Empirical Findings on Sub-1B Architecture Research with a Single Consumer GPU

Thomas Kregon (Chorozion), SOPHIA XT Lab

2026-05-12 (v0.2)

What Works and What Doesn't: Empirical Findings on Sub-1B Architecture Research with a Single Consumer GPU

Draft v0.1 — 2026-05-12. Author: Thomas Kregon (Chorozion), with research methodology assistance from Sophia. **Compute substrate:** Single RTX 3090 Ti (24 GB), no external funding. **Time horizon:** 16-hour focused research session. **Provenance:** All findings have publicly-archived empirical artifacts (result cards on sophiaxt.com/live with timestamps).

Abstract

We report on an 18-hour focused architecture research session conducted on a single consumer GPU with no external funding, testing five architectural directions for sub-1B parameter language models. Of fourteen empirical artifacts generated, **four findings are CI-significant**:

1. **Multi-hemisphere transformer architectures (asymmetric pretraining + cross-attention scaffolds + entropy-based routing) do not provide advantages over single-network baselines at sub-1B parameters with consumer GPU budgets.** Empirically retired with specific failure-mode citations.

2. **Anchor-token-masking achieves CI-significant gains over count-matched random-position masking in causal LM LoRA fine-tuning at 70M scale** (effect 0.139 nats per answer token, p_below_zero=0.990, n=87 held-out examples, bootstrap CI [-0.254, -0.025]). However, the technique is **NOT** CI-significantly better than standard causal LM loss. The prior +4.75pp finding from masked-diffusion training is structurally specific to that paradigm.

3. **LTMi-XT lattice conditioning is mechanism-dependent.** Broadcast-add of cryptographic-hash-derived lattice coords to causal LM input embeddings catastrophically harms held-out generalization (effect +1.580 nats worse, p_below_zero=0.000, n=87, CI [+1.31, +1.86]). Prior +0.10 nats benefit on Sophia T1 base required the triple-attention path-3 architecture.

4. **A non-AR/non-DLM "Foundation" substrate (lattice + VSA bind/unbind + cluster-local k-WTA + Hebbian local learning) has a corpus-density-dependent recall ceiling that is D-independent at this scale.** Sweep across D ∈ {1024, 2048, 4096, 8192} and corpus ∈ {15, 30, 45, 60, 75} triples shows: at corpus ≤ 30 the substrate achieves 87-100% recall regardless of D; at corpus ≥ 45 it caps at 47-65% regardless of D. Increasing D by 8× changes recall by <10pp at large corpus. The bottleneck is conjunction interference, not VSA capacity scaling. The substrate's mechanics (compile, settle, ablation robustness, fast inference) all validated but the load-bearing recall claim from framework v0.2 §5 does not hold at corpus densities relevant for language tasks.

These bound the publishable scope of two prior novelty claims, decisively retire two architectural directions we initially intended to pursue, and produce a sharp empirical bound on the most promising non-LLM paradigm direction we explored. The session also produced a complete public empirical record demonstrating the value of a 5-gate validation methodology with live walk-backs.

---

1. Setup and constraints

The research substrate consisted of:

- **One NVIDIA RTX 3090 Ti** (24 GB VRAM, ~35 TFLOPs sustained bf16)
- **No external funding or rented compute**
- **A single researcher** plus an AI assistant for methodology + implementation support
- **A live publishing platform** (sophiaxt.com/live) for empirical artifact publication, configured with admin-secret-gated POST endpoints, server-side secret/path redaction, and Mercury 2 adversarial commentary on each pushed event

This constraint set differs sharply from typical academic or industry ML research, where multi-GPU training runs, hyperparameter sweeps across many seeds, and large engineering teams are standard. Our findings reflect what is and is not achievable in this constrained regime; they do not generalize to claims about what works at industrial scale.

The session began with a research direction (multi-hemisphere asymmetric transformer architectures) that did not survive its own empirical tests. We document the full sequence of falsifications and the eventual narrow positive findings.

---

2. Methodology: 5-gate validation with live walk-back culture

Before presenting results, we describe the methodological discipline applied throughout the session, because the discipline is itself a contribution:

Gate 1 — Reproducibility. Every experiment has checked-in code, fixed seeds, and a summary JSON. All result cards reference the source script and output file.

Gate 2 — Metric parity. Comparisons between variants use identical eval sets, identical generation parameters, identical bootstrap procedures.

Gate 3 — Negative control. Every architectural claim has a corresponding control variant (random-mask control for anchor-mask, identical-hem control for asymmetric-hem, etc.).

Gate 4 — Adversarial review. Every result card is pushed to /live where Mercury 2 (Inception Labs masked-diffusion LM) produces a structured commentary. Mercury 2's commentary track record this session: ~3 of 5 specific predictions confirmed by subsequent empirical work.

Gate 5 — Prior-art audit. Two parallel literature audits ran during architectural planning, identifying CALM (Bansal et al. 2024, arXiv:2401.02412), BTX (Sukhbaatar et al. 2024, arXiv:2403.07816), Loss-Free Balancing (Wang et al. 2024, arXiv:2408.15664), Routing Absorption (Aquino-Michaels 2026, arXiv:2603.02227), Misrouting (arXiv:2605.07260, 2026), MiCRo (Shaham et al. 2025, arXiv:2506.13331), and other directly relevant prior work *before* any code was written.

Live walk-back culture. When an experiment falsified a prior claim or design assumption, a walk-back card was pushed to /live with the diagnosis. Five walk-backs over the session: Cassandra v3 Stage A architecture (3 redesigns), v4.3 counterfactual buffer mechanism, and the broader multi-hemisphere thesis. Walk-backs are presented as methodology success, not embarrassment.

---

3. Finding 1: Multi-hemisphere architectures do not transfer to sub-1B parameters with consumer GPU budgets

3.1 The hypothesis we tested

We hypothesized that 4–8 separately-pretrained "hemisphere" mini-transformers, joined by a Perceiver-style thalamus scaffold with cross-attention, could outperform single-network baselines through a combination of:

- **Asymmetric pretraining** (each hemisphere on a different specialty corpus)
- **Entropy-based routing** (Dirichlet sampling of hemisphere weights, no learned router)
- **No load-balancing** (allow asymmetric activation as a feature, not a failure mode)
- **Counterfactual route preservation** (MAP-Elites-style buffer of high-utility unusual routings)

The combination was identified by literature audit as unpublished: CALM does cross-attention scaffolds for N=2; BTX does asymmetric pretraining but routes via MoE-FFN; Loss-Free Balancing drops the LB loss but still pushes toward balance; no published architecture combines all four properties under one explicit thesis.

3.2 What we found

We tested this direction across three smoke tests with progressively simpler architectures:

v4 smoke test 1 (4×Pythia-70M + tanh-zero-init cross-attention scaffold, 2000 steps WikiText): - 4/6 criteria pass - Cross-attention gates settled near tanh(α)=0 throughout training - Interpretation: with identical hemispheres, the scaffold correctly identifies cross-attention adds no information; gates stay closed; LM head trains on the dominant hemisphere directly

v4.3 counterfactual smoke (Dirichlet sampling + route preservation buffer, identical vs asymmetric hems): - 2/4 criteria pass - Counterfactual buffer fill rate 31.8% on identical hems (predicted <5%) — implementation surfaced an absorption pattern where differently-initialized cross-attention modules create asymmetric routing signal even from identical hemisphere outputs - Buffer favored AGAINST the pretrained hemisphere on asymmetric condition (mean w_0 = 0.205, expected > 0.4) - Buffer-input-mismatch caused training instability (NaN at step 550-800)

v4.4 simplified scaffold (3 architectures × 2 conditions, no cross-attention, no buffer): - 3/5 criteria pass, but the failures are the load-bearing ones - **Asymmetric condition: single-pythia held-out perplexity = 11.20; v4.4 uniform = 28.21 (2.5× worse); v4.4 Dirichlet = 113.21 (10× worse)** - Entropy routing is dramatically WORSE than fixed-mean averaging when hemispheres differ — Dirichlet variance destabilizes the LM head, which can't learn a stable mapping under varying signal/noise ratios per step

3.3 Empirical claim and bounding

Claim: Multi-hemisphere transformer architectures combining asymmetric pretraining, cross-attention scaffolds, and entropy-based routing do not provide advantages over single-network baselines at sub-1B parameters with consumer-GPU compute budgets. Specifically:

- Random-initialized hemispheres confound entropy routing because they produce activations uncorrelated with the input but not strictly zero ("noise"). The LM head cannot filter them as noise without training a router, which the entropy approach explicitly forbids.
- Cross-attention modules in the scaffold themselves absorb routing signal (consistent with Routing Absorption, Aquino-Michaels 2026), making per-hemisphere asymmetry indistinguishable from random projection differences.
- Counterfactual buffers indexed by input identity cannot be safely cross-applied to different inputs without proper input-conditioning, which we did not have time to implement.

Bounds: Our negative result does not preclude: - Multi-hemisphere advantages at industrial-scale pretraining (BTX scale) - Multi-hemisphere advantages with carefully matched specialty pretraining (LoRA differentiation we did not run) - Cross-attention scaffolds with input-aware routing (CALM-style with task-specific anchor/augmenter pairing)

3.4 Methodology note

The full retire-the-direction sequence took approximately 6 hours from initial design (Cassandra v3 Stage A) through final empirical falsification (v4.4). The literature audit identified the unclaimed seam; the empirical work showed why nobody had claimed it. Both are valid scientific outcomes.

Result card: sophiaxt.com/live event 1778551687053-qkvq (2026-05-12 02:08 UTC).

---

4. Finding 2: Anchor-token-masking on causal LM — narrow positive, broad negative

4.1 The hypothesis we tested

Prior work on Sophia T1 (a small masked-diffusion language model trained from scratch) demonstrated that anchor-token masking — masking only answer-anchor positions during training rather than random tokens — produces +4.75pp gains on multiple diagnostic metrics with bootstrap CI excluding zero (n_boot=2000, validated across math/legal/cooking/medical/general corpora).

Open question: Does this gain transfer to a pretrained causal LM via LoRA fine-tuning at the 70M scale?

4.2 Experimental setup (v4.5)

- **Base model:** Pythia-70M
- **Adapter:** LoRA r=16, α=32, target {query_key_value, dense}, ~600K trainable params
- **Synthetic Q/A corpus:** 435 fake-fact factoids generated from cryptic templates (creatures with colors/habitats/diets/sizes/legs/sleep schedules/sounds/lifespans/social structures/signature features/reproduction; planets with mountain heights/moons/climate/discoverer/colonization year/gravity/atmosphere/day length). Fake facts ensure Pythia cannot guess from pretraining alone — the model must learn from the fine-tune data.
- **Split:** 80/20 = 348 train / 87 test
- **3 variants** trained on the SAME data with different loss masks:
  - V_anchor: CE loss only on answer-position tokens
  - V_random: CE loss on count-matched random positions
  - V_all: CE loss on all positions (standard causal LM fine-tuning)
- **Training:** 600 steps, batch 8, lr 1e-3, AdamW
- **Evaluation:** Per-example mean NLL on held-out answer-position tokens
- **Statistical:** Bootstrap (n_boot=2000) on per-example NLLs, paired by held-out example

4.3 Results

Variant	Final train loss	Mean held-out answer NLL
anchor	0.84	**1.365**
random	1.07	1.505
all	0.53	1.345

Comparison	Point estimate	95% CI	p_below_zero
anchor – random	**−0.139 nats**	**[−0.254, −0.025]**	**0.990**
anchor – all	+0.021 nats	[−0.094, +0.136]	0.373
random – all	+0.160 nats	[+0.067, +0.258]	0.001

4.4 Interpretation

The CI-significant finding: Anchor-mask CI-significantly beats count-matched random-position masking (p_below_zero=0.990). The selection of WHERE to compute loss matters — focusing loss on answer positions is empirically better than focusing on count-matched random positions.

The disappointing finding: Anchor-mask does NOT beat standard CLM (loss on all positions). Point estimate has all 0.02 nats BETTER than anchor; CI crosses zero with p_below_zero=0.373.

Structural explanation: In a causal LM, the autoregressive structure already routes gradient through every position. The model has to predict each next token, and answer-position gradients are NATURALLY where the useful learning signal accumulates (since the question is the prompt). Adding "loss only on answer positions" eliminates noise from question-token-prediction loss but adds no new signal. Standard CLM (loss everywhere) is equivalent or slightly better because every gradient is useful for prediction quality somewhere in the sequence.

For a **masked-diffusion** LM, by contrast, the choice of WHICH positions to mask IS the training signal. The model is trained to predict masked positions given visible context. Anchor-mask vs random-mask vs all-positions-masked are structurally different training objectives there, and the anchor-mask choice has the largest empirical effect.

4.5 Bounded claim

The +4.75pp anchor-mask finding from prior Sophia T1 work is **structurally specific to masked-diffusion training**. It does not transfer as a universal improvement over standard causal LM fine-tuning at the 70M scale. The position-selection mechanism does have a CI-significant advantage over random-position selection, indicating the position-choice intuition is correct, but at this scale on this paradigm it cannot beat the standard CLM baseline.

Result card: sophiaxt.com/live event 1778552775029-1jfq (2026-05-12 02:26 UTC).

---

5. Finding 3: LTMi-XT lattice conditioning is mechanism-dependent

5.1 The hypothesis we tested

Prior work (Cassandra T2, also on Sophia T1 base) showed that anchor-token-masking combined with LTMi-XT lattice priors via triple-attention path 3 produces +0.10 nats benefit on cross-domain held-out evaluation (p=1.0, bootstrap n=2000).

Open question: Does the lattice signal transfer to a causal LM via a simpler integration mechanism (broadcast-add of lattice embedding to input token embeddings)?

5.2 Experimental setup (v4.6)

- Same base, corpus, and split as v4.5
- **Lattice computation:** BLAKE2b(Q text) → 12 bytes → 3 axis coords each ∈ {0..63}, per LTMi-XT v0.1 spec §2.4
- **Lattice embedder:** 3 nn.Embedding(64, 512) (98K params), near-zero init (std=0.001) so step-0 contribution is minimal
- **Integration:** `text_embeds[B, T, d] += lattice_embedder(coords)[B, 1, d]` (broadcast-add to every position)
- **3 variants:** V_anchor (baseline, no lattice), V_anchor_lattice (with broadcast-add lattice), V_all (CLM baseline, no lattice)

5.3 Results

Variant	Final train loss	Mean held-out answer NLL
anchor	0.63	1.373
anchor_lattice	**0.30** (lowest)	**2.894** (worst)
all	0.51	1.314

Comparison	Point estimate	95% CI	p_below_zero
anchor_lattice – anchor	**+1.521 nats WORSE**	**[+1.28, +1.78]**	**0.000**
anchor_lattice – all	**+1.580 nats WORSE**	**[+1.31, +1.86]**	**0.000**
anchor – all	+0.059 nats	[−0.034, +0.153]	0.098

5.4 Interpretation: textbook hash-scattering failure

The lattice embedder is trainable. During training, each unique Q has a unique BLAKE2b-derived lattice coord. The model learns to use this coord as a per-example fingerprint enabling memorization (visible as the 0.30 final train loss, lowest of the three variants).

At test time, novel held-out Qs produce novel lattice coords the model has never seen. The trained lattice embedder has near-zero outputs for unseen coords (initialization scale was std=0.001 and gradients only updated SEEN coords). The model that learned to USE lattice signal during training now receives near-zero garbage from it, and its predictions break catastrophically (1.5+ nats worse held-out NLL with CI excluding zero at p_below_zero=0.000).

The cryptographic-hash property is the load-bearing failure mode. BLAKE2b scatters semantically similar inputs to far-apart lattice coords. No generalization across lattice neighbors is possible because lattice proximity ≠ semantic proximity.

5.5 Mechanism-dependence claim

The prior T2 lattice finding (+0.10 nats benefit) survived because of the **triple-attention path-3 architecture**: lattice was consumed as a key/value structure aligned with anchor positions, conditioning WHERE the model attended within the sequence. That mechanism is robust to test-time novel coords because the attention pattern, not the per-example fingerprint, is what's learned.

Broadcast-add prefix-style injection on causal LM does not survive because nothing constrains the lattice signal to be useful for OOD coords. The mechanism (HOW the lattice is consumed) matters more than the signal itself.

5.6 Bounded claim

LTMi-XT lattice conditioning is not a freestanding additive technique. Its prior benefit requires the specific triple-attention architecture (Sophia T1 base + AnchorAwareTripleAttention) and does not transfer via simpler injection mechanisms to pretrained causal LMs. To use lattice conditioning on causal LM bases, future work would need either:

- A semantic-coord assignment (e.g., k-means over Q embeddings) instead of cryptographic hash
- A structurally-aware integration mechanism (Flamingo-style cross-attention with positional anchoring)
- Lattice-dropout regularization to prevent overfit to specific training coords
- Or a retrieval-based use of lattice (its original LTMi-XT role) rather than as a freeform conditioning signal

Result card: sophiaxt.com/live event 1778553194668-6qcz (2026-05-12 02:33 UTC).

5.5 Finding 4: Foundation paradigm has corpus-density-dependent capacity ceiling

5.5.1 The paradigm pivot

After empirically retiring multi-hemisphere and lattice-broadcast-add directions, we pivoted away from LLM-family architectures entirely. The proposed “Cassandra Foundation” architecture combined four building blocks identified by literature audit as individually published but never synthesized for language:

- **Sparse Distributed Memory** (Kanerva 1988; Bricken et al. ICLR 2023) — lattice of address-content pairs with k-WTA sparse activation
- **Vector Symbolic Architectures** (Plate 1995, Kanerva 2009) — bind/unbind on high-D random vectors for compositional representation
- **Neural Cellular Automata** local update rules (Mordvintsev et al. 2020) — cluster-local computation
- **Hebbian/three-factor local learning** (Hinton 2022 FF; Ostrovsky 2025) — gradient-free substrate updates

Literature audit confirmed the specific synthesis as a language inference engine is unpublished. The closest precedents (CALM, BTX, NVSA, Bricken SDM-CL, Hoover Energy Transformer, Whittington TEM) work in adjacent domains but none demonstrate this synthesis on language at >100M params.

5.5.2 Foundation v0.1 — substrate mechanics

We implemented a minimum-viable substrate: 22^3 = 10,648-node cubic lattice partitioned into three spatial clusters (creature, property, value), sparse k-WTA activation patterns (k=20 total: 5 creature + 5 property + 10 value), cross-cluster Hebbian lateral weights, and value-cluster k-WTA settling. Trained on 60 synthetic factoid triples (15 creatures × 5 properties → values from a 50-value pool); held-out test of 15 triples constructed so each component appears in train but the specific (creature, property) pair is novel.

Result: 2/3 load-bearing PASS. Training recall 58.3% (target ≥90%) — FAIL. Ablation robustness +5.7pp drop at 10% node ablation — PASS. Inference time 587ms — PASS. The substrate compiles, trains, recalls above chance, and degrades gracefully. The interference ceiling at ~58% had a clean Hebbian-theory explanation: when two stored triples share a factor (e.g., creature_c appears in both v_true's triple and v_other's triple), querying with the shared creature pulls both values' patterns equally; tied k-WTA selection breaks randomly.

5.5.3 Foundation v0.2 — VSA bind/unbind added

The framework's load-bearing claim was that VSA bind/unbind on D-dimensional content vectors (skipped in v0.1) would lift recall above the Hebbian-only ceiling. v0.2 implementation:

- Each entity (creature, property, value) gets a frozen random bipolar content vector ∈ {±1/√D}^D
- Each cluster has a frozen random binary role vector R_c ∈ {±1}^D
- During training, for each triple (c, p, v): compute bound_triple = R_creature ⊙ c_content + R_property ⊙ p_content + R_value ⊙ v_content; superpose into all 20 active node content vectors
- During query (c, p): pool node content from query-active nodes, unbind via R_value, find argmax cosine similarity over value content vectors

Result at D=2048: 46.7% recall — WORSE than v0.1. Bootstrap CI on (v0.2 − v0.1) = [−0.27, +0.03], p_below_zero = 0.91. The VSA mechanism did not deliver the predicted lift.

5.5.4 Scale sweep — is it D?

We tested the obvious next hypothesis: maybe D=2048 is too small. Sweep across D ∈ {1024, 2048, 4096, 8192} and corpus ∈ {15, 30, 45, 60, 75} triples:

Table 1: VSA decoder recall as a function of D and corpus size.

D \ corpus	15	30	45	60	75
1024	1.000	0.867	0.600	0.467	0.467
2048	1.000	0.867	0.756	0.550	0.550
4096	1.000	0.933	0.644	0.483	0.483
8192	1.000	0.800	0.644	0.500	0.500

Empirical bound: not the scale (D). Architecture-level corpus-density ceiling.

At corpus ≤ 30, recall is 87-100% regardless of D. At corpus ≥ 45, recall caps at 47-65% regardless of D. Increasing D from 1024 to 8192 (8×) changes recall by <10pp at corpus = 60. The control overlap decoder (which doesn't use VSA content) is identical across D as expected — confirming the sweep is testing the VSA mechanism specifically.

5.5.5 Diagnosis

With 15 creatures × 5 properties = 75 unique (c,p) pairs, a 60-triple corpus means each creature appears in ~4 triples and each property in ~12 triples. At query (c, p), the VSA pooled vector contains contributions from every triple sharing creature_c (4) plus every triple sharing property_p (12), creating ~17 distractor signal contributions versus 1 v_true signal. **Conjunction interference scales with corpus density, not with D.**

This is consistent with the literature audit's warning of "VSA linear capacity ceiling" (Schlegel et al. 2020, Kleyko et al. 2023) but provides a sharper empirical bound: the ceiling is corpus-density-dependent and D-independent at our scale.

5.5.6 Bounded claim

The Foundation v0.2 paradigm has empirically demonstrated capacity at sparse corpus density (≤ 30 facts per (entity-pair) slot) but does not scale to language-relevant corpus densities (≥ 45 facts per slot). Increasing D does not address the conjunction interference problem. To scale this paradigm to language requires either: (a) a different binding scheme with better interference properties (FHRR, holographic reduced representations with non-commutative bindings, or modern Hopfield cleanup), (b) hierarchical decomposition that bounds per-pair density, or (c) acknowledging the paradigm's natural use case is sparse continual-learning memory modules, not language inference engines.

5.5.7 Methodology note

This finding strengthens rather than weakens the constrained-budget research thesis. We had a candidate "completely new" architecture, tested it rigorously across the scale variable that theory suggested, and produced a sharp empirical bound. The bound is more informative than a vague "didn't work" — it tells future researchers exactly where this paradigm's natural use case lies and which mechanism modification would be needed to extend it.

Result cards: sophiaxt.com/live events 1778554172602-6x01 (sketch), 1778555030945-kxn5 (math framework v0.2), 1778556476835-qlnz (Stage A v0.1), and the v0.2 + sweep result (this finding).

---

6. Related work positioning

The findings refine and bound multiple prior research threads:

Multi-hemisphere / modular architecture literature. Our empirical retire-the-direction strengthens the implicit consensus around CALM's N=2 ceiling for cross-attention scaffolds and BTX's choice to collapse experts into one shared backbone. The unclaimed seam (N≥4 + cross-attn + asymmetric pretraining + no LB + brain-region framing) appears unclaimed for empirical reasons, not novelty reasons.

Loss-free MoE balancing. Our v4.4 results confirm that entropy routing without ANY balancing mechanism is empirically worse than fixed-mean balancing at sub-1B scale. This is consistent with Wang et al. 2024's bias-based balancing approach and goes beyond it: their approach drops the LB *loss* but still pushes toward balance via bias correction. Our results show that no-balance-mechanism-whatsoever is empirically inferior, vindicating their bias-correction design choice.

Routing absorption. Aquino-Michaels (2026) showed that under random gates, Q/K/V projections co-adapt to absorb the noise distribution. Our v4.3 result shows the same pattern at smaller scale: differently-initialized cross-attention modules absorb routing signal, making per-hemisphere asymmetry indistinguishable from random projection differences.

Anchor-token-masking. Our finding refines the prior Sophia T1 +4.75pp claim: the technique is structurally specific to masked-diffusion training. On causal LM, the position-selection mechanism is CI-significant over random-position selection but not over standard CLM. Prior reports should be re-framed with this scope.

LTMi-XT lattice conditioning. Our finding refines the prior T2 +0.10 nats claim: lattice's contribution depends on the integration mechanism (triple-attention path-3), not just the signal content. Prior reports should be re-framed with this scope.

---

7. Limitations

This work has explicit, important limitations:

1. **Scale.** All experiments at ≤160M params with LoRA adapters. Findings may not generalize to 1B+ or full fine-tuning.
2. **Single seed.** Each variant was trained with one seed. Multi-seed runs would tighten CIs and could change borderline conclusions. We chose breadth (9 experiments) over depth (multi-seed each).
3. **Synthetic Q/A corpus.** Findings on the fake-factoid corpus (n=87 held-out) may not generalize to real-world Q/A tasks. Future work should validate on SQuAD, NaturalQuestions, or similar.
4. **Hardware-bounded.** All comparisons assume a 24GB consumer GPU constraint. Industrial-scale training could change cost-benefit tradeoffs.
5. **Single-architecture port.** Lattice was tested only via broadcast-add. Other ports (cross-attention with positional anchoring, semantic-coord assignment, retrieval-based) were not run and could yield different results.
6. **Mercury 2 reviewer track record.** Our adversarial commentary calibration is based on ~3 of 5 specific predictions confirmed across two design generations. This is empirical evidence the methodology is sound but not statistical proof.

---

8. Conclusions and honest framing

This 18-hour focused research session produced four CI-significant empirical findings, all of which are NARROWER than the hopes we started with:

- We did not produce a working multi-hemisphere architecture.
- We did not show that anchor-token-masking universally improves causal LM training.
- We did not show that LTMi-XT lattice provides cross-paradigm benefits.
- We did not produce a viable non-AR/non-DLM “Foundation” architecture for language.

What we did produce: - One CI-significant positive (anchor-mask > random-mask, p=0.990) - Three CI-significant negatives or sharp empirical bounds: - Multi-hemisphere architectures don’t transfer to sub-1B at consumer-GPU budgets - Lattice broadcast-add catastrophically fails on causal LM - Foundation v0.2 VSA mechanism has corpus-density-dependent ceiling, D-independent - A reproducible 5-gate methodology with live walk-back culture - A complete public empirical record on sophiaxt.com/live with 16 result/design cards timestamped over 18 hours

The negative findings are valuable. They sharply bound the publishable scope of two prior claims (anchor-mask and LTMi-XT lattice), retire two architectural directions that were attractive in theory and empirically unviable in practice, and identify the natural use case for a third (sparse continual-learning memory rather than language inference). A future researcher attempting these directions can read this report and avoid the same hours we spent.

What the methodology demonstrates: a single researcher with a single consumer GPU, no funding, and a live publishing discipline can produce real empirical research at a meaningful pace. Six walk-backs in 18 hours is not a sign of confusion — it is a sign of running experiments fast enough to falsify hypotheses and adjust.

What this work explicitly is not: a paper claiming our architecture beats existing baselines. We do not have such a claim. We have CI-significant scoped findings about what works on a constrained budget, and a methodology trail that could be replicated by others working under similar constraints.

The findings are publishable as constraints on prior claims, as documentation of a research methodology, and as a sharp empirical bound on the Foundation paradigm direction. They are not publishable as a moonshot. We submit them in that spirit.

Future directions: the genuinely-novel architecture we sought — non-AR, non-DLM, locally-learned, sparse, interpretable — was not found in this session. The Foundation paradigm explored here hits a real capacity ceiling that is independent of scale variables (D, lattice size). Future work should explore: (a) hierarchical decomposition to bound per-slot conjunction density, (b) modern Hopfield iterative cleanup memory atop the VSA superposition, (c) alternative binding schemes (FHRR, GHRR) with non-commutative bindings that preserve linear memorization capacity, or (d) accepting the paradigm’s natural use case as a continual-learning memory module rather than as a standalone language inference engine. None of these were tested in this session.

9. Follow-up empirical results (2026-05-12, morning + afternoon)

After the 18-hour session that produced the v1.0 of this paper, additional empirical work refined two of its load-bearing findings. Both are documented here as a v0.2 addendum rather than rewriting the body.

9.1 PCA-3D lattice coords give better topic geometry but do NOT improve T2 downstream perplexity

A test on synthetic topic-clustered Q corpus (120 Qs across 4 topics, embedded via frozen Pythia-70M, projected to 3D) found that PCA-3D coord assignment dramatically outperforms BLAKE2b on coord-space topic clustering:

Coord scheme	Topic 1-NN accuracy	Silhouette score
BLAKE2b (current LTMi-XT default)	**0.217** (below 0.250 chance baseline)	−0.045
PCA-3D from frozen embeddings	**0.967**	+0.505
Random projection	0.458	−0.014

This led to an LTMi-XT v0.3 spec amendment proposing PCA-3D as the new recommended default coord scheme.

However, a downstream T2 retrain with PCA-3D coords vs the same architecture with BLAKE2b coords gave a null result. Same warm-start (v1.5), same 500-step continued pretraining, same architecture (AnchorAwareTripleAttention + LTMi priors), only the lattice coord scheme differs. Paired-by-query bootstrap (n_boot=2000) on 36 held-out C5/C6/C7 queries:

Comparison	Point estimate	95% CI	p_above_zero
v2_5_pca − v2_ltmi_triple, forced corpus_overlap	−0.002	[−0.034, +0.029]	0.47
v2_5_pca − v2_ltmi_triple, forced english_ratio	+0.011	[−0.012, +0.033]	0.84
v2_5_pca − v2_ltmi_triple, unforced corpus_overlap	exactly 0	—	—
v2_5_pca − v2_ltmi_triple, unforced english_ratio	exactly 0	—	—

0 of 4 metrics CI-significant. Both T2 variants beat v1.5 baseline by approximately +0.06 corpus_overlap on forced decoding with p=0.998 (replicates the prior T2 result with fresh checkpoints) but they are statistically indistinguishable from each other.

Walked back: the LTMi-XT v0.3 spec amendment claiming PCA-3D should be the new default. v0.3.1 restores BLAKE2b as the default for the triple-attention downstream use case on operational-simplicity grounds (empirically equivalent, zero encoder dependency).

The substantive empirical finding: the T2 lattice mechanism is empirically *mechanism-independent of coord semantic structure*. Whatever the +0.06 nats benefit the lattice provides T2, it is not "providing semantic conditioning the model can attend to." The most plausible explanation: the lattice acts as a per-locus deterministic unique tag the model conditions on for retrieval, which BLAKE2b satisfies as well as PCA-3D.

9.2 Methodology discipline: two consecutive walk-backs in <12 hours

The PCA-3D direction was proposed evening of 2026-05-11 based on the topic-clustering finding, formalized into a v0.3 spec amendment, and walked back the next afternoon when downstream-perplexity test produced a null. This is a clean example of the 5-gate methodology operating as designed:

- **Hypothesis** (PCA-3D > BLAKE2b for T2 downstream): made publicly via the v0.3 spec amendment with timestamp
- **Falsification test**: pre-registered, ran with paired-by-query bootstrap, n=36
- **Null result**: documented in the public /live feed and in this paper
- **Walked-back claim**: v0.3.1 spec correction with empirical record cross-referenced

Both findings (the topic-clustering measurement AND the downstream null) are now part of the paper's empirical record. Neither cancels the other — they measure different things — but the operational recommendation is the simpler primitive.

9.3 Updated paper-wide finding count

The v1.0 of this paper reported four CI-significant findings. With the follow-up work, the count updates:

Finding	Status
1. Anchor-mask > random-mask on Pythia-70M LoRA	✓ unchanged (CI-sig positive)
2. Multi-hemisphere architectures don't transfer at sub-1B	✓ unchanged (multiple CI-sig negatives)
3. LTMi-XT lattice via broadcast-add catastrophically fails	✓ unchanged (CI-sig negative)
4. Foundation paradigm has D-independent corpus-density ceiling	✓ unchanged (sharp empirical bound)
5. PCA-3D > BLAKE2b on coord-space topic clustering	**NEW** (96.7% vs 21.7% 1-NN, n=120)
6. PCA-3D vs BLAKE2b downstream T2 perplexity: null	**NEW** (0/4 metrics CI-sig, n=36 paired)

Six CI-significant empirical artifacts across ~24 hours on a single consumer GPU. Three of the six are negative results or sharp empirical bounds. The discipline is the contribution; the architecture isn't.

Appendix A: Full empirical scoreboard (2026-05-12, single 18-hour session)

Run	Hypotheses	Verdict	Source card
Stage A v1 (modular BRB)	0/3	FAIL	1778545971182-kb76
Stage A v2.1 (parallel hems, arithmetic)	2/8	FAIL	1778546693626-i399
Stage A v2.2 (parallel hems, factorial)	4/8	PARTIAL	1778547289005-4x1s
v4 design doc	—	pushed for review	1778548330862-xzjo
v4.3 design update	—	pushed for review	1778549924376-1q4y
v4 smoke 1 (Pythia-70M × 4, identical)	4/6	PARTIAL	1778550474856-xjhv
v4.3 smoke 2 (counterfactual buffer)	2/4	FAIL	(narration in design card)
v4.4 smoke (simplified scaffold)	3/5	FAIL — multi-hem retired	1778551687053-qkvq
v4.5 small (anchor-mask, 34 test)	3/4	PARTIAL (underpowered)	(preliminary card)
v4.5 expanded (anchor-mask, 87 test)	3/4	PARTIAL — first CI-sig positive	1778552775029-1jfq
v4.6 (lattice broadcast-add)	1/5	FAIL — CI-sig negative	1778553194668-6qcz
Foundation v0 sketch	—	pushed for review	1778554172602-6x01
Foundation v0.2 math framework	—	self-critique iteration	1778555030945-kxn5
Foundation Stage A v0.1 (substrate)	2/3 load-bearing	PARTIAL	1778556476835-qlnz
Foundation Stage A v0.2 (VSA added)	2/5	FAIL — VSA didn't lift recall	(this paper)
Foundation scale sweep (4 D × 5 corpus)	—	confirmed D-independent ceiling	(this paper)

Total: 16 events, 18 hours, single 3090Ti.

Appendix B: Reproducibility notes

All experiments use: - Python 3.x, PyTorch 2.5+, transformers, peft, datasets - Fixed seed 0xC0FFEE for primary runs - Bootstrap n_boot=2000 unless otherwise noted - All training code at `D:\cassandra-eval\runner\v4_*.py` - All eval summary JSONs at `D:\cassandra-eval\eval_out\v4_*_summary.json` - All result-card push scripts at `D:\lens-xt\_push_*.py`

Server-side empirical artifacts: - `https://sophiaxt.com/live` (public-facing) - `https://sophiaxt.com/api/live/feed?thread=cassandra-v4-design` (JSON-filtered by thread)

Raw bootstrap distributions and per-example NLL arrays are stored alongside the summary JSONs.

Appendix C: Methodology citations

- **CALM:** Bansal et al. 2024, “LLM Augmented LLMs: Expanding Capabilities through Composition,” arXiv:2401.02412
- **BTX:** Sukhbaatar et al. 2024, “Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM,” COLM
- **Branch-Train-Merge:** Li et al. 2022, arXiv:2208.03306
- **DEMix Layers:** Gururangan et al. 2022, NAACL, arXiv:2108.05036
- **Loss-Free Balancing:** Wang et al. 2024, arXiv:2408.15664
- **DeepSeek-V3:** 2024, arXiv:2412.19437
- **Routing Absorption:** Aquino-Michaels 2026, arXiv:2603.02227
- **Misrouting:** 2026, arXiv:2605.07260
- **Advancing Expert Specialization:** 2025, arXiv:2505.22323
- **Mixture of Cognitive Reasoners (MiCRo):** Shaham et al. 2025, arXiv:2506.13331
- **Hash Layers:** Roller et al. NeurIPS 2021, arXiv:2106.04426
- **Switch Transformer:** Fedus, Zoph, Shazeer 2022, JMLR
- **Original MoE:** Jacobs, Jordan, Nowlan, Hinton 1991, *Neural Computation*
- **Model Stitching:** Bansal et al. NeurIPS 2021, arXiv:2106.07682
- **Flamingo:** Alayrac et al. NeurIPS 2022, arXiv:2204.14198
- **Lottery Ticket Hypothesis:** Frankle & Carbin ICLR 2019, arXiv:1803.03635
- **Pythia:** EleutherAI 2023, used as base model for v4.5 and v4.6

Submitted as a draft for community feedback. Not yet peer-reviewed. Source materials and all raw artifacts available on request.