K-hash (18.3 to 14.9), while EGPT-Dual drops only 0.3pp. Standard-attention recurrent models can be halted modestly (e.g. a token-stability schedule saves ∼10% with negligible PPL change), but they degrade substantially under aggressive halting. Full halting results: Appendix P, O.

Replication across scales. Figure 5 compares EGPT-Dual and RecGPT under iteration reduction at 400M, 800M, and 1B (full numerical table in Appendix O). The robustness partition holds at all three scales and *widens* with scale: at T=1 RecGPT's PPL grows from 48× baseline (400M) to 53× (1B), while EGPT-Dual stays at 1.4–2.6×. Per-scale numerical breakdown is deferred to Appendix O.

5 How energy attention shapes the write space

Section 3 showed that energy-attention blocks keep intermediate states prediction-aligned. Here we identify an exact architectural constraint that separates energy attention from standard attention, show its empirical signature at all three scales, and state precisely what the energy view of the update does and does not explain.

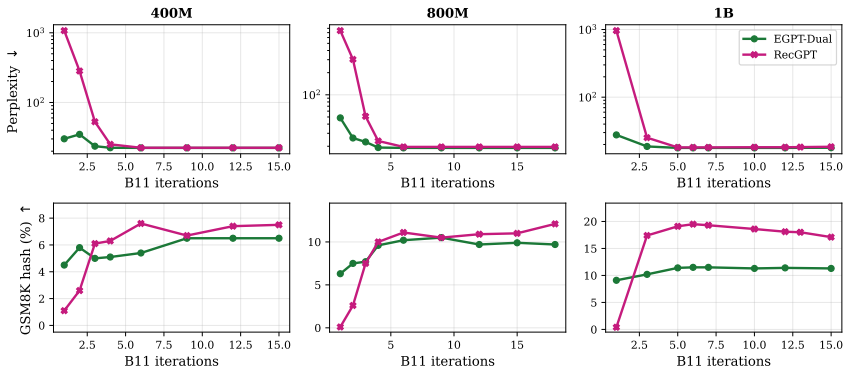


Figure 5: Robustness across scales: EGPT-Dual vs. RecGPT at T=1 (OOD), T=3 (OOD), and baseline (ID: T=9 at 400M/800M, T=6 at 1B). **Top**: perplexity (log scale). EGPT-Dual stays near baseline at all scales; RecGPT collapses to 700–1000+ at T=1. **Bottom**: GSM8K hash (%). EGPT-Dual retains 60–76% of baseline math reasoning at T=1; RecGPT drops to <2%. The gap widens with scale.

5.1 An exact constraint: value tying couples write space to read space

For one attention head, EAttn writes $\Delta \boldsymbol{h}_A^{\mathrm{EAttn}} = (\boldsymbol{\Pi}/\sqrt{d_h})\, J \sum_B a_{AB} \boldsymbol{g}_B$ with $J := \boldsymbol{W}_Q^\top \boldsymbol{W}_K$. Every update of a block therefore lies in $\mathrm{col}(\boldsymbol{\Pi} J)$, the image under $\boldsymbol{\Pi}$ of that block's own read space $R := \mathrm{row}(\boldsymbol{W}_Q)$. This holds exactly, by construction, in every energy-attention block. Standard attention writes through $\boldsymbol{W}_O \boldsymbol{W}_V$, with $\boldsymbol{W}_V$ independent of $(\boldsymbol{W}_Q, \boldsymbol{W}_K)$, so its write space has no architectural relation to what it reads (App. E.1). Intuitively, an energy-attention block can only move the state along directions it already attends in; it cannot open a separate write channel unrelated to its queries. We hypothesise that this constraint is what keeps intermediate states prediction-aligned, and test it through its weight-space signature. Isolating tying from the other changes energy attention makes (removing $\boldsymbol{W}_V$ and $\boldsymbol{W}_O$, routing through $\boldsymbol{\Pi}$) requires a targeted ablation, which we leave to future work.

Testing the constraint in the trained weights. The constraint above is architectural, so it only matters if training actually uses it. We check this directly in the weights. If an energy-attention block reads prediction-relevant directions and can only write into what it reads, its write operator should overlap the directions the LM head reads more than a random matrix would. A standard-attention block has no such constraint, so its write operator should look random. The constraint also suggests a second signature: each block's writes must match what the next block reads, so training should pull the write operators of successive blocks toward each other.

We test both predictions on the models of §3 (Table 4). For *alignment*, we take the top-k and bottom-k right-singular directions of the LM head $\boldsymbol{W}_U$, the directions it reads most and least strongly, and measure how much of the write operator lies in them, $\mathrm{align}(W, L_k) = \|\boldsymbol{P}_{L_k} W\|_F / \|W\|_F$, reported as excess over a random matrix ($\sqrt{k/d}$). For *sharing*, we report the cosine similarity between the write operators of different recurrent blocks. EGPT is the variant with a single preconditioner $\boldsymbol{\Pi}$ shared by attention and FFN. Definitions are in App. E.2.

Both predictions hold at every scale. **Alignment:** EGPT-Dual's write operator $\boldsymbol{\Pi}_a J$ overlaps the LM head well above chance at both ends of its spectrum, with excess $+\mathbf{0.13}$ to $+\mathbf{0.24}$ in the top-32 directions and $+\mathbf{0.34}$ to $+\mathbf{0.43}$ in the bottom-32, fading once the band widens to the top 512 (-0.08 to $+0.05$). RecGPT's write operator stays close to random ($\leq +0.09$ anywhere). Energy attention therefore writes into the directions the LM head reads most strongly, which adjust the dominant predictions, and into those it reads only weakly, where it can plausibly hold intermediate information with little effect on the output. **Sharing:** the write operators of different EGPT-Dual blocks have cosine 0.035 to 0.12, far above the chance level of about $1/d \approx 0.0005$ for matrices this size, while RecGPT's stay at chance (≤ 0.002). The energy FFN's write operator $\boldsymbol{\Pi}_f \boldsymbol{W}_2^\top \boldsymbol{W}_1$ is shared even more (0.13 to 0.26). Sharing weakens with scale, whereas the alignment does not.

Table 4: Cross-block write-operator sharing and LM-head alignment at three scales (same 8-GPT + 4-recurrent layout; iter$= [3, 4, 6, 9]$ at 400M/800M, iter$= [2, 3, 4, 6]$ at 1B; $\sim$126B tokens). EGPT-Dual uses separate $\mathbf{\Pi}_a, \mathbf{\Pi}_f$ per block; RecGPT uses standard softmax attention with independent $\boldsymbol{W}_V$. $\bar{\rho}^{\text{write}}$: cross-block cosim of the full write operator ($\mathbf{\Pi}_a J$ for EGPT-Dual, $\boldsymbol{W}_O\boldsymbol{W}_V$ for RecGPT). $\text{align}_{\pm}(k)$: write-operator energy excess over random $\sqrt{k/d}$ in the top-k or bottom-k right-singular subspace of $\boldsymbol{W}_U$. Full per-k breakdown in Table 9 (App. E.5).

				align_+ excess (top-k)		align_- excess (bot-k)	
Model	**Scale**	**Write op**	$\bar{\rho}^{\text{write}}$	k=512	k=32	k=512	k=32
EGPT	400M	$\mathbf{\Pi}J$	0.131	+0.083	**+0.291**	−0.010	+0.182
EGPT-Dual	400M	$\mathbf{\Pi}_a J$	0.121	−0.078	**+0.210**	**+0.156**	**+0.434**
RecGPT	400M	$\boldsymbol{W}_O\boldsymbol{W}_V$	−0.000	+0.029	+0.088	−0.001	+0.046
EGPT-Dual	800M	$\mathbf{\Pi}_a J$	0.075	+0.046	**+0.241**	**+0.184**	**+0.336**
RecGPT	800M	$\boldsymbol{W}_O\boldsymbol{W}_V$	0.000	+0.039	+0.053	+0.021	+0.070
EGPT-Dual	1B	$\mathbf{\Pi}_a J$	0.035	+0.002	**+0.127**	**+0.203**	**+0.355**
RecGPT	1B	$\boldsymbol{W}_O\boldsymbol{W}_V$	0.002	+0.068	+0.079	+0.011	+0.044

Taken together, trained energy-attention blocks show the signatures the write-space constraint predicts, and standard attention shows none of them; within each energy block the hidden state also moves steadily toward the readout directions (App. E.3). This is consistent with the prediction-aligned intermediate states of §3 and with the robustness to early stopping in §4.

Why the energy FFN and the consecutive layout matter: step size. The write-space constraint keeps the *direction* of each update aligned with the readout, but a discrete step can still overshoot, as EAttn-FFN-Prll's $2.93\times$ update shows (Table 3). The two paths to robustness in §4.1 both limit step size. The energy FFN's gate σ' is bounded (within roughly $[0, 1.1]$ for GELU), whereas SwiGLU's swish gate is not, so the SwiGLU product grows quadratically with its input while the energy FFN grows only linearly. The consecutive layout instead splits each iteration into two smaller residual steps. Either choice keeps update norms below $0.8\times$ in every robust model (Table 3).

What the energy view does and does not explain. The trained models do not minimise their energy: the certified-descent condition fails in every block (Table 5) and the measured energy rises across iterations (App. F), so they are energy-structured rather than energy-minimising. The energy view still guarantees that the normalised dynamics cannot diverge for any $\mathbf{\Pi}$ (Prop. 1), a property shared by any normalised recurrent block, including RecGPT. It also shows how an indefinite $\mathbf{\Pi}$ can still converge, through rotation from its antisymmetric part (App. B.11; full treatment in App. B).

6 Discussion and Conclusion

Our aim was to understand how energy components change recurrent computation, not to beat standard baselines at full iteration count. Energy-attention blocks refine the prediction at every iteration, while standard-attention blocks defer it to the final ones. Energy attention is necessary for robustness to fewer iterations, and either the energy FFN or the consecutive layout keeps step sizes small enough to preserve it: the robust models degrade by $+3.9\%$ to $+9.8\%$ PPL under iteration reduction, against $+122\%$ to $+1736\%$ for standard attention. We propose value tying, which confines each block's writes to its own read space, as the mechanism, and find that the trained models are energy-structured rather than energy-minimising.

Limitations. EGPT-Dual lags on GSM8K at 1B (12.0 vs. RecGPT 18.3), a gap absent at 400M/800M that may reflect undertraining or the write-space constraint. LogitLens measures readout decodability rather than computational progress, but iteration reduction confirms the two modes functionally. Isolating value tying with a targeted ablation is future work.

AI USE STATEMENT

We used generative AI tools to help edit and restructure the manuscript, including rephrasing text, checking that numbers reported in the text match the tables, part of the implementation and formatting references. Reference metadata was verified against arXiv. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat any dataset, support qualitative or thematic data analysis, or interpret results. All AI-assisted text or artifacts produced were reviewed by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies language-model architectures of up to 1B parameters trained on publicly available web and mathematics corpora (Nemotron-CC, MegaMath-Web-Pro, FineMath). It involves no human subjects and no data beyond these public datasets, and we do not foresee risks specific to this work beyond those of language models in general.

REPRODUCIBILITY STATEMENT

All models are trained from scratch with the same data mixture and hyperparameters (App. I, Table 15). Architectures and update rules are defined in Section 2 and Table 1, with per-scale sizes, iteration schedules and parameter counts in Tables 13 and 14. Benchmarks and metrics are listed in App. J. The LogitLens protocol is described in Section 3.3, with full per-iteration tables in App. K; halting criteria and their calibration are in App. P; FLOPs are measured with `torch.utils.flop_counter`. Proofs of all propositions are in App. B.

REFERENCES

Ralph Abraham, Jerrold E. Marsden, and Tudor Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75 of *Applied Mathematical Sciences*. Springer, New York, 2 edition, 1988. ISBN 978-0-387-96790-5. doi: 10.1007/978-1-4612-1029-0.

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, and Se-Young Yun. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. In *International Conference on Machine Learning Workshop on Automated Machine Learning*, 2021.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Nima Dehmamy, Benjamin Hoover, Bishwajit Saha, Leo Kozachkov, Jean-Jacques Slotine, and Dmitry Krotov. Nrgpt: An energy-based alternative for gpt. In *International Conference on Learning Representations*, 2026.

Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. Looped transformers for length generalization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 11398–11442. PMLR, 2023.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J Zaki, and Dmitry Krotov. Energy transformer. In *International Conference on Learning Representations*, 2024.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory. *arXiv preprint arXiv:2507.06211*, 2025.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations, 2017.

Yan Liu, Renren Jin, Ling Shi, Zheng Yao, and Deyi Xiong. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models, 2024. URL https://arxiv.org/abs/2403.07747.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.

NVIDIA, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, Ann Guan, Anna Shors, Anubhav Mandarwal, Arham Mehta, Arun Venkatesan, Ashton Sharabiani, Ashwath Aithal, Ashwin Poojary, Ayush Dattagupta, Balaram Buddharaju, Banghua Zhu, Barnaby Simkin, Bilal Kartal, Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Brian Yu, and Bryan Catanzaro. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016.

Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models. *arXiv preprint arXiv:2604.12946*, 2026.

Hubert Ramsauer, Bernhard Schafl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, 2022.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P. Xing. Megamath: Pushing the limits of open math corpora, 2025. URL https://arxiv.org/abs/2504.02807.

Rui-Jie Zhu et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

APPENDIX OVERVIEW

The appendix is organised in four groups. Each section is referenced from the main text where it is used.

Theory of the energy components. App. A derives the Gaussian-mixture and vMF readings of the attention energy and states the gradient convention. App. B gives the sphere-completeness result, the Lyapunov conditions and their failure in trained models, and the local stability argument. App. C derives the energy FFN and compares it with SwiGLU, and App. D relates the dual preconditioner to a Helmholtz decomposition.

Mechanism diagnostics. App. E defines the read and write spaces and the alignment and sharing measures behind Table 4, and reports them at all three scales. App. F decomposes the learned preconditioner and shows that the energy rises across iterations. App. G tests an explicit tangent (Rayleigh) projection at 160M.

Setup and full-iteration quality. App. H motivates the hybrid design with a 4-block study. App. I gives architectures, parameter counts and training hyperparameters, and App. J lists the benchmarks and the full 400M results.

Computation modes, robustness and halting. App. K and App. L give the full LogitLens tables and figures. App. M and App. N report perplexity and benchmarks under iteration reduction for every 400M model, and App. O does the same at 1B. App. P gives the halting criteria and results, App. Q compares uniform and heterogeneous schedules, App. R reduces iterations in each block and in all blocks jointly, and App. S reports seed variance. Extended related work is in App. T.

A INTERPRETATIONS OF THE ATTENTION ENERGY

This appendix derives the GMM and vMF interpretations referenced in the main text, including the precise role of the singular spectrum of $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$. We work in the post-RMSNorm coordinates $\boldsymbol{g}_A = \mathrm{RN}(\boldsymbol{h}_A) \in \sqrt{d}\, S^{d-1}$ and use the SVD $\boldsymbol{W}_Q^\top \boldsymbol{W}_K = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ are orthogonal and $\boldsymbol{\Sigma} \succeq 0$ is diagonal with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$. SVD with $\boldsymbol{\Sigma} \succeq 0$ is canonical: any sign that would arise in an asymmetric eigendecomposition is absorbed into the orthogonal factors $\boldsymbol{U}, \boldsymbol{V}$.

A.1 ATTENTION ENERGY AS A GAUSSIAN-MIXTURE LOG-LIKELIHOOD

The attention kernel hides a Gaussian. Applying the polarization identity $\boldsymbol{Q}_A^\top \boldsymbol{K}_B = \frac{1}{2}(\|\boldsymbol{Q}_A\|^2 + \|\boldsymbol{K}_B\|^2 - \|\boldsymbol{Q}_A - \boldsymbol{K}_B\|^2)$, the $\|\boldsymbol{Q}_A\|^2$ piece factors out of the LSE over B as a per-query constant and the $\|\boldsymbol{K}_B\|^2$ piece becomes a per-key mixing weight $\pi_B \propto \exp[\|\boldsymbol{K}_B\|^2/(2\sqrt{d_h})]$, giving

$$E_A^{\mathrm{attn}}(\boldsymbol{h}) \;=\; -\log \sum_{B<A} \pi_B\, \mathcal{N}\big(\boldsymbol{Q}_A;\, \boldsymbol{K}_B,\, \sqrt{d_h}\, \boldsymbol{I}\big) \;-\; \tfrac{\|\boldsymbol{Q}_A\|^2}{2\sqrt{d_h}}, \tag{4}$$

a Gaussian-mixture negative log-likelihood of $\boldsymbol{Q}_A$ over the *discrete set of observed* past keys $\{\boldsymbol{K}_B\}_{B<A}$ (not a continuous density), plus a per-query quadratic $-\|\boldsymbol{Q}_A\|^2/(2\sqrt{d_h}) = -\boldsymbol{g}_A^\top \boldsymbol{W}_Q^\top \boldsymbol{W}_Q \boldsymbol{g}_A/(2\sqrt{d_h})$. This quadratic has the same algebraic shape as the FFN energy $E^{\mathrm{FF}} = -\boldsymbol{g}^\top \boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{g})$ (identity activation, tied weights); we absorb it into the FFN channel under the assumption that the FFN can realize it (otherwise the absorption changes the gradient). EAttn is then an interpretive lens on the update as a non-parametric log-likelihood step: previous tokens are the data, the current query is the test point, and each iteration improves the fit. Restricted to the post-RMSNorm sphere, the attention weights $p(B \mid A)$ take a von Mises–Fisher form over the observed keys, with mean direction and concentration set by the singular structure of $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$ (derived below).

Dense Associative Memories. A different interpretation relates energy attention to Dense Associative Memories (Krotov & Hopfield, 2016). In this setting the query serves as a state vector that evolves in time according to energy descent dynamics on the energy landscape describing a collection memory vectors. Each key corresponds to a memory. The state vector is attracted to each memory, and the strength of attraction defines attention scores (Krotov et al., 2025).

A.2 GRADIENT CONVENTION

Hoover et al. (2024) and Dehmamy et al. (2026) both define the update rule as $\dot{\boldsymbol{h}} = -\boldsymbol{\Pi}\nabla_{\boldsymbol{g}} E$, gradient w.r.t. $\boldsymbol{g} = \mathrm{RN}(\boldsymbol{h})$, not $\boldsymbol{h}$. However, we find that with RMSNorm, the proper gradient w.r.t. $\boldsymbol{h}$ performs just as well, if not better than $\boldsymbol{g}$, resulting in an update $\dot{\boldsymbol{h}} = -\boldsymbol{\Pi}\nabla_{\boldsymbol{h}} E$. The key difference is that this new update rule, related Rayleigh Quotients, $\hat{\boldsymbol{h}} = -\boldsymbol{\Pi}\nabla_{\boldsymbol{h}} E$ contains the Jacobian $\nabla\mathrm{RN} = I - \hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^T$, explicitly forcing the dynamics to stay on the hypersphere. For brevity, we use the old update rule above.

A.3 SVD-ALIGNED COORDINATES

Define $\tilde{\boldsymbol{g}}_A := \boldsymbol{U}^\top \boldsymbol{g}_A$ and $\tilde{\boldsymbol{k}}_B := \boldsymbol{V}^\top \boldsymbol{g}_B$. Both lie on $\sqrt{d}\, S^{d-1}$ since $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal. The attention logit becomes

$$\boldsymbol{Q}_A^\top \boldsymbol{K}_B \;=\; \boldsymbol{g}_A^\top\, \boldsymbol{W}_Q^\top \boldsymbol{W}_K\, \boldsymbol{g}_B \;=\; \tilde{\boldsymbol{g}}_A^\top\, \boldsymbol{\Sigma}\, \tilde{\boldsymbol{k}}_B \;=\; (\boldsymbol{S}\tilde{\boldsymbol{g}}_A)^\top (\boldsymbol{S}\tilde{\boldsymbol{k}}_B), \qquad \boldsymbol{S} := \boldsymbol{\Sigma}^{1/2}. \tag{5}$$

The vectors $\boldsymbol{S}\tilde{\boldsymbol{g}}_A$ and $\boldsymbol{S}\tilde{\boldsymbol{k}}_B$ live on an ellipsoid with semi-axes $\sqrt{d}\,\sigma_i^{1/2}$. Polarization gives

$$2\,(\boldsymbol{S}\tilde{\boldsymbol{g}}_A)^\top (\boldsymbol{S}\tilde{\boldsymbol{k}}_B) \;=\; \|\boldsymbol{S}\tilde{\boldsymbol{g}}_A\|^2 + \|\boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2 - \|\boldsymbol{S}\tilde{\boldsymbol{g}}_A - \boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2, \tag{6}$$

so the kernel decomposes as a Mahalanobis Gaussian times boundary terms:

$$\exp\Big[\tfrac{\boldsymbol{Q}_A^\top \boldsymbol{K}_B}{\sqrt{d_h}}\Big] \;=\; \exp\Big[-\tfrac{\|\boldsymbol{S}\tilde{\boldsymbol{g}}_A - \boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2}{2\sqrt{d_h}}\Big] \cdot \exp\Big[\tfrac{\|\boldsymbol{S}\tilde{\boldsymbol{g}}_A\|^2 + \|\boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2}{2\sqrt{d_h}}\Big]. \tag{7}$$

A.4 Energy decomposition

Substituting equation 7 into the LSE, the $\|\boldsymbol{S}\tilde{\boldsymbol{g}}_A\|^2$ piece is B-independent and factors out:

$$E_A^{\text{attn}}(\boldsymbol{h}) \;=\; \underbrace{-\,\tfrac{\|\boldsymbol{S}\tilde{\boldsymbol{g}}_A\|^2}{2\sqrt{d_h}}}_{\text{Bingham}} \;-\; \log \sum_{B<A} \exp\Big[-\tfrac{\|\boldsymbol{S}\tilde{\boldsymbol{g}}_A - \boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2}{2\sqrt{d_h}} \;+\; \tfrac{\|\boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2}{2\sqrt{d_h}}\Big]. \tag{8}$$

The Bingham term is a quadratic on the sphere depending only on the query and the singular structure of $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$; it satisfies $\sigma_d\, d/(2\sqrt{d_h}) \le \|\boldsymbol{S}\tilde{\boldsymbol{g}}_A\|^2/(2\sqrt{d_h}) \le \sigma_1\, d/(2\sqrt{d_h})$ since $\|\tilde{\boldsymbol{g}}_A\|^2 = d$. The remaining LSE is a Mahalanobis-Gaussian mixture log-likelihood with key-dependent prior weights $\pi_B \propto \exp[\|\boldsymbol{S}\tilde{\boldsymbol{k}}_B\|^2/(2\sqrt{d_h})]$.

A.5 The attention weights take a vMF form

The conditional probability of attending to key B is $p(B \mid A) \propto \exp[\boldsymbol{Q}_A^\top \boldsymbol{K}_B/\sqrt{d_h}]$. The Bingham term in $\tilde{\boldsymbol{g}}_A$ is constant in B and cancels, leaving

$$p(B \mid A) \;\propto\; \exp\Big[\tfrac{1}{\sqrt{d_h}}\, \tilde{\boldsymbol{g}}_A^\top \boldsymbol{\Sigma}\, \tilde{\boldsymbol{k}}_B\Big] \;=\; \exp\Big[\boldsymbol{c}_A^\top\, \tilde{\boldsymbol{k}}_B\Big], \qquad \boldsymbol{c}_A := \tfrac{\boldsymbol{\Sigma}\tilde{\boldsymbol{g}}_A}{\sqrt{d_h}}. \tag{9}$$

This is a von Mises–Fisher kernel on $\sqrt{d}\, S^{d-1}$, evaluated at the discrete set of observed keys (not a continuous density), with mean direction $\boldsymbol{c}_A/\|\boldsymbol{c}_A\|$ and concentration

$$\kappa_A \;=\; \sqrt{d}\,\|\boldsymbol{c}_A\| \;=\; \|\boldsymbol{\Sigma}\,\tilde{\boldsymbol{g}}_A\|\,\sqrt{d/d_h} \;\in\; \big[\,\sigma_d\, d/\sqrt{d_h},\;\; \sigma_1\, d/\sqrt{d_h}\,\big]. \tag{10}$$

Both the mean direction and the concentration depend on A through how $\tilde{\boldsymbol{g}}_A$ aligns with the singular vectors of $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$: queries aligned with large-singular-value directions produce sharper attention, queries aligned with small-singular-value directions produce flatter attention. The weights keep this vMF form for arbitrary singular spectrum — anisotropy is absorbed into the parameters $(\boldsymbol{c}_A, \kappa_A)$, not into the functional form.

Anisotropic case. For general $\boldsymbol{\Sigma}$, the weights keep the vMF form equation 10 and the energy retains the Bingham quadratic. The Bingham term is invisible in the attention *probabilities* (it cancels in the conditional) but visible in the energy and its gradient: it contributes a query-dependent bias to the EAttn update that pulls $\tilde{\boldsymbol{g}}_A$ toward the dominant singular directions of $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$, independent of which keys are present. The strength of this bias is controlled by the conditioning σ_1/σ_d.

B Sphere stability and the Lyapunov candidate: full treatment

This appendix expands on §5.1. An explicit Rayleigh projection that enforces the tangent form of the update trains well at 160M (App. G). It covers the comparison with Parcae's spectral framework, the formal Sphere Completeness proposition with proof, the Lyapunov-single and Lyapunov-dual statements with their PSD coupling conditions, the discrete-time consequences linking the continuous theory to the empirical $2.93\times$ overshoot, and the structural hierarchy that organises the factorial.

B.1 Why energy iteration is stable on the LayerNorm sphere

The LTI analysis of Prairie et al. (2026) treats the recurrent block as $\boldsymbol{h}_{t+1} = \bar{\boldsymbol{A}}\boldsymbol{h}_t + \bar{\boldsymbol{B}}\boldsymbol{e} + \bar{\mathcal{R}}(\boldsymbol{h}_t, \boldsymbol{e})$, linearises by dropping $\bar{\mathcal{R}}$, and identifies large *spectral radius* $\rho(\bar{\boldsymbol{A}}) := \max_i |\lambda_i(\bar{\boldsymbol{A}})|$ of the linear injection as the source of residual stream explosion in prior recurrent architectures. They restore stability by parameterising $\boldsymbol{A}$ as a negative diagonal matrix, forcing $\rho(\bar{\boldsymbol{A}}) < 1$.

Our learned $\boldsymbol{\Pi}$ contains substantial antisymmetric mass and is not sign-definite ($\|\mathbf{A}\|_F/\|\mathbf{S}\|_F \approx 1$, App. F). The corresponding linearised operator $\bar{\boldsymbol{A}}_{\text{EGPT}} = \boldsymbol{I} - \boldsymbol{\Pi}_a \boldsymbol{M}_a - \lambda \boldsymbol{\Pi}_f \boldsymbol{M}_f$ need not satisfy $\rho < 1$, yet EGPT-Dual trains and infers stably. The mechanism is in the term Parcae discards: $\bar{\mathcal{R}}$ contains the LayerNorm that every energy module reads, putting the dynamics on a compact sphere on which divergence is geometrically impossible.

Geometric setup. With $\boldsymbol{g} = \boldsymbol{h}/\mathrm{RMS}(\boldsymbol{h}) \in \sqrt{d}\, S^{d-1}$ and Rayleigh projector $\boldsymbol{P_g} = \boldsymbol{I} - \frac{1}{d}\boldsymbol{g}\boldsymbol{g}^\top$ (which is also the Jacobian of $\boldsymbol{h} \mapsto \boldsymbol{g}$ up to scale), the chain rule gives the Riemannian gradient $\mathrm{grad}_{S^{d-1}} E(\boldsymbol{g}) = \boldsymbol{P_g} \nabla_{\boldsymbol{g}} E(\boldsymbol{g})$ identically. Define $\widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}} := \boldsymbol{P_g} \boldsymbol{\Pi} \boldsymbol{P_g}$; the EGPT update $-\widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E$ is a tangent vector field on the sphere for any $\boldsymbol{\Pi}$ — a *free tangency* that the architecture acquires by composing gradient with Rayleigh projector.

Proposition 1 (Sphere completeness)**.** *Let $E : \sqrt{d}\, S^{d-1} \to \mathbb{R}$ be C^2 (so that the vector field below is C^1, hence locally Lipschitz) and let $\boldsymbol{\Pi} \in \mathbb{R}^{d \times d}$ be arbitrary, with no spectral, symmetry, or sign constraint. The flow $\dot{\boldsymbol{g}} = -\widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E(\boldsymbol{g})$ on $\sqrt{d}\, S^{d-1}$ is complete for all $t \in \mathbb{R}$, and every trajectory has a non-empty compact ω-limit set. No condition on $\rho(\boldsymbol{\Pi})$ or $\|\boldsymbol{\Pi}\|_2$ is required for non-divergence.*

The proof is one paragraph: a C^1 tangent vector field on a compact manifold has unique, globally defined integral curves, and forward trajectories of bounded flows have non-empty ω-limit sets by sequential compactness (full proof below). Proposition 1 is the geometric counterpart of Parcae's spectral condition: where Parcae constrains the linear injection so that $\mathbb{R}^d$-residuals cannot diverge, we observe that the LayerNorm image is already a compact submanifold on which divergence is geometrically impossible.

Discrete-time consequences. The propositions are continuous-time. The forward pass executes explicit Euler steps; tracking the flow requires $\|\widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V\| \ll \|\boldsymbol{g}\|$. EFFN bounds $\|\nabla_{\boldsymbol{g}} E^{\mathrm{FF}}\|$ via the saturation of $\sigma(\boldsymbol{W}_1 \boldsymbol{g})$, controlling per-step displacement; SwiGLU has no such bound, and an indefinite $\boldsymbol{\Pi}$ can then amplify the update. This is the discrete-time counterpart of the $2.93\times$ overshoot in EAttn-FFN-Prll (Table 3): EFFN, or the consecutive layout (two smaller steps with independent residuals), supplies the discrete-step-size budget the continuous theory does not. Sphere completeness itself depends only on a compact normalisation image, not LayerNorm specifically: a 160M EGPT-Dual trained with RMSNorm + explicit Rayleigh projection $\boldsymbol{P_g} = \boldsymbol{I} - \frac{1}{d}\boldsymbol{g}\boldsymbol{g}^\top$ on the update matches LayerNorm to within seed variance (App. G).

B.2 Vector fields on compact manifolds

We recall the standard result (see, e.g., Abraham et al., 1988, §4.1) that a C^1 tangent vector field on a compact smooth manifold M is *complete* — its integral curves $\gamma : \mathbb{R} \to M$ are defined for all $t \in \mathbb{R}$. The argument is geometric: M is closed and bounded as a subset of any embedding space, so a trajectory that fails to extend to $t = t^*$ would have to exit every compact subset of M, contradicting compactness of M itself. The same argument shows that ω-limit sets are non-empty: the forward trajectory $\{\gamma(t) : t \geq 0\}$ has accumulation points in M by sequential compactness, and these accumulation points form a non-empty closed invariant set.

These facts apply unchanged when the vector field has indefinite divergence or contains rotational components — they require only C^1 regularity and tangency, not any form of dissipativity.

B.3 Proof of Proposition 1

Proof. Set $X(\boldsymbol{g}) := -\widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E(\boldsymbol{g}) = -\boldsymbol{P_g}\, \boldsymbol{\Pi}\, \boldsymbol{P_g}\, \nabla_{\boldsymbol{g}} E(\boldsymbol{g})$.

Tangency. For any $\boldsymbol{u} \in \mathbb{R}^d$, $\boldsymbol{P_g}\boldsymbol{u} \in T_{\boldsymbol{g}} S^{d-1}$ since $\boldsymbol{g}^\top \boldsymbol{P_g}\boldsymbol{u} = \boldsymbol{g}^\top(\boldsymbol{u} - \frac{1}{d}\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{u}) = \boldsymbol{g}^\top \boldsymbol{u} - \frac{1}{d}(\boldsymbol{g}^\top \boldsymbol{g})(\boldsymbol{g}^\top \boldsymbol{u}) = 0$ using $\boldsymbol{g}^\top \boldsymbol{g} = d$. Therefore $X(\boldsymbol{g}) \in T_{\boldsymbol{g}} S^{d-1}$.

Regularity. $\nabla_{\boldsymbol{g}} E$ is C^1 because E is C^2; $\boldsymbol{g} \mapsto \boldsymbol{P_g}$ is smooth (its entries are polynomials in the entries of $\boldsymbol{g}$); $\boldsymbol{\Pi}$ is constant. Composition and matrix-vector product preserve C^1 regularity, so X is C^1, hence locally Lipschitz, on $\sqrt{d}\, S^{d-1}$. The softmax attention energy and the energy FFN with a smooth activation (e.g. GELU) satisfy this.

Completeness and ω-limits. $\sqrt{d}\, S^{d-1}$ is compact. By the standard theorem on continuous vector fields on compact manifolds (stated above), integral curves of X are defined for all $t \in \mathbb{R}$ and forward trajectories have non-empty ω-limit sets.

No part of the argument uses any property of $\boldsymbol{\Pi}$ beyond its being a fixed real matrix; in particular $\rho(\boldsymbol{\Pi})$, $\|\boldsymbol{\Pi}\|_2$, and the symmetry/skewness of $\boldsymbol{\Pi}$ are unconstrained. □

B.4 Lyapunov function for single-Pi EGPT

Proposition 2 (Lyapunov function for single-$\mathbf{\Pi}$ EGPT)**.** *Let $V(\boldsymbol{g}) := E^{\mathrm{Attn}}(\boldsymbol{g}) + \lambda E^{\mathrm{FF}}(\boldsymbol{g})$ and consider the single-$\mathbf{\Pi}$ flow $\dot{\boldsymbol{g}} = -\widetilde{\mathbf{\Pi}}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V$. For any $\mathbf{\Pi}$ with symmetric part $\mathbf{\Pi}^s := \frac{1}{2}(\mathbf{\Pi} + \mathbf{\Pi}^\top) \succeq 0$,*

$$\dot{V}(\boldsymbol{g}(t)) \;=\; -(\boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V)^\top \mathbf{\Pi}^s (\boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V) \;\leq\; 0, \tag{11}$$

with equality iff $\boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V \in \ker(\mathbf{\Pi}^s)$. The antisymmetric part $\mathbf{\Pi}^A := \frac{1}{2}(\mathbf{\Pi} - \mathbf{\Pi}^\top)$ contributes zero to $\dot{V}$.

Proof. $\dot{V} = \nabla_{\boldsymbol{g}} V^\top \dot{\boldsymbol{g}} = -\nabla_{\boldsymbol{g}} V^\top \boldsymbol{P}_{\boldsymbol{g}} \mathbf{\Pi} \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V = -w^\top \mathbf{\Pi} w$ for $w := \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} V$. For any matrix $\mathbf{\Pi}$ and any vector w, $w^\top \mathbf{\Pi} w = w^\top \mathbf{\Pi}^s w$ (antisymmetric quadratic forms vanish identically), so $\dot{V} = -w^\top \mathbf{\Pi}^s w \leq 0$ when $\mathbf{\Pi}^s \succeq 0$. The equality case follows. □

B.5 Lyapunov function for EGPT-Dual

Proposition 3 (Lyapunov function for EGPT-Dual)**.** *For the two-channel EGPT-Dual flow $\dot{\boldsymbol{g}} = -\widetilde{\mathbf{\Pi}}_{a\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{Attn}} - \lambda \widetilde{\mathbf{\Pi}}_{f\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{FF}}$, with V as in Proposition 2,*

$$\dot{V} \;=\; -\boldsymbol{u}^\top \mathbf{\Pi}_a^s \boldsymbol{u} \;-\; \boldsymbol{v}^\top \mathbf{\Pi}_f^s \boldsymbol{v} \;-\; \boldsymbol{u}^\top (\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top) \boldsymbol{v}, \tag{12}$$

where $\boldsymbol{u} := \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{Attn}}$ and $\boldsymbol{v} := \lambda \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{FF}}$. V is a Lyapunov function whenever the block matrix $\left(\begin{smallmatrix} \mathbf{\Pi}_a^s & \frac{1}{2}(\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top) \\ \frac{1}{2}(\mathbf{\Pi}_a + \mathbf{\Pi}_f^\top) & \mathbf{\Pi}_f^s \end{smallmatrix}\right)$ is PSD — in particular when $\mathbf{\Pi}_a^s, \mathbf{\Pi}_f^s \succeq c\boldsymbol{I}$ and $\|\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top\| < 2c$.

The within-channel terms $\boldsymbol{u}^\top \mathbf{\Pi}_a^s \boldsymbol{u}$ and $\boldsymbol{v}^\top \mathbf{\Pi}_f^s \boldsymbol{v}$ inherit the same antisymmetric-is-free structure as in the single-$\mathbf{\Pi}$ case. The cross-channel term $\boldsymbol{u}^\top (\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top) \boldsymbol{v}$ is the price of two distinct preconditioners, with indefinite sign in general; the spectral condition above bounds it. EGPT-Dual is therefore not full Helmholtz, its conservative part being restricted to the form $\mathbf{\Pi}_i^A \nabla E^i$ rather than an arbitrary divergence-free field — a restriction that preserves a tractable Lyapunov criterion.

B.6 Proof of Proposition 3

Proof. With $\boldsymbol{u} := \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{Attn}}$ and $\boldsymbol{v} := \lambda \boldsymbol{P}_{\boldsymbol{g}} \nabla_{\boldsymbol{g}} E^{\mathrm{FF}}$, the time derivative of $V = E^{\mathrm{Attn}} + \lambda E^{\mathrm{FF}}$ along the flow is

$$\begin{aligned}
\dot{V} &= \nabla_{\boldsymbol{g}} V^\top \dot{\boldsymbol{g}} = -(\nabla_{\boldsymbol{g}} E^{\mathrm{Attn}} + \lambda \nabla_{\boldsymbol{g}} E^{\mathrm{FF}})^\top \boldsymbol{P}_{\boldsymbol{g}} \big[\mathbf{\Pi}_a \boldsymbol{u} + \mathbf{\Pi}_f \boldsymbol{v}\big] \\
&= -(\boldsymbol{u} + \boldsymbol{v})^\top (\mathbf{\Pi}_a \boldsymbol{u} + \mathbf{\Pi}_f \boldsymbol{v}) \\
&= -\boldsymbol{u}^\top \mathbf{\Pi}_a \boldsymbol{u} - \boldsymbol{u}^\top \mathbf{\Pi}_f \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{\Pi}_a \boldsymbol{u} - \boldsymbol{v}^\top \mathbf{\Pi}_f \boldsymbol{v} \\
&= -\boldsymbol{u}^\top \mathbf{\Pi}_a^s \boldsymbol{u} - \boldsymbol{v}^\top \mathbf{\Pi}_f^s \boldsymbol{v} - \boldsymbol{u}^\top (\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top) \boldsymbol{v}.
\end{aligned}$$

The first two terms used $w^\top \mathbf{\Pi} w = w^\top \mathbf{\Pi}^s w$ for any matrix $\mathbf{\Pi}$; the cross term combined $\boldsymbol{u}^\top \mathbf{\Pi}_f \boldsymbol{v}$ and $\boldsymbol{v}^\top \mathbf{\Pi}_a \boldsymbol{u} = \boldsymbol{u}^\top \mathbf{\Pi}_a^\top \boldsymbol{v}$. Writing $\boldsymbol{w} := (\boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top$,

$$-\dot{V} = \boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{w}, \qquad \boldsymbol{M} = \begin{pmatrix} \mathbf{\Pi}_a^s & \frac{1}{2}(\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top) \\ \frac{1}{2}(\mathbf{\Pi}_a + \mathbf{\Pi}_f^\top) & \mathbf{\Pi}_f^s \end{pmatrix},$$

and V is Lyapunov iff $\boldsymbol{M} \succeq 0$. By Schur complement, $\boldsymbol{M} \succeq 0$ when $\mathbf{\Pi}_a^s, \mathbf{\Pi}_f^s \succeq c\boldsymbol{I}$ for some $c > 0$ and $\|\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top\| < 2c$. □

B.7 Lyapunov diagnostic on trained models

The condition of Proposition 3 is checkable on trained models. For each block we compute the symmetrized block matrix $\boldsymbol{M}_{\mathrm{sym}}$ and report its smallest eigenvalue $\lambda_{\min}(\boldsymbol{M}_{\mathrm{sym}})$. Negative values indicate directions in which V may transiently increase along the flow.

All $\lambda_{\min}$ values are negative at every scale, confirming that $V = E^{\mathrm{Attn}} + \lambda E^{\mathrm{FF}}$ is *not* a strict Lyapunov function for the trained dynamics. This is consistent with the Helmholtz decomposition

Table 5: Lyapunov diagnostic: $\lambda_{\min}(\boldsymbol{M}_{\text{sym}})$ per recurrent block. All values are negative, indicating the strict Lyapunov condition (Prop. 3) does not hold.

Model	B8	B9	B10	B11
EGPT-Dual 400M	−15.07	−12.37	−9.72	−8.97
EGPT-Dual 800M	−20.47	−18.32	−15.85	−12.58
EGPT-Dual 1B	−18.80	−21.45	−20.05	−16.48

(Table 10) showing the energy increases by $+5$ to $+11$ per token per iteration. At 400M and 800M, $|\lambda_{\min}|$ decreases monotonically from B8 to B11 (the block with the most iterations), suggesting that longer iteration chains induce projections closer to the positive-definite boundary. V is therefore *not* a Lyapunov function for the trained dynamics: the certified-descent condition (Prop. 3) fails on every block (Table 5), and the realized energy drift is positive and bounded (Table 10), with magnitude consistent with $|\lambda_{\min}|$. This is an honest negative result for *global monotone descent*, not a contradiction with stability. What the dynamics rests on instead is (i) the unconditional non-divergence of Prop. 1 (compactness, no condition on $\mathbf{\Pi}$), and (ii) a *local* spectral criterion at the states the model actually reaches (App. B.11): the linearisation of the flow at an attractor is $-\mathbf{\Pi}_T \boldsymbol{H}_T$, whose spectrum—not $\lambda_{\min}(\boldsymbol{M}_{\text{sym}})$—governs convergence, and which can be stable even when $\boldsymbol{M}_{\text{sym}}$ is indefinite (*gyroscopic stabilisation*). The indefinite regime genuinely admits limit cycles (explicit in-class construction, App. B.10), so the sign condition was never vacuous; trained models simply do not operate in the cycle regime.

B.8 Lyapunov diagnostic: measurement protocol

The Lyapunov condition of Proposition 3 can be checked directly on a trained EGPT-Dual checkpoint. For each looped block $\ell \in \{B8, B9, B10, B11\}$, extract the learned projections $\mathbf{\Pi}_a^{(\ell)}, \mathbf{\Pi}_f^{(\ell)} \in \mathbb{R}^{d\times d}$ and form

$$\mathbf{\Pi}_a^s = \tfrac{1}{2}(\mathbf{\Pi}_a + \mathbf{\Pi}_a^\top), \quad \mathbf{\Pi}_f^s = \tfrac{1}{2}(\mathbf{\Pi}_f + \mathbf{\Pi}_f^\top), \quad \mathbf{\Pi}_{\text{cross}} = \tfrac{1}{2}(\mathbf{\Pi}_f + \mathbf{\Pi}_a^\top), \tag{13}$$

and assemble the $2d \times 2d$ symmetrized block matrix

$$\boldsymbol{M}_{\text{sym}}^{(\ell)} = \begin{pmatrix} \mathbf{\Pi}_a^s & \mathbf{\Pi}_{\text{cross}} \\ \mathbf{\Pi}_{\text{cross}}^\top & \mathbf{\Pi}_f^s \end{pmatrix}. \tag{14}$$

Compute and report the smallest eigenvalue $\lambda_{\min}\big(\boldsymbol{M}_{\text{sym}}^{(\ell)}\big)$ per block. By Proposition 3, V is a strict Lyapunov function whenever $\lambda_{\min} > 0$ on every block; negative values quantify the worst-case rate at which V can transiently increase along the cross-channel coupling direction.

Cross-reference with Helmholtz decomposition. Table 5 shows all blocks have $\lambda_{\min} < 0$, predicting that V can increase along the cross-channel coupling direction. Table 10 (App. F) confirms this empirically: the energy drift is $+5$ to $+11$ per token per iteration. The magnitude of $\lambda_{\min}$ ($\approx$ 9–21) is consistent with the observed drift being bounded but non-negligible.

B.9 When the preconditioner is indefinite

When $\mathbf{\Pi}^s$ is indefinite (Section 5 of the main text), V is not monotone along trajectories and the LaSalle invariance principle does *not* apply: LaSalle requires a function that is non-increasing along the flow, which is exactly what indefiniteness destroys, and the accumulation set need not lie in $\{\boldsymbol{g} : \dot{E} = 0\}$ in the invariant-set sense. What remains unconditionally is Birkhoff's theorem (Prop. 1): ω-limit sets are non-empty, compact, connected, and invariant. These can be equilibria or, in principle, limit cycles—an explicit in-class construction with indefinite $\mathbf{\Pi}^s$ exhibiting a hyperbolically attracting cycle is given in App. B.10, so the sign condition is genuinely not vacuous. We do not observe cycles in trained models; App. B.11 gives the local spectral criterion ($\rho(D\mathbf{F}|_T) < 1$) that certifies convergence to equilibria at the states the models actually reach, and explains why an indefinite $\boldsymbol{M}_{\text{sym}}$ (Table 5) is compatible with the observed convergence.

B.10 AN IN-CLASS ATTRACTING LIMIT CYCLE

Compactness (Prop. 1) buys boundedness, never convergence. The following explicit member of the flow class $\dot{\boldsymbol{g}} = -\boldsymbol{P_g}\boldsymbol{\Pi}\boldsymbol{P_g}\nabla_{\boldsymbol{g}}V$ has an *attracting limit cycle*, showing that a spectral or sign condition is genuinely required—the "limit cycles in principle" remark above is realised by a concrete construction inside the model class.

Proposition 4 (Attracting limit cycle with indefinite $\boldsymbol{\Pi}^s$)**.** *Let $d = 3$, $V(\boldsymbol{g}) = \frac{1}{2}g_3^2$, and $\boldsymbol{\Pi} = [\boldsymbol{e}_3]_\times + \varepsilon\,\mathrm{diag}(1,1,-1)$ with $\varepsilon > 0$, where $[\boldsymbol{e}_3]_\times$ generates rotation about $\boldsymbol{e}_3$. Then $\boldsymbol{\Pi}^s = \varepsilon\,\mathrm{diag}(1,1,-1)$ is indefinite ($\mathrm{spec}(\boldsymbol{\Pi}^s) = \{\varepsilon,\varepsilon,-\varepsilon\}$, so $\lambda_{\min} = -\varepsilon < 0$), yet the flow has a hyperbolically attracting limit cycle $\{g_3 = \sqrt{d/2}\}$ carrying pure rotation: every trajectory that does not start on the equator $\{g_3 = 0\}$ or at a pole (which are equilibria) converges to it.*

Proof. With $w := \boldsymbol{P_g}\nabla_{\boldsymbol{g}}V = g_3\boldsymbol{e}_3 - (g_3^2/d)\boldsymbol{g}$, direct computation gives $\dot{V} = -w^\top\boldsymbol{\Pi}^s w = \varepsilon\, g_3^2\big(1 - \frac{g_3^2}{d}\big)\big(1 - \frac{2g_3^2}{d}\big)$, so $\dot{V} > 0$ for $0 < g_3^2 < d/2$ (energy ascends) and $\dot{V} < 0$ for $g_3^2 > d/2$. The circles $\{g_3 = \pm\sqrt{d/2}\}$ are invariant; by rotational equivariance about $\boldsymbol{e}_3$ the drift has no component along the circle, so motion on it is pure rotation at angular speed $g_3^2/d = \frac{1}{2} \neq 0$. Linearising $s := g_3^2 - d/2$ gives $\dot{s} = -\varepsilon s$: the cycle is hyperbolically attracting for every $\varepsilon > 0$. □

Two consequences. **(i)** The attractor contains no critical point of V ($\boldsymbol{P_g}\nabla_{\boldsymbol{g}}V \neq 0$ on the cycle): with indefinite $\boldsymbol{\Pi}^s$, "where the dynamics goes" and "where the energy is critical" decouple. **(ii)** The energy is non-monotone along trajectories (ascent below the cycle, descent above) while the dynamics is perfectly stable — bounded, sign-varying energy drift is what this regime predicts, quantitatively the same phenomenon as the $+5$ to $+11$ drift of Table 10.

B.11 LOCAL SPECTRAL STABILITY: WHAT CERTIFIES CONVERGENCE

$\lambda_{\min}(\boldsymbol{M}_{\mathrm{sym}}) < 0$ is a worst-case statement over all of $\mathbb{R}^{2d}$; it rules out global *monotone* descent and nothing more. Note $\boldsymbol{M}$ is *constant* — built from $(\boldsymbol{\Pi}_a, \boldsymbol{\Pi}_f)$ alone; token/state dependence enters only through $w(t)$ (unlike the token-embedding stability matrix of prior energy-transformer analyses). Local stability of the attractors the model actually reaches is governed by a different, tangent-space object.

Proposition 5 (Linearisation on the sphere)**.** *Let $\boldsymbol{g}^\star$ be a Riemannian critical point of V ($\boldsymbol{P}_{\boldsymbol{g}^\star}\nabla V(\boldsymbol{g}^\star) = 0$, i.e. $\nabla V(\boldsymbol{g}^\star) = \mu\boldsymbol{g}^\star$ with $\mu = \boldsymbol{g}^{\star\top}\nabla V(\boldsymbol{g}^\star)/d$). The linearisation of the flow restricted to $T_{\boldsymbol{g}^\star}$ is*

$$D\mathbf{X}|_T = -\,\boldsymbol{\Pi}_T\boldsymbol{H}_T, \qquad \boldsymbol{\Pi}_T := (\boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{P})|_T, \qquad \boldsymbol{H}_T := (\boldsymbol{P}\nabla^2 V\boldsymbol{P} - \mu\boldsymbol{I})|_T,$$

where $\boldsymbol{H}_T$ is exactly the Riemannian Hessian of $V|_{\sqrt{d}\,S^{d-1}}$. For EGPT-Dual, $D\mathbf{X}|_T = -(\boldsymbol{\Pi}_{a,T}\boldsymbol{H}_T^{\mathrm{Atn}} + \lambda\boldsymbol{\Pi}_{f,T}\boldsymbol{H}_T^{\mathrm{FF}})$.

Proposition 6 (Local exponential stability; gyroscopic stabilisation)**.** *If $\boldsymbol{H}_T \succ 0$ (strict local minimum on the sphere) and every eigenvalue of $\boldsymbol{\Pi}_T\boldsymbol{H}_T$ has positive real part, then $\boldsymbol{g}^\star$ is locally exponentially stable, with rate* $\min \mathrm{Re}\,\mathrm{spec}(\boldsymbol{\Pi}_T\boldsymbol{H}_T)$.

- Sufficient condition: $\boldsymbol{u}^\top\boldsymbol{\Pi}\boldsymbol{u} > 0$ *for all nonzero $\boldsymbol{u} \in T_{\boldsymbol{g}^\star}$ (i.e. $\boldsymbol{\Pi}^s \succ 0$ restricted to the tangent space) implies* $\mathrm{Re}\,\mathrm{spec}(\boldsymbol{\Pi}_T\boldsymbol{H}_T) > 0$. *Proof: $\boldsymbol{\Pi}_T\boldsymbol{H}_T$ is similar to $\boldsymbol{B} := \boldsymbol{H}_T^{1/2}\boldsymbol{\Pi}_T\boldsymbol{H}_T^{1/2}$; for an eigenpair $\boldsymbol{B}\boldsymbol{y} = \lambda\boldsymbol{y}$,* $\mathrm{Re}\,\lambda = \hat{\boldsymbol{y}}^*\boldsymbol{B}^s\hat{\boldsymbol{y}}$ *with $\boldsymbol{B}^s = \boldsymbol{H}_T^{1/2}\boldsymbol{\Pi}_T^s\boldsymbol{H}_T^{1/2} \succ 0$ by congruence.*
- Not necessary — gyroscopic stabilisation. *Take $\boldsymbol{\Pi} = \left(\begin{smallmatrix} 1 & \omega \\ -\omega & -\delta \end{smallmatrix}\right)$, $\boldsymbol{H} = \boldsymbol{I}$:* $\mathrm{spec}(\boldsymbol{\Pi}^s) = \{1, -\delta\}$ *is indefinite, yet $-\boldsymbol{\Pi}\boldsymbol{H}$ is Hurwitz iff $\delta < 1$ and $\omega^2 > \delta$ (e.g. $\omega = 1.2, \delta = 0.4$: eigenvalues $-0.3 \pm 0.975i$). The state spirals into the equilibrium while the energy* rises on a positive fraction of steps*: the antisymmetric part rotates the ascent direction into the descent directions fast enough that the net motion contracts. This is precisely the trained regime — antisymmetric mass ≈ symmetric mass (App. F),* $\lambda_{\min}(\boldsymbol{\Pi}^s) < 0$*, bounded upward energy drift, and converging states, all at once.*
- Cheap necessary condition: $\mathrm{tr}(\boldsymbol{\Pi}_T^s\boldsymbol{H}_T) = \sum_i \mathrm{Re}\,\lambda_i > 0$ (*$\boldsymbol{\Pi}^s$ positive* on average *against the Hessian, even if not PSD).*

Proposition 7 (Discrete-time certificate: the implemented iteration)**.** *For the idealised sphere iterate* $\mathbf{F}(\boldsymbol{g}) = \sqrt{d}\,\mathrm{RN}\big(\boldsymbol{g} - \widetilde{\boldsymbol{\Pi}}_{\boldsymbol{g}}\nabla_{\boldsymbol{g}}V\big)$ *at a fixed point* $\boldsymbol{g}^\star$*:* $D\mathbf{F}|_T = \boldsymbol{I}_T - \boldsymbol{\Pi}_T\boldsymbol{H}_T$*, and* $\boldsymbol{g}^\star$ *is locally exponentially stable for the iteration iff the tangent spectral radius* $\rho(D\mathbf{F}|_T) < 1$*. For the implemented block the same criterion applies to the Jacobian of the actual iteration map, measurable directly by autodiff JVPs with no PSD assumption. It is the finite-step analogue of* $\mathrm{Re}\,\mathrm{spec}(\boldsymbol{\Pi}_T\boldsymbol{H}_T) > 0$*; the EFFN saturation bound on* $\|\nabla_{\boldsymbol{g}}E^{\mathrm{FF}}\|$ *keeps the effective step inside the contraction region, consistent with the* $2.93\times$ *overshoot of the parallel layout (Table 3).*

Resolution of the apparent contradiction. Global sign-definiteness of $\boldsymbol{M}_{\mathrm{sym}}$ certifies global *monotone* descent; its failure (Table 5) rules out exactly that and nothing more. The model converges the way a damped gyroscope does — non-monotone energy, contracting state — not the way gradient descent does. Table 6 organises the four levels of guarantee: only the last is the operative claim, and the first is an honest negative result.

Table 6: What each Lyapunov-type quantity certifies. Only the local-spectral row is the operative claim for trained models; the global worst-case row fails (Table 5), which is the honest negative result the diagnostic reports.

Quantity	**Certifies**	**Status**
$\lambda_{\min}(\boldsymbol{M}_{\mathrm{sym}}) \geq 0$	global monotone descent (LaSalle)	**fails** (Tab. 5)
$\lambda_{\max}(\boldsymbol{M}_{\mathrm{sym}}) \leq 0$	global monotone ascent (LaSalle on $-V$)	fails (indefinite)
$q_t = w_t^\top \boldsymbol{M} w_t / \|w_t\|^2$	realized descent/ascent balance	$q_t < 0$ on avg (Tab. 10)
$\rho(D\mathbf{F}\|_T) < 1$	convergence of the implemented iteration	**operative claim**

Checkpoint-only measurements (no retraining). (i) $\lambda_{\max}(\boldsymbol{M}_{\mathrm{sym}}) > 0$ on every block, with $|\lambda_{\max}| \approx |\lambda_{\min}|$ (Table 7): $\boldsymbol{M}_{\mathrm{sym}}$ is genuinely *indefinite*, not negative semidefinite (which would give benign monotone ascent, equally convergent on the sphere). (ii) A natural next measurement is the tangent spectral radius $\rho(D\mathbf{F}|_T)$ of the implemented block iteration at its reached states (JVP power iteration; a few hundred tokens suffice); $\rho < 1$ is the direct discrete certificate of local convergence, and we predict $\rho < 1$ for the distributed-refinement blocks and $\rho \geq 1$ where overshoot was observed.

Table 7: Both eigenvalues of $\boldsymbol{M}_{\mathrm{sym}}$ on an accessible EGPT-Dual 400M checkpoint (8 GPT + 4 energy, RMSNorm + Rayleigh, iter $[3, 4, 6, 9]$), a sibling run of the Table 5 models that reproduces the $|\lambda_{\min}| \approx$9–19 band. $\lambda_{\max} > 0$ on every block with $|\lambda_{\max}| \approx |\lambda_{\min}|$: $\boldsymbol{M}_{\mathrm{sym}}$ is genuinely *indefinite*, not negative semidefinite. Neither monotone-descent ($\boldsymbol{M} \succeq 0$) nor monotone-ascent ($\boldsymbol{M} \preceq 0$) holds; stability is the gyroscopic/local-spectral regime of Props. 5–7.

Block	**iters**	$\lambda_{\min}(\boldsymbol{M}_{\mathrm{sym}})$	$\lambda_{\max}(\boldsymbol{M}_{\mathrm{sym}})$
B8	$\times 3$	-10.45	$+12.06$
B9	$\times 4$	-11.24	$+11.35$
B10	$\times 6$	-10.65	$+10.04$
B11	$\times 9$	-19.13	$+18.26$

C ENERGY FFN: FULL GATE-SUBSTITUTION ANALYSIS

The forward pass of EFFN, derived in Section 2, Equation 1, is

$$\mathrm{EFFN}(\boldsymbol{h}) \;=\; \underbrace{\boldsymbol{\Pi}\boldsymbol{W}_2\,\sigma(\boldsymbol{W}_1\boldsymbol{g})}_{\text{standard MLP}} \;+\; \underbrace{\boldsymbol{\Pi}\boldsymbol{W}_1^\top\big[\sigma'(\boldsymbol{W}_1\boldsymbol{g}) \odot \boldsymbol{W}_2^\top\boldsymbol{g}\big]}_{\text{gated branch}}.$$

The gated branch matches SwiGLU's algebraic form $\boldsymbol{W}_o[\,\text{gate} \odot \text{value}\,]$ under three substitutions: (1) the swish gate is replaced by σ': for GELU, $\sigma'(z) = \Phi(z) + z\Phi'(z)$ is smooth and bounded on roughly $[0, 1.1]$ and qualitatively swish-like; for ReLU, σ' is a hard binary gate; (2) the value projection is tied to $\boldsymbol{W}_2^\top$ rather than an independent $\boldsymbol{W}_u$; (3) the output projection is tied to $\boldsymbol{\Pi}\boldsymbol{W}_1^\top$ rather than an independent $\boldsymbol{W}_o$. EFFN realises a SwiGLU-style gated computation with $\approx 2\,d\,d_{ff} + d^2$

parameters versus SwiGLU's $\approx 3\,d\,d_{ff}$, while conservativity ($\mathrm{FFN} = -\boldsymbol{\Pi}\nabla E^{\mathrm{FF}}$) additionally forces the standard-MLP term that SwiGLU omits.

D EGPT-DUAL VS HELMHOLTZ DECOMPOSITION

Decomposing each projection into symmetric and antisymmetric parts, $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}_i^s + \boldsymbol{\Pi}_i^a$ with $\boldsymbol{\Pi}_i^s = \frac{1}{2}(\boldsymbol{\Pi}_i + \boldsymbol{\Pi}_i^\top)$ and $\boldsymbol{\Pi}_i^a = \frac{1}{2}(\boldsymbol{\Pi}_i - \boldsymbol{\Pi}_i^\top)$, the two-projection update

$$\Delta \boldsymbol{h} \;=\; -\big[\,\boldsymbol{\Pi}_a^s\,\nabla E^{\mathrm{Atn}} + \lambda\,\boldsymbol{\Pi}_f^s\,\nabla E^{\mathrm{FF}}\,\big] \;-\; \big[\,\boldsymbol{\Pi}_a^a\,\nabla E^{\mathrm{Atn}} + \lambda\,\boldsymbol{\Pi}_f^a\,\nabla E^{\mathrm{FF}}\,\big] \tag{15}$$

splits into a dissipative gradient-flow term (symmetric metric) and a conservative Hamiltonian-bracket term (skew operator), realizing per-channel the gradient/curl split prescribed by the Helmholtz decomposition. With the identifications $\boldsymbol{M}_i := \boldsymbol{\Pi}_i^s$ and $\boldsymbol{J}_i := -\boldsymbol{\Pi}_i^a$ this is a two-channel GENERIC system in which the metric and Poisson operator are recovered without loss as the symmetric and antisymmetric parts of one learnable $\boldsymbol{\Pi}_i$. The only structural restriction relative to a fully general two-channel Helmholtz flow is the identification $B_i \equiv E_i$: the conservative dynamics on each module are driven by the same potential as the dissipative dynamics, rather than by an independent Hamiltonian. EGPT thus occupies the parameter-efficient subspace of Helmholtz flows in which gradient and curl share a potential per channel, paying only $O(d^2)$ extra parameters per module beyond the single-projection baseline.

E SUBSPACE-ALIGNMENT DIAGNOSTICS: SETUP, PROTOCOL, AND RESULTS

This appendix expands on §5.1: it gives the formal read/write-space decomposition behind the feedback-loop argument, the definition of prediction alignment, and the full empirical results.

E.1 READ AND WRITE SPACES, PREDICTION ALIGNMENT

Read and write spaces. For one head, $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_h \times d}$. Attention reads the residual stream through $\mathrm{row}(\boldsymbol{W}_Q)$, the *read space* $R := \mathrm{row}(\boldsymbol{W}_Q)$ (union over heads). Standard attention writes back via $\Delta \boldsymbol{h}_A^{\mathrm{Attn}} = \boldsymbol{W}_O \sum_B a_{AB} \boldsymbol{W}_V \boldsymbol{g}_B$, so updates lie in $W^{\mathrm{Attn}} := \mathrm{col}(\boldsymbol{W}_O \boldsymbol{W}_V)$; because $\boldsymbol{W}_O, \boldsymbol{W}_V$ are independent of $\boldsymbol{W}_Q, \boldsymbol{W}_K$, W^{Attn} has no architectural relation to R. EAttn's value tying $\boldsymbol{V} = \boldsymbol{W}_Q^\top \boldsymbol{K}$ writes

$$\Delta \boldsymbol{h}_A^{\mathrm{EAttn}} \;=\; \frac{\boldsymbol{\Pi}}{\sqrt{d_h}}\, J \sum_B a_{AB} \boldsymbol{g}_B \;\in\; \mathrm{col}(\boldsymbol{\Pi}\, J), \qquad J := \boldsymbol{W}_Q^\top \boldsymbol{W}_K, \tag{16}$$

so $W^{\mathrm{EAttn}} := \mathrm{col}(\boldsymbol{\Pi} J) = \boldsymbol{\Pi} R^\top$: the write space is the read space, rotated by $\boldsymbol{\Pi}$.

Prediction alignment. With weight tying $\boldsymbol{W}_U = \boldsymbol{W}_E^\top$, the LM head reads through $L := \mathrm{col}(\boldsymbol{W}_E) = \mathrm{row}(\boldsymbol{W}_U)$. A state $\boldsymbol{h}$ is *prediction-aligned at level* ϵ if

$$\epsilon(\boldsymbol{h}) \;:=\; \frac{\|(\boldsymbol{I} - \boldsymbol{P}_L)\boldsymbol{h}\|}{\|\boldsymbol{h}\|} \;<\; \epsilon, \qquad \boldsymbol{P}_L = \boldsymbol{W}_E(\boldsymbol{W}_E^\top \boldsymbol{W}_E)^{-1}\boldsymbol{W}_E^\top. \tag{17}$$

Small ϵ means $\boldsymbol{W}_U \boldsymbol{h}$ captures most of $\boldsymbol{h}$; large ϵ means substantial $\boldsymbol{h}$ content is invisible to the readout, occupying $L^\perp$.

Feedback-loop argument. The architectural fact above forces each EAttn block to write only through $\mathrm{col}(\boldsymbol{\Pi} J) = \boldsymbol{\Pi} R^\top$, the image of its *own* read space. This tying is exact within a block; each block has its own $\boldsymbol{W}_Q, \boldsymbol{W}_K$, so nothing ties different blocks together architecturally. Standard attention has independent $\boldsymbol{W}_V$, and each block's write space $\mathrm{col}(\boldsymbol{W}_O \boldsymbol{W}_V)$ is unconstrained relative to its own J or to other blocks'. Because end-to-end training only constrains the *final* state $\boldsymbol{h}^{(T)}$, standard attention is free to use its d_h-dimensional write channel idiosyncratically per block. In EAttn, each block's writes must lie in what it reads, and what it reads is what the previous block wrote; end-to-end training should therefore pull the write operators of successive blocks toward each other. This cross-block alignment is an expected consequence of training, not an architectural identity, and we test it empirically. *Where* that shared channel sits relative to $\boldsymbol{W}_U$'s spectrum is a separate empirical question; the data (§E.2) show a bimodal pattern, not the naive "aligned with the dominant directions of L."

What ρ^J does and does not measure. $\rho^J_{ij} = \cos(\mathrm{vec}(J_i), \mathrm{vec}(J_j))$ is a *cross-block* agreement metric, not a similarity to L. It is gauge-invariant under $(\boldsymbol{W}_Q, \boldsymbol{W}_K) \to (M^{-\top}\boldsymbol{W}_Q, M\boldsymbol{W}_K)$ for $M \in GL(d_h)$ that leaves attention scores fixed but makes raw $\boldsymbol{W}_Q, \boldsymbol{W}_K$ comparisons meaningless. There is an architectural asymmetry: in EAttn the value tying $\boldsymbol{V} = \boldsymbol{W}_Q^\top \boldsymbol{K} = J\boldsymbol{g}$ makes J *simultaneously* the scoring kernel and the value-subspace generator, so ρ^J measures both kernel and value-space agreement at once. In standard attention, $\boldsymbol{W}_V$ is independent of J; ρ^J then reports kernel agreement only, and the analog cross-block measure for value subspaces is $\rho^{\boldsymbol{W}_O\boldsymbol{W}_V}$ (gauge-fixed by including the output projection). Likewise, $\mathrm{align}(\boldsymbol{\Pi}J, L_k)$ is the EAttn write-space alignment with L; for standard attention the analog is $\mathrm{align}(\boldsymbol{W}_O\boldsymbol{W}_V, L_k)$.

E.2 Cross-block functional similarity

We measure three models: (a) **EGPT**: 8-layer GPT backbone + 4 EGPT blocks with iter$=[3, 4, 6, 9]$, d=1280, RMSNorm+Rayleigh, 400M; (b) **RecGPT**: same 8-GPT + 4-recurrent layout with iter$=[3, 4, 6, 9]$, all softmax attention ($\boldsymbol{W}_V$ independent, no $\boldsymbol{\Pi}$), d=1280 — structurally identical to (a) but without value tying; (c) **deep GPT**: 24-layer standard GPT, d=1024, 354M. Models (a) and (b) share the same width, the same iteration schedule, and the same training budget, isolating the effect of $\boldsymbol{W}_V = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$.

Metrics. For each block (or layer) i, define the *residual update* $\boldsymbol{\delta}_i = \boldsymbol{h}_i^{\mathrm{out}} - \boldsymbol{h}_i^{\mathrm{in}} \in \mathbb{R}^d$ at each token position. For each pair of blocks (i, j), we report the mean cosine similarity

$$\rho^\delta_{ij} \;=\; \frac{1}{T}\sum_{t=1}^{T} \frac{\boldsymbol{\delta}_i(\boldsymbol{g}_t)\cdot\boldsymbol{\delta}_j(\boldsymbol{g}_t)}{\|\boldsymbol{\delta}_i(\boldsymbol{g}_t)\|\,\|\boldsymbol{\delta}_j(\boldsymbol{g}_t)\|},$$

averaged over T token positions from 8 held-out text passages. For weight matrices $\boldsymbol{W}_i, \boldsymbol{W}_j$ we report $\cos(\boldsymbol{W}_i^\flat, \boldsymbol{W}_j^\flat)$ where $\flat$ denotes flattening.

Multi-head treatment. All weight metrics use *combined* (not per-head) matrices. For H heads with head dimension $d_h = d/H$, the combined matrices satisfy

$$J \;=\; \boldsymbol{W}_Q^\top \boldsymbol{W}_K \;=\; \textstyle\sum_h \boldsymbol{W}_{Q_h}^\top \boldsymbol{W}_{K_h}, \qquad \boldsymbol{W}_O\boldsymbol{W}_V \;=\; \textstyle\sum_h \boldsymbol{W}_{O_h}\boldsymbol{W}_{V_h},$$

so both are rank-$\leq d$ aggregate write operators, automatically summed over all heads. This is the correct gauge-invariant choice: the per-head decomposition is ambiguous under any head-wise rotation; the sum is invariant to it.

L_k and alignment. L_k denotes the k-dimensional subspace spanned by the top-k right singular vectors of the LM head $\boldsymbol{W}_U \in \mathbb{R}^{|V|\times d}$. These are the directions of $\mathbb{R}^d$ that $\boldsymbol{W}_U$ is most sensitive to. We use k=1024, which captures 92% of $\boldsymbol{W}_U$'s Frobenius energy for the d=1280 EGPT model. **Baseline**: a random $d\times d$ matrix W satisfies $\mathrm{align}(W, L_k) \approx \sqrt{k/d}$ (each column has expected squared projection k/d onto L_k). For k=1024, d=1280: $\sqrt{0.8} \approx 0.894$; for $k{=}d$ (deep GPT): $\sqrt{1} = 1.0$ trivially. Deviations above/below this baseline indicate non-random alignment with L.

Note on multi-block requirement. Cross-block metrics (ρ^δ, ρ^J, ρ^{WOV}) require at least two recurrent blocks to be meaningful. The EGPT model used here has 8 GPT backbone layers followed by **4 EGPT blocks with iter**$=[3, 4, 6, 9]$; only these four blocks enter the off-diagonal averages. The 8 GPT backbone layers are excluded because (i) they have $\boldsymbol{W}_V \neq \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ and would conflate two distinct mechanisms, and (ii) comparing backbone GPT blocks to recurrent EGPT blocks is the wrong question. RecGPT uses the same 4-block recurrent sub-network (positions 8–11) with standard softmax attention; all backbone layers for both models are standard GPT and are excluded.

E.3 Per-iteration $L^\perp$ trajectory (D2)

Setup. Using k=256 right-singular vectors of $\boldsymbol{W}_U$ as an approximation of L_k (capturing 54% of $\boldsymbol{W}_U$ energy for EGPT and 70% for deep GPT at k=256; absolute ϵ values are *upper bounds* on the true $L^\perp$ fraction), we measure $\epsilon(\boldsymbol{h}) = \|(\boldsymbol{I} - \boldsymbol{P}_{L_k})\boldsymbol{h}\|/\|\boldsymbol{h}\|$ at the input to each recurrent block and at each intermediate iteration.

Results.

- **EGPT (8-backbone+4-EGPT(3,4,6,9)):** Backbone GPT layers increase ϵ from 0.79 (B0) to 0.93 (B7, last backbone layer) — they push hidden-state content into $L^{\perp}$. Each EGPT block then *decreases* ϵ: the four EGPT blocks collectively bring it from 0.93 to 0.81. Within each block, ϵ decreases monotonically with iteration count (e.g. the 9-iteration block: ϵ drops $0.815 \to 0.791$ over 9 iterations), consistent with each step pulling $\boldsymbol{h}$ towards L.
- **Deep GPT (24-layer):** ϵ drops steeply in layers 0–3 ($0.76 \to 0.45$), stays near 0.44 through the middle layers (minimum prediction-misalignment), then *rises* in layers 16–23 ($0.48 \to 0.66$). The late-layer $L^{\perp}$ accumulation is consistent with these layers parking intermediate results in the scratch space invisible to the LM head, to be decoded by the final layer.

These patterns directly support the prediction of §5.1: EGPT iterations reduce $L^{\perp}$ content (write-space $\approx L$); standard attention can and does accumulate $L^{\perp}$.

E.4 LM-HEAD SINGULAR VALUE DISTRIBUTION AND BOTTOM-k ALIGNMENT

To properly test the scratch-space hypothesis, we need to know (i) the effective rank of the LM head $\boldsymbol{W}_U$ — how many singular directions carry non-negligible prediction energy — and (ii) whether each model's write operator aligns with the *bottom* singular vectors (the near-null space of $\boldsymbol{W}_U$) more or less than expected for a random matrix.

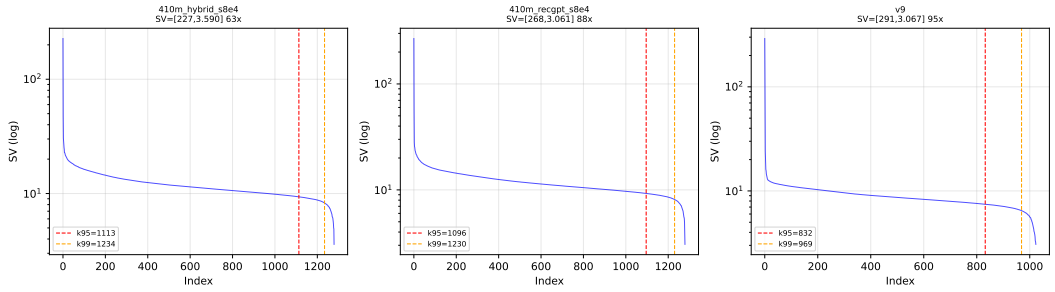


Figure 6: Singular value spectrum of $\boldsymbol{W}_U$ for the matched pair and deep GPT. Red dashed: k_{95} (index where cumulative energy crosses 95%); orange dashed: k_{99}. A fast decay means the bottom singular vectors are near-zero and constitute genuine scratch space; a slow decay means $\boldsymbol{W}_U$ is nearly full-rank and there is no meaningful $L^{\perp}$.

Bottom-k alignment. Let $\boldsymbol{V}_{\perp}^{(k')} \in \mathbb{R}^{d \times k'}$ denote the k' right-singular vectors of $\boldsymbol{W}_U$ with the *smallest* singular values. $\text{align}_{\perp}(W, k') := \|\boldsymbol{V}_{\perp}^{(k')\top} W\|_F / \|W\|_F$ measures overlap with the near-null space. Random baseline: $\sqrt{k'/d}$. Excess $=$ $\text{align}_{\perp} - \sqrt{k'/d}$: positive means the write operator concentrates into near-null space (scratch space usage); negative means it avoids it.

Singular value distribution. Both 410M models have d=1280 and $k_{95} \approx 1110$: 87% of singular dimensions are needed to capture 95% of $\boldsymbol{W}_U$'s energy (decay ratio $\approx$63–88$\times$). This means $\boldsymbol{W}_U$ is *nearly full-rank*; the true $L^{\perp}$ subspace is tiny ($\approx$46 dimensions carry $<1\%$ of $\boldsymbol{W}_U$'s energy at k_{99}). The scratch-space mechanism cannot rely on a large invisible subspace: it operates within the tail of L.

E.5 SCALING THE BIMODAL PATTERN: EGPT-DUAL AT 400M, 800M, AND 1B

Here we report the LM-head alignment pattern at 400M, 800M and 1B using the EGPT-Dual architecture (separate $\boldsymbol{\Pi}_{\text{attn}}$ and $\boldsymbol{\Pi}_{\text{mlp}}$ per block) and its matched RecGPT counterpart. All models share the same 8-GPT + 4-recurrent layout (iter$=[3, 4, 6, 9]$ at 400M/800M; iter$=[2, 3, 4, 6]$ at 1B) and iteration dropout.

Cross-block cosim. Table 8 reports the off-diagonal mean over the 4 recurrent blocks (B8–B11).

Table 8: Cross-block cosine similarity at three scales. EGPT-Single uses one shared $\mathbf{\Pi}$ per block; EGPT-Dual uses separate $\mathbf{\Pi}_{\text{attn}}$, $\mathbf{\Pi}_{\text{mlp}}$; RecGPT uses standard softmax attention ($\boldsymbol{W}_O\boldsymbol{W}_V$, no $\mathbf{\Pi}$). $\bar{\rho}^{\mathbf{\Pi}_a J}$: the projected attention write operator (the architecturally faithful analog of single-$\mathbf{\Pi}$'s $\bar{\rho}^{J}$). †EGPT-Dual $\bar{\rho}^{\delta}$ has two-group anti-correlation (B8 vs B9–B11); within-group cosim is 0.42–0.80.

		EGPT-Single	EGPT-Dual			RecGPT		
Metric	**Write op**	400M	400M	800M	1B	400M	800M	1B
d		1280	1280	1792	2048	1280	1792	2048
$\bar{\rho}^{\delta}$ (functional update)†	$\cos(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j)$	-0.043	-0.072	0.094	**0.246**	0.138	0.105	0.044
$\bar{\rho}^{J}$ (bare attention kernel)	$\cos(\boldsymbol{W}_Q^\top \boldsymbol{W}_K)$	**0.296**	0.038	0.115	0.082	-0.000	0.001	-0.000
$\bar{\rho}^{\mathbf{\Pi}_a J}$ (proj. attn write-op)	$\cos(\mathbf{\Pi}_a \boldsymbol{W}_Q^\top \boldsymbol{W}_K)$	0.131	0.121	0.075	0.035	-0.000	0.000	0.002
$\bar{\rho}^{\mathbf{\Pi}_f \boldsymbol{W}_2^\top \boldsymbol{W}_1}$ (proj. FFN)	$\cos(\mathbf{\Pi}_f \boldsymbol{W}_2^\top \boldsymbol{W}_1)$	—	0.127	**0.260**	0.252	N/A	N/A	N/A
$\bar{\rho}^{\boldsymbol{W}_2^\top \boldsymbol{W}_1}$ (raw FFN)	$\cos(\boldsymbol{W}_2^\top \boldsymbol{W}_1)$	—	0.002	-0.003	-0.001	0.000	0.000	-0.000
$\bar{\rho}^{\mathbf{\Pi}_{\text{attn}}}$ (raw proj)	$\cos(\mathbf{\Pi}_{\text{attn}})$	—	0.010	0.032	0.010	N/A	N/A	N/A

LM-head alignment at scale. Table 9 reports the write-operator alignment with the top-k and bottom-k right-singular subspaces of $\boldsymbol{W}_U$, as excess over the random baseline $\sqrt{k/d}$. EGPT-Dual write operator: $\mathbf{\Pi}_{\text{attn}} J$; RecGPT write operator: $\boldsymbol{W}_O\boldsymbol{W}_V$.

Table 9: LM-head alignment excess over random baseline $\sqrt{k/d}$ at 400M (d=1280), 800M (d=1792), and 1B (d=2048). **Bold**: strongest bimodal signal per column. The bimodal pattern (positive excess at both top-32 and bottom-32, deficit in the middle) persists across all three scales for EGPT-Dual and is absent in RecGPT.

	Top-k excess (dominant vocab)						**Bottom-k excess (near-null)**					
	EGPT-Dual			RecGPT			EGPT-Dual			RecGPT		
k	400M	800M	1B	400M	800M	1B	400M	800M	1B	400M	800M	1B
512	-0.078	+0.046	+0.002	+0.029	+0.039	+0.068	**+0.156**	**+0.184**	**+0.203**	-0.001	+0.021	+0.011
128	+0.112	+0.203	+0.107	+0.081	+0.063	+0.095	**+0.375**	**+0.361**	**+0.345**	+0.039	+0.059	+0.040
32	**+0.210**	**+0.241**	**+0.127**	+0.088	+0.053	+0.079	**+0.434**	**+0.336**	**+0.355**	+0.046	+0.070	+0.044

Interpretation. The bimodal alignment pattern persists at every scale:

- **Bottom-k excess scales consistently**: EGPT-Dual bot-32 excess = +0.43 (400M), +0.34 (800M), +0.36 (1B); RecGPT stays at +0.04–0.07 across all scales. The gap is 5–10×.
- **Mid-spectrum deficit persists**: EGPT-Dual top-512 excess is near zero or negative (-0.08, $+0.05$, $+0.00$); RecGPT is uniformly positive ($+0.03$–$+0.07$). EGPT avoids the bulk middle of $\boldsymbol{W}_U$'s spectrum at every scale.
- **Projected write-operator sharing ($\bar{\rho}^{\mathbf{\Pi}_a J}$).** At 400M, EGPT-Single ($\bar{\rho}^{\mathbf{\Pi} J} = 0.131$) and EGPT-Dual ($\bar{\rho}^{\mathbf{\Pi}_a J} = 0.121$) have nearly identical projected attention write-operator sharing: dual $\mathbf{\Pi}_a$ absorbs alignment that single $\mathbf{\Pi}$ leaves in bare J (0.296 vs 0.038), but the *net write operator* is equally shared. At larger scales, $\bar{\rho}^{\mathbf{\Pi}_a J}$ decreases (0.075 at 800M, 0.035 at 1B), suggesting dual projection allows per-block differentiation of the attention channel while the bimodal LM-head signature (bot-32 excess $\geq +0.34$) survives independently. The projected FFN kernel $\mathbf{\Pi}_f \boldsymbol{W}_2^\top \boldsymbol{W}_1$ remains the strongest cross-block sharing signal at scale (0.13–0.26).

These results are consistent with the prediction of §5.1: energy-attention write operators concentrate at both extremes of $\boldsymbol{W}_U$'s spectrum at every scale, while standard-attention write operators do not. The cross-block sharing of the projected write operator ($\bar{\rho}^{\mathbf{\Pi}_a J}$) decreases with scale, but the bimodal *alignment with* $\boldsymbol{W}_U$ does not: the two properties are dissociable.

F Helmholtz decomposition of the learned projection

We decompose the learned projection $\Pi = S + A$ into symmetric (S) and antisymmetric (A) parts and report the diagnostic quantities referenced in §5: $\cos\theta$ between the update direction $\Pi(\nabla E)$

and the gradient ∇E averaged over blocks/iterations on 20 WikiText validation samples; the *excess descent* z-score excess $= (\boldsymbol{g}^\top S\boldsymbol{g})/(\sigma_W \|\boldsymbol{g}\|^2)$ with $\sigma_W = \|S\|_F/d$, distributed $\mathcal{N}(0,1)$ for random W; the antisymmetric-to-symmetric ratio $\|A\|_F/\|S\|_F$; and the per-iteration energy change ΔE.

Table 10: Helmholtz decomposition of the learned Π (averaged over blocks/iterations, 20 WikiText samples). The energy *rises* across iterations and $\boldsymbol{\Pi}$ is empirically indefinite ($\|A\|/\|S\| \approx 1$), ruling out strict descent. Energy-attention models still learn projections whose symmetric part is biased toward descent (excess $+11.6\sigma$ for EGPT-Dual, $+16.3\sigma$ at 800M), whereas Attn-FFN-Prll's projection is random-like (-0.7σ) yet achieves the best PPL.

Model	Π **type**	$\cos\theta$	**Excess** (σ)	$\|A\|/\|S\|$	ΔE**/iter**	**PPL**
EGPT-Dual (400M)	Dual	-0.065	$+11.6$	0.97	$+5.0$	22.28
EGPT-Single (400M)	Single	-0.052	$+3.3$	0.96	$+4.4$	22.41
EAttn-FFN-Prll (400M)	Single	-0.055	$+9.8$	1.00	–	21.81
Attn-FFN-Prll (400M)	Single	$+0.007$	-0.7	1.00	–	21.82
EGPT-Dual (800M)	Dual	-0.064	$+16.3$	0.97	$+11.3$	19.00

G RMSNorm + explicit Rayleigh projection (160M)

To test that sphere completeness (Proposition 1) is the actual stability mechanism (not LayerNorm-specific), we replace the implicit projection of LayerNorm with RMSNorm $\boldsymbol{g} = \boldsymbol{h}/\mathrm{RMS}(\boldsymbol{h})$ followed by an explicit Rayleigh projection $\boldsymbol{P_g} = \boldsymbol{I} - \frac{1}{d}\boldsymbol{g}\boldsymbol{g}^\top$ applied to the update. The model is otherwise identical to EGPT-Dual.

Table 11 reports results on three 160M-parameter variants (1 recurrent EGPT block, 12 iterations, d=768, ≈160M total including the 77M embedding, 30k steps, 7.9B BPE tokens).

Table 11: Rayleigh projection at 160M scale. PPL: WikiText word perplexity. Avg 10: mean accuracy over 10 classification tasks. GSM8K^h: 5-shot flexible-extract exact match. Differences in Avg 10 and GSM8K are within a single standard error and should be read as noise.

Norm	**Rayleigh**	**PPL** ↓	**Avg 10** ↑	**GSM8K^h** ↑
RMSNorm	No	81.55	42.8%	1.67%
LayerNorm	No	81.00	41.6%	1.82%
RMSNorm+Ray	**Yes**	**80.42**	42.3%	1.67%

RMSNorm + Rayleigh achieves the best perplexity (80.42), improving by 1.1 points over plain RMSNorm and 0.6 points over LayerNorm, while matching both on classification tasks. Enforcing the tangent form of the update explicitly is therefore cheap and trains well, and the gain does not depend on the specific form of LayerNorm's implicit projection.

H 4-block study

Before scaling to deeper models, we verify the basic premise with a minimal setup: 4 layers (d=1280, ∼200–240M parameters), each iterated T=6 times (except GPT at T=1), trained for 30k steps.

Table 12: 4-block experiment (d=1280, 30k steps). All recurrent models use T=6 with iteration dropout ±2. Effective layers = blocks × iterations.

Model	**Params**	**Eff. layers**	**PPL** ↓	**Avg%** ↑	**GSM8K%** ↑
GPT (T=1)	233M	4	31.52	54.2	2.1
RecGPT	233M	24	**26.58**	56.5	**4.2**
Attn-FFN-Prll	240M	24	27.07	**56.9**	4.0
EGPT	201M	24	34.65	53.5	1.5
EGPT (T=1)	201M	4	48.15	51.0	1.4

Three observations emerge. First, **iteration helps substantially**: RecGPT (26.58 PPL) improves 5 points over the non-looped GPT baseline (31.52), confirming that weight-shared iteration is a powerful mechanism at small depth. Second, **EGPT underperforms** at this scale (34.65 PPL), worse even than GPT despite 6× the effective depth. The energy attention and energy FFN impose representational constraints that are too restrictive when all four layers must use them. Third, **Attn-FFN-Prll is close to RecGPT** (27.07 vs. 26.58), showing that the parallel layout update is nearly as effective as consecutive layout when combined with standard attention.

These results motivate the **hybrid design**: rather than making all blocks recurrent, we use 8 standard GPT blocks for general feature extraction, followed by 4 recurrent blocks that perform refinement. This gives the model sufficient standard capacity to reach competitive quality, while the recurrent blocks provide the opportunity for adaptive compute.

I TRAINING DETAILS

I.1 MODEL ARCHITECTURE

All models use a 12-layer decoder-only transformer with RMSNorm (epsilon 10^{-5}), Rotary Position Embeddings (RoPE, dimension 64), and tied input/output embeddings. We train at three scales, summarized in Table 13.

Table 13: Architecture hyperparameters for the three training scales.

Hyperparameter	**400M scale**	**800M scale**	**1B scale**
Hidden dimension d	1280	1792	2048
Attention heads	20	28	16
Head dimension	64	64	128
MLP intermediate size	5120	7168	8192
Number of layers	12	12	12
Vocabulary size	100,352	100,352	100,352
Max sequence length	4096	4096	4096
Iteration schedule (B8–B11)	$[3, 4, 6, 9]$	$[3, 4, 6, 9]$	$[2, 3, 4, 6]$
Iteration dropout	$\pm[1, 2, 2, 4]$	$\pm[1, 2, 2, 4]$	$\pm[1, 1, 2, 2]$

For the 12-layer hybrid models, blocks B0–B7 are standard GPT blocks (single pass, $T{=}1$) and blocks B8–B11 are looped with counts $[3, 4, 6, 9]$.

Table 14: Parameter counts for each architecture. Energy attention removes W_V and W_O; the energy FFN uses two weight matrices instead of three (no gate projection). Attn-FFN-Prll adds a $d{\times}d$ projection to standard softmax blocks.

Architecture	**Params**	**Notes**
GPT (no recurrence)	443.1M	Standard 12-layer baseline
RecGPT	443.1M	Same params as GPT (iterated)
Attn-FFN-Prll	449.6M	+ projection Π per recurrent block
EGPT-Dual	416.8M	EAttn ($-W_V, W_O$), energy FFN (-1 matrix)
EGPT-Sandwich	435.3M	Softmax B0–7, EAttn B8–10, softmax B11
EAttn-FFN-Cons	429.9M	EAttn, SwiGLU, no projection
EAttn-FFN-Prll	436.5M	EAttn, SwiGLU, $+\Pi$
EAttn-EFFN-Cons	403.7M	EAttn, energy FFN, no projection
Attn-EFFN-Prll	423.4M	Softmax attn, energy FFN, $+\Pi$

I.2 TRAINING CONFIGURATION

I.3 ITERATION DROPOUT

During training, each block's iteration count is perturbed by $\pm[1, 2, 2, 4]$ for blocks B8–B11, exposing the model to counts in $\{2{-}4\}$, $\{2{-}6\}$, $\{4{-}8\}$, $\{5{-}13\}$ respectively. All recurrent architectures use the same dropout configuration.

Table 15: Training hyperparameters (identical for all architectures).

Hyperparameter	Value
Data mixture	Nemotron-CC (NVIDIA et al., 2025) 70% MegaMath-Web-Pro (Zhou et al., 2025) 15% FineMath (Liu et al., 2024) 15%
Tokenizer	BPE tiktoken (100k vocab)
Optimizer	AdamW (β_1=0.9, β_2=0.95, ϵ=10^{-8})
Peak learning rate	2×10^{-3}
LR schedule	Cosine decay to $0.1\times$ peak
Warmup / decay steps	2,000 / 28,000
Total training steps	30,000
Weight decay	0.1
Gradient clipping	1.0
Precision	BF16
Micro batch size	2 per GPU
Gradient accumulation	8 steps

J Full benchmark results

Evaluation benchmarks. The 11 benchmarks comprising Avg% are: HellaSwag (Zellers et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018), PIQA (Bisk et al., 2019), BoolQ (Clark et al., 2019), COPA (Roemmele et al., 2011), SciQ (Welbl et al., 2017), WinoGrande (Sakaguchi et al., 2019), RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), and LAMBADA (Paperno et al., 2016). We use `acc_norm` for HellaSwag, ARC-E, ARC-C, PIQA, OpenBookQA, and SciQ; `acc` for BoolQ, COPA, WinoGrande, RACE, and LAMBADA. GSM8K (Cobbe et al., 2021) is evaluated separately (8-shot CoT, hash-match).

Table 16: Complete 400M benchmark results for all models (acc_norm used where available). Models in rows, benchmarks in columns.

Model	PPL↓	HSwag	ARC-E	ARC-C	PIQA	BoolQ	MMLU	Lamb.	GSM^{f}	GSM^{h}
GPT	22.05	47.5	58.1	32.0	69.6	58.3	29.8	41.7	4.5	5.9
EGPT-Dual	22.28	47.5	60.1	32.2	70.5	53.0	29.8	39.1	5.6	6.5
EGPT-Sandwich	22.17	47.6	60.7	32.6	69.6	50.1	30.6	42.1	4.6	5.2
RecGPT	22.37	47.5	60.6	32.4	70.1	**60.8**	29.8	**42.9**	7.4	6.7
Attn-FFN-Prll	**21.82**	48.2	**62.4**	31.8	69.9	**62.2**	29.8	41.0	**8.8**	**9.3**
EAttn-FFN-Cons	**21.80**	**48.5**	60.7	**33.0**	**70.5**	57.3	30.2	40.1	5.2	5.1
EAttn-FFN-Prll	21.81	**48.5**	61.7	32.2	**70.5**	56.6	30.8	40.5	3.3	6.8
EAttn-EFFN-Cons	22.55	46.7	59.0	32.0	70.3	56.0	30.4	40.4	3.9	4.2
Attn-EFFN-Prll	22.19	46.9	60.4	31.6	**70.5**	56.5	30.8	**43.0**	4.0	7.1

HSwag, ARC-E, ARC-C, PIQA use acc_norm. BoolQ, Lamb. use acc. GSM^{f}=flex, GSM^{h}=hash (8-shot CoT).

K Detailed LogitLens tables

Tables 18–19 show the full per-iteration LogitLens analysis for B11 of key models.

Table 17: Median correct-token rank at block boundaries (400M). Lower is better. Energy attention models enter B11 with low rank (already near convergence); softmax models defer prediction to B11.

	B8 in	**B8 out**	**B9 out**	**B10 out**	**B11 t=3**	**B11 final**
Energy attention models						
EGPT-Dual	16	9	7	3.4	2.5	2.5
EAttn-EFFN-Cons	38	20	13	3.9	2.4	2.3
EAttn-FFN-Cons	110	42	20	11	2.4	2.3
EAttn-FFN-Prll	81	24	19	5.5	5.2	2.4
Softmax attention models						
Attn-FFN-Prll	210	103	55	27	11	2.8
Attn-EFFN-Prll	45	19	10	8.3	3.8	2.6
RecGPT	3127	1636	571	258	4.5	2.5

Table 18: EGPT-Dual B11 LogitLens (400M). Converges by t=5.

Iter	**Entropy**	**Rank**	**Conv.**	**CKA%**	**Top-1%**	**KL**
input	2.05	3.4	–	–	67.1	0.591
t=1	2.94	3.1	0.516	97.0	69.7	0.342
t=3	3.21	2.5	0.237	97.9	82.4	0.071
t=5	3.32	2.4	0.056	99.9	97.7	0.003
t=9	3.35	2.5	0.025	100.0	100.0	–

Table 19: RecGPT B11 LogitLens (400M). Requires 5+ iterations.

Iter	**Entropy**	**Rank**	**Conv.**	**CKA%**	**Top-1%**	**KL**
input	6.14	258	–	–	3.4	8.133
t=1	4.49	93	0.730	90.2	6.8	5.647
t=3	2.80	4.5	0.432	96.0	52.5	0.610
t=5	3.34	2.4	0.124	99.6	95.0	0.013
t=9	3.49	2.5	0.036	100.0	100.0	–

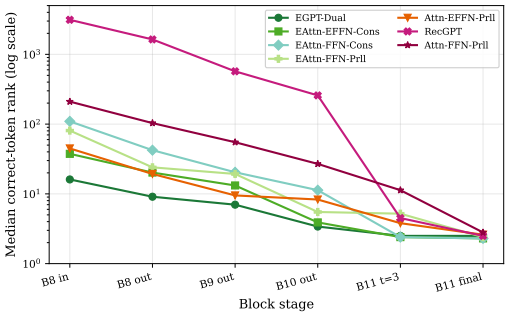
Table 20: EAttn-FFN-Cons B11 LogitLens (400M). Robust despite no projection.

Iter	**Entropy**	**Rank**	**Conv.**	**CKA%**	**Top-1%**	**KL**
input	3.41	11.3	–	–	37.7	1.833
t=1	2.40	6.2	0.448	96.5	46.0	1.218
t=3	2.70	2.4	0.274	93.3	84.2	0.122
t=5	3.32	2.3	0.116	99.5	97.4	0.003
t=9	3.34	2.3	0.024	100.0	100.0	–

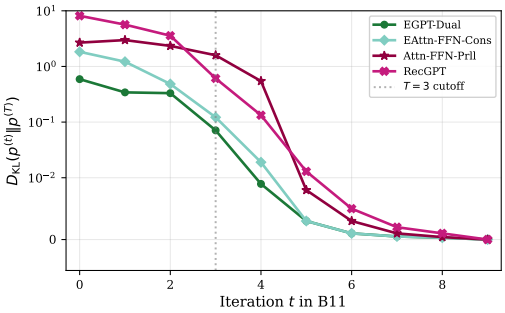
Table 21: Attn-FFN-Prll B11 LogitLens (400M). Best quality but collapses at T=3.

Iter	**Entropy**	**Rank**	**Conv.**	**CKA%**	**Top-1%**	**KL**
input	3.87	26.9	–	–	31.3	2.680
t=1	2.20	33.0	0.705	100.0	29.2	2.977
t=3	1.82	11.3	0.450	100.0	43.3	1.590
t=5	3.50	2.8	1.108	100.0	95.9	0.008
t=9	3.50	2.8	0.026	100.0	100.0	–

L TWO-MODES FIGURES (RANK TRAJECTORY AND KL CONVERGENCE)



(a) Median correct-token rank through recurrent blocks. Energy attention models progressively refine predictions; softmax models defer construction to the final block.



(b) KL divergence from the final prediction in B11. EGPT converges 1–2 iterations faster than RecGPT.

Figure 7: Two computation modes in looped transformers (400M). **(a)** Distributed refinement (energy attention) progressively improves predictions across blocks; concentrated construction (standard) defers to the final block. **(b)** Within B11, energy attention models converge faster, explaining their robustness to iteration reduction.

M PER-ARCHITECTURE PPL UNDER ITERATION REDUCTION

Table 22 lists PPL at T=3 (OOD), T=6 (ID), and T=9 (ID, training configuration) for all seven 400M architectures. This is the table version of the robustness partition discussed in Section 3.2; Figure 8 plots the same partition across all B11 iteration counts.

Table 22: Perplexity under reduced B11 iterations (400M). Trained with T=9 $\pm$ 4 (ID range: $T \in [5, 13]$). Bold = <15% degradation. T=3 is OOD.

B11 T		**Robust** (<15%)			**Partial** (100–150%)		**Collapse** (>500%)	
		EGPT-Dual	**EAttn-EFFN-Cons**	**EAttn-FFN-Cons**	**Attn-EFFN-Prll**	**RecGPT**	**EAttn-FFN-Prll**	**Attn-FFN-Prll**
3	OOD	**23.65**	**23.43**	**23.94**	49.30	52.91	142.9	400.7
6	ID	**22.29**	**22.58**	**21.81**	22.23	22.38	21.83	21.83
9	ID	22.28	22.55	21.80	22.19	22.37	21.81	21.82

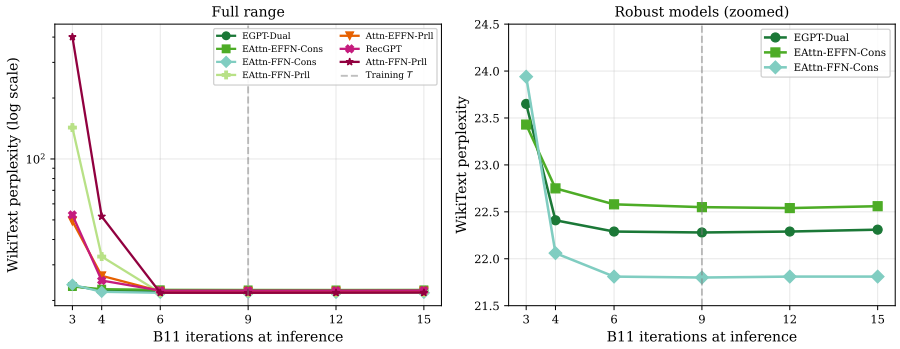


Figure 8: Iteration scaling at 400M. **Left**: reducing B11 iterations, three models with energy attention degrade gracefully while all others collapse. **Right**: zoomed view of robust models shows 4–10% degradation at T=3 (OOD), recovering by T=6 (ID).

N FULL BENCHMARK ROBUSTNESS UNDER ITERATION REDUCTION

Table 23 reports per-benchmark scores at T=1, T=2, T=3, and T=9 (baseline) for all eight recurrent architectures at 400M. This is the complete version of the main-body summary (Figure 3 and Table 22).

Table 23: Full benchmark robustness under iteration reduction (400M). B11 iterations are varied from T=9 (training configuration) down to T=1. All other blocks remain at training iterations $[3, 4, 6]$.

Model	B11 T	PPL ↓	HSwag%	ARC-C%	PIQA%	MMLU%	GSM8K^{h}%
EGPT-Dual	9	22.3	47.5	32.2	70.5	29.8	6.5
	3	23.6	46.4	31.7	70.5	29.6	5.0
	2	34.9	45.8	31.4	68.7	29.1	5.8
	1	30.1	44.6	31.1	68.2	28.9	4.5
EAttn-EFFN-Cons	9	22.5	46.7	32.0	70.3	30.4	4.2
	3	23.4	45.8	32.0	70.0	30.2	3.3
	2	25.1	45.1	31.2	70.0	29.8	3.5
	1	37.3	43.0	29.9	68.1	27.7	3.0
EAttn-FFN-Cons	9	21.8	48.5	33.0	70.5	30.2	5.1
	3	23.9	46.8	32.4	70.7	29.7	5.1
	2	35.8	44.3	32.0	69.5	28.9	4.2
	1	95.1	41.8	31.1	67.1	29.4	3.9
RecGPT	9	22.4	47.5	32.4	70.1	29.8	6.7
	3	52.9	42.1	30.7	68.5	29.3	6.1
	2	283	39.0	30.2	64.1	28.0	2.6
	1	1074	37.6	29.2	62.5	26.6	1.1
Attn-FFN-Prll	9	21.8	48.2	31.8	69.9	29.8	9.3
	3	401	41.2	30.0	65.2	28.4	5.5
	2	1818	39.1	30.8	64.0	27.9	2.4
	1	2071	37.5	29.7	62.8	29.9	0.8
EAttn-FFN-Prll†	9	21.9	48.5	32.2	70.5	30.1	4.7
	3	192	42.5	32.0	66.2	28.6	3.3
	2	6078	40.9	32.2	65.7	28.2	0.4
	1	8275	39.0	31.1	63.8	25.7	1.2
Attn-EFFN-Prll	9	22.2	46.9	31.6	70.5	30.8	7.1
	3	49.3	43.5	30.7	70.2	31.0	6.7
	2	75.4	41.7	30.6	68.1	30.4	5.0
	1	213	40.4	31.0	66.1	29.3	4.1

All benchmarks use acc_norm where available. GSM8K^{h} = hash prompt, 8-shot CoT.
†Dual-$\mathbf{\Pi}$ variant (PPL 192 at T=3, +775%). The factorial figure uses the single-$\mathbf{\Pi}$ variant (PPL 143, +555%).

O 1B SCALE RESULTS

We train EGPT-Dual and RecGPT at 1B scale (d=2048, 16 heads, intermediate 8192, 12 layers with iterations $[1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 4, 6]$) using the same hyperparameters as 400M/800M (lr=2e-3, 30k steps). EGPT-Dual has 944M parameters (energy FFN uses 2 weight matrices); RecGPT has 1.01B parameters.

Per-scale robustness breakdown. At T=3 (OOD), EGPT-Dual's PPL increase stays moderate (**+6%** at 400M, **+20%** at 800M, **+5%** at 1B) while RecGPT collapses (**+136%**, **+161%**, **+39%**). At T=1, EGPT-Dual retains functional GSM8K hash at all scales (**4.5%**, **6.3%**, **9.1%**), while RecGPT drops to effectively zero (**1.1%**, **0.1%**, **0.4%**).

O.1 FULL BENCHMARKS AT 1B

EGPT-Dual achieves better perplexity (17.79 vs. 18.12) despite fewer parameters, and leads on HellaSwag, ARC-C, and PIQA. RecGPT leads on GSM8K (18.3 vs. 12.0 hash), consistent with the 400M/800M pattern where softmax models are stronger on mathematical reasoning.

O.2 TTS ROBUSTNESS AT 1B

The two computation modes replicate at 1B. EGPT-Dual at T=3 shows only **5% PPL increase** with benchmarks preserved or slightly improved (HSwag 55.7 vs. 56.5, ARC-C 36.4 vs. 36.7). Even at T=1, EGPT-Dual retains **52.6% HellaSwag** and **9.1% GSM8K hash**. RecGPT is more robust at 1B than at 400M/800M at T=3 (+39% vs. +136%), likely due to the reduced iteration schedule $[2, 3, 4, 6]$, but still collapses completely at T=1 (PPL 963, GSM8K 0.4%).

Iteration scheduling at 1B. The 1B models use baseline schedule $[2, 3, 4, 6]$ (EGPT-Dual: 20.1 TFLOPs; RecGPT: 17.5 TFLOPs, measured with `torch.utils.flop_counter`). Table 26 shows the "global half" schedule $[2, 2, 3, 5]$ results with measured FLOPs.

EGPT-Dual saves **12.8%** of measured FLOPs with all benchmarks within noise of baseline ($\pm$0.3pp). RecGPT saves a comparable 11.8% but loses **3.4pp on GSM8K hash** (14.9 vs. 18.3) and 0.8pp on MMLU, showing that concentrated construction models sacrifice generation quality when iterations are reduced. At T=1 (no iteration), EGPT-Dual reaches PPL 27.7 (+56%) while RecGPT collapses to PPL 963, confirming the two computation modes at 1B scale.

Table 24: 1B benchmark results at full training iterations.

Benchmark	EGPT-Dual (944M)	RecGPT (1.01B)
WikiText PPL ↓	17.79	18.12
Lambada Acc ↑	49.3	47.7
HellaSwag ↑	56.5	55.6
ARC-Easy ↑	69.5	69.9
ARC-Challenge ↑	36.7	35.9
PIQA ↑	74.2	72.3
BoolQ ↑	53.4	61.7
COPA ↑	70.0	72.0
SciQ ↑	91.5	91.5
WinoGrande ↑	54.6	54.9
RACE ↑	34.0	33.8
OpenBookQA ↑	38.2	34.6
MMLU (5-shot) ↑	35.3	37.2
GSM8K flex ↑	11.5	16.4
GSM8K hash ↑	12.0	18.3

Table 25: 1B robustness under iteration reduction. B11 iterations varied from T=6 (training) down to T=1. EGPT-Dual remains functional at T=1 (PPL 27.7, GSM8K 9.1%); RecGPT collapses (PPL 963, GSM8K 0.4%).

Model	B11 T	PPL ↓	HSwag%	ARC-C%	PIQA%	MMLU%	GSM8K^{h}%
EGPT-Dual	6	17.8	56.6	36.7	73.0	35.2	12.0
	3	18.7	55.7	36.4	73.0	35.3	10.2
	1	27.7	52.6	35.7	72.0	35.0	9.1
RecGPT	6	18.1	56.0	36.3	72.6	37.2	18.3
	3	25.2	52.4	36.6	70.7	36.5	17.4
	1	963	42.8	31.3	64.3	27.2	0.4

Table 26: Iteration scheduling at 1B: global half $[2, 2, 3, 5]$ vs. baseline $[2, 3, 4, 6]$. FLOPs measured with `torch.utils.flop_counter` at seq_len 4096. EGPT-Dual saves 12.8% FLOPs with no benchmark regression. RecGPT loses 3.4pp on GSM8K hash at the same schedule.

Model	Schedule	TFLOPs	Saved	PPL	HSwag%	PIQA%	MMLU%	GSM8K^{h}%
EGPT-Dual	$[2, 3, 4, 6]$	20.07	–	17.79	56.5	72.9	35.3	12.0
EGPT-Dual	$[2, 2, 3, 5]$	**17.49**	**12.8%**	17.80	56.6	73.1	35.1	11.7
RecGPT	$[2, 3, 4, 6]$	17.49	–	18.12	55.6	72.8	37.2	18.3
RecGPT	$[2, 2, 3, 5]$	**15.43**	**11.8%**	18.34	54.9	72.7	36.4	14.9

P ADAPTIVE HALTING: FULL RESULTS

We evaluate three online halting criteria for energy attention models, selecting thresholds on 100 C4 validation samples (out-of-distribution with respect to the evaluation benchmarks), then running full benchmarks at the resulting schedules to confirm transfer.

Convergence rate. After each iteration t, compute $\rho_t = \|\mathbf{h}^{(t)} - \mathbf{h}^{(t-1)}\| / \|\mathbf{h}^{(t-1)}\|$. When ρ_t drops below a threshold, stop iterating. This requires no additional computation beyond a norm.

Token prediction stability. Project $\mathbf{h}^{(t)}$ through the LM head (LogitLens) and check whether $\arg\max p^{(t)} = \arg\max p^{(t-1)}$ for a target fraction of positions.

Fixed early exit. Cap all blocks at $T_{\max}$ iterations regardless of convergence.

Table 27: Adaptive halting at 400M: thresholds selected on C4 validation, evaluated on standard benchmarks. TFLOPs measured with `torch.utils.flop_counter` at seq_len 4096. Energy attention models save 12–21% of measured FLOPs with <1.1% PPL degradation. RecGPT degrades substantially under aggressive halting.

Model	Criterion	Schedule	TFLOPs	Saved	PPL	ΔPPL	GSM8K^{h}%
EGPT-Dual	baseline	[3,4,6,9]	11.55	–	22.28	–	6.5
EGPT-Dual	LogitLens	[2,3,5,6]	9.34	19.1%	22.30	+0.1%	6.7
EGPT-Dual	conv < 0.1	[3,4,5,4]	9.34	19.1%	22.40	+0.5%	5.2
EAttn-FFN-Cons	baseline	[3,4,6,9]	9.78	–	21.80	–	5.1
EAttn-FFN-Cons	conv < 0.15	[3,4,6,5]	8.63	11.8%	21.83	+0.1%	5.7
EAttn-FFN-Cons	fixed T=4	[3,4,4,4]	7.77	20.6%	22.03	+1.1%	6.0
RecGPT	baseline	[3,4,6,9]	10.07	–	22.37	–	6.7
RecGPT	token > 0.95	[3,4,6,6]	–	∼10%	22.38	+0.03%	6.5
RecGPT	fixed T=4	[3,4,4,4]	7.97	20.9%	25.02	+11.8%	5.8

Table 28: Adaptive halting at 800M. The pattern from 400M replicates: energy attention enables large compute savings while softmax models degrade substantially under aggressive halting.

Model	Criterion	Schedule	Saved	PPL	ΔPPL	GSM8K^{h}%
EGPT-Dual	baseline	[3,4,6,9]	–	19.00	–	10.5
EGPT-Dual	conv < 0.1	[3,4,5,5]	17%	19.03	+0.2%	10.7
EGPT-Dual	fixed T=4	[3,4,4,4]	23%	19.18	+0.9%	10.2
RecGPT	baseline	[3,4,6,9]	–	19.56	–	10.5
RecGPT	token > 0.95	[3,4,6,6]	10%	19.57	+0.05%	11.1
RecGPT	fixed T=4	[3,4,4,4]	23%	22.89	+17.0%	9.5

Q UNIFORM VS. HETEROGENEOUS SCHEDULES

Our main experiments use a heterogeneous schedule $[3, 4, 6, 9]$ that allocates more iterations to later blocks. Table 29 compares against uniform T=3 and uniform T=6.

The uniform T=3 results show both EGPT-Dual and Attn-FFN-Prll are essentially flat across all TTS values (PPL varies by <0.1 when B11 is changed from T=1 to T=9), but at a cost in MMLU and GSM8K. At uniform T=6, quality recovers and the robustness gap reappears, confirming the difference between energy and standard attention is not an artifact of the heterogeneous schedule.

Table 29: Uniform vs. heterogeneous iteration schedules (400M). All uniform models use iteration dropout ± 2 per block. Uniform T=3 models are robust by construction (flat across all T), but sacrifice quality and math reasoning. Uniform T=6 recovers quality but reintroduces the robustness gap. Note: Attn-FFN-Prll (parallel layout) is used for the softmax uniform comparison; RecGPT (consecutive) is the hetero baseline.

Model	Schedule	Iters	PPL	MMLU%	GSM8K^{h}%	B11=1 PPL	Robust?
EGPT-Dual	[3,4,6,9]	30	22.28	29.8	6.5	30.13	Yes
EGPT-Dual	[3,3,3,3]	20	22.48	29.3	4.5	22.53	Trivially
EGPT-Dual	[6,6,6,6]	32	22.24	30.2	5.8	34.86	Mild
RecGPT	[3,4,6,9]	30	22.37	29.8	6.7	1073.9	No
Attn-FFN-Prll	[3,3,3,3]	20	21.82	28.5	7.1	21.98	Trivially
Attn-FFN-Prll	[6,6,6,6]	32	22.10	31.6	8.0	295.3	No

R PER-BLOCK ANALYSIS

Table 30 varies iterations across all four blocks independently for five architectures, revealing the contribution of each block.

Early blocks are essential: removing B8–B10 ("only B11") degrades all models severely, even with full B11 iterations. The LogitLens-guided schedule $[2, 3, 5, 6]$ saves 20% compute with negligible quality loss across all energy attention models. Redistribution helps: shifting iterations from B11 to earlier blocks matches or improves the baseline for all models.

Table 30: Comprehensive per-block analysis (400M). Baseline: $[3, 4, 6, 9]$, 30 effective layers. "Global half" cuts all blocks by ∼50%. Bold PPL = <1% degradation from baseline.

		EGPT-Dual		EAttn-FFN-Cons		EAttn-EFFN-Cons		RecGPT		Attn-FFN-Prll	
Config	**Schedule**	**PPL**	**G%**	**PPL**	**G%**	**PPL**	**G%**	**PPL**	**G%**	**PPL**	**G%**
Baseline	[3,4,6,9]	**22.28**	6.5	**21.80**	5.1	**22.55**	4.2	22.37	6.7	**21.82**	9.3
LogitLens	[2,3,5,6]	**22.30**	6.7	**21.82**	5.2	**22.63**	4.1	22.55	6.8	22.13	6.7
Min dropout	[2,2,4,5]	**22.35**	5.5	**21.87**	5.7	**22.68**	4.4	23.06	5.7	22.71	5.6
Global half	[2,2,3,5]	**22.45**	5.8	**21.98**	5.2	**22.77**	3.3	24.52	4.7	23.99	5.5
Redistribute	[4,5,7,6]	**22.29**	5.7	**21.81**	5.6	**22.59**	3.4	**22.37**	6.8	**21.80**	9.6
B11 = 3	[3,4,6,3]	23.65	5.0	23.94	5.1	23.43	3.3	52.91	6.1	400.7	5.5
B11 = 1	[3,4,6,1]	30.13	4.5	95.12	3.9	37.27	3.0	1074	1.1	2071	0.8
B10 = 1	[3,4,1,9]	26.51	4.0	28.13	2.8	24.98	4.0	75.69	2.0	35.32	2.1
Only B11	[1,1,1,9]	32.57	3.9	28.73	2.6	26.58	4.0	1836	2.0	45.93	2.1
No recurrence	[1,1,1,1]	83.20	2.9	75.23	1.7	38.57	2.5	379.2	1.4	5231	0.8

G% = GSM8K hash%. Redistribute shifts compute from B11 to earlier blocks at matched total iterations.

S SEED VARIANCE AND AN INDEPENDENT 1B RUN

The main results are single training runs. To check that the robustness partition is not a seed artefact, we trained additional 400M runs with different seeds and a second, independently trained 1B EGPT-Dual run. Because of compute limits, the additional 400M runs use shorter training budgets than the main models (15.7B and ∼40B tokens instead of 126B).

Iteration reduction. Table 31 reports WikiText PPL degradation when B11 is reduced from $T{=}9$ to $T{=}3$ (from $T{=}6$ to $T{=}3$ for the 1B run). Every additional run stays far below the standard-attention and parallel-SwiGLU collapse range ($+122\%$ to $+1736\%$, Table 22). The seed spread is large at short training, and at 15.7B tokens the ordering of EGPT-Dual and EAttn-EFFN-Cons is the reverse of the fully trained 400M models. We therefore treat the qualitative partition as seed-robust and the ordering within the robust group as not established.

Table 31: Iteration-reduction robustness across seeds and runs (WikiText PPL, B11 reduced to $T{=}3$). The 3-seed rows report mean ± std over seeds.

Run	Tokens	PPL, full T	PPL, $T{=}3$	Degradation
EGPT-Dual 400M, 3 seeds	15.7B	–	–	$+14.2 \pm 6.7\%$
EAttn-EFFN-Cons 400M, 3 seeds	15.7B	–	–	$+3.3 \pm 1.2\%$
EGPT-Dual 400M, seed 42	∼40B	26.10	27.46	$+5.2\%$
EGPT-Dual 400M, seed 1234	∼40B	26.06	29.64	$+13.8\%$
EGPT-Dual 1B, second run	126B	18.60	18.68	$+0.4\%$

Iteration elasticity of the second 1B run. The second 1B run was trained with B11 at $T{=}6$ and iteration dropout covering $T \in [3, 9]$. Its PPL is nearly flat across that range and degrades gracefully beyond it (Table 32).

Table 32: WikiText PPL of the second 1B EGPT-Dual run as B11 iterations vary. $T{=}6$ is the training setting; $T{=}12$ is outside the dropout range.

B11 iterations	3	6 (train)	9	12 (OOD)
PPL	18.68	18.60	18.63	18.87
Change vs. $T{=}6$	$+0.4\%$	–	$+0.1\%$	$+1.4\%$

Benchmark variance. Table 33 compares the two ∼40B-token 400M seeds at a matched step. The gaps are small relative to the architecture effects in the main text.

Table 33: Benchmark variance between two EGPT-Dual 400M seeds at a matched step (∼19.3k steps, ∼40.5B tokens), and the second 1B run at full training. Avg is the 11-task average of Table 2; GSM8K-CoT uses flexible extract.

Run	WikiText PPL	Avg	MMLU	GSM8K-CoT
400M seed 42	26.10	50.49	27.21	3.56
400M seed 1234	26.06	49.63	27.61	3.71
Seed $\lvert\Delta\rvert$	0.04	0.86	0.40	0.15
1B second run (126B)	18.60	55.61	31.73	10.16

T EXTENDED RELATED WORK

Recurrent and looped transformers. Universal Transformers (Dehghani et al., 2019) share weights across all layers and use Adaptive Computation Time (Graves, 2016) for early halting. Deep Equilibrium Models (Bai et al., 2019) iterate to a fixed point via implicit differentiation.

Giannou et al. (2023) show that looped transformers can simulate arbitrary programs, while Fan et al. (2025) demonstrate improved compositional generalisation with loop-structured computation. Recent work scales looped LMs to billions of parameters and trillions of tokens: Ouro (Zhu et al., 2025) reaches 2.6B/7.7T; Parcae (Prairie et al., 2026) establishes scaling laws and gives a spectral stability framework via $\rho(\bar{\boldsymbol{A}}) < 1$ on the linear injection; Blayney et al. (2026) show that looped models self-organise into "stages of inference." Recurrent-depth language models (Geiping et al., 2025) iterate a shared block in latent space to scale test-time compute for reasoning, and Mixture-of-Recursions (Bae et al., 2025) learns token-level recursion depths. We deliberately use *recurrent* (not *looped*) for our setting to reflect that only a small number of late blocks share weights and iterate, rather than every layer of the network. Our stability story is geometric (compactness of the RMSNorm image) rather than spectral (constraint on ρ); the two are complementary, not competing.

Adaptive compute. PonderNet (Banino et al., 2021) learns a halting probability per step. Confident Adaptive Language Modelling (Schuster et al., 2022) uses confidence-based early exit across layers. Mixture-of-Depths (Raposo et al., 2024) routes tokens to skip layers entirely. These methods bolt an explicit halting / routing mechanism onto a fixed architecture. Our finding is structural: certain attention designs produce adaptive compute as an *emergent* property of the recurrent dynamics, without any halting mechanism, because their intermediate states are already prediction-aligned (§3, §5.1).

Energy-based attention. Dense Associative Memories (Krotov & Hopfield, 2016) connect attention to energy minimisation: a single attention pass reads as a step toward a local minimum of an associative-memory energy (Ramsauer et al., 2021). Energy Transformers (Hoover et al., 2024) extend this to a full iterative trajectory descending the energy. Our contribution is to identify which structural feature drives the empirical refinement gap: the value tying $\boldsymbol{V} = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ that this construction *forces*, which ties each block's residual write space to its own read space. We propose (§5.1) that this subspace coupling, rather than energy minimisation, is the mechanism behind distributed refinement; Dehmamy et al. (2026)'s decoder-only adaptation is the specific instantiation we use.

Representation analysis and intermediate-state probing. LogitLens (Belrose et al., 2023) and Tuned Lens approaches project hidden states through the output head to track how predictions form across layers. We apply this technique across *iterations* within a recurrent block, revealing two qualitatively different computation modes (distributed refinement and concentrated construction; §3). The closest prior conceptual work is Blayney et al. (2026)'s "stages of inference," which finds that looped models recapitulate feedforward layer roles across iterations; our finding is that this behaviour is not universal — it depends on the value-tying property of the attention update — and that the alternative regime (concentrated construction) is what loses test-time-compute robustness.

Scratch space in attention. Darcet et al. (2024) observed that vision transformers exhibit high-norm outlier tokens on uninformative image patches and that adding explicit non-decoded positions eliminates this artefact. The mechanism is related to what we identify in §5.1: standard attention has implicit scratch space (in writeback directions invisible to the readout) and uses it spontaneously when training tolerates it. EAttn's value tying constrains this scratch space architecturally; for tasks that benefit from non-vocabulary intermediate computation (e.g. multi-step arithmetic), this may be a net cost and a direction for future work.