

# Supplementary Material

## THEIA: Learning Complete Kleene Three-Valued Logic in a Pure-Neural Modular Architecture

Augustus Haoyang Li  
Irvine Valley College, Irvine, CA, USA  
hli132@ivc.edu

This supplement gives the details behind the main article. Sections refer to the main article as “main Section X”. Code, data generators and diagnostics are at <https://github.com/xxx12e/theia-k3>.

### A Hypotheses and Analysis Choices

**H1 (reliability of the step architecture under discretized composition).** In the mod-3 chain task, the new state is (previous state + local verdict) mod 3, so no absorbing state fixes the rest of the chain. Phase 1 trains the step module on single steps (local accuracy). Phase 2 trains the transition network (state, local verdict  $\rightarrow$  next state), and Phase 3 trains end to end on 5-step chains with Gumbel-softmax straight-through discretization. At 500 steps, with all else fixed, a 3-layer flat MLP step module (0.80M or 2.75M parameters) falls to chance. A ResMLP grid crosses 4 or 8 blocks with expansion 2 or 4, with 5 seeds each. Three configurations have about 2.78M parameters, and  $4b \times 4d$  has 2.85M. Accuracy rises with depth, but only 3/20 trials reach  $\geq 99\%$  in the *as-specified aggregate*. The *strict aggregate* (2/19) leaves out one seed that needed a plateau-restart. A pre-LN Transformer reaches  $99.24\% \pm 0.34\%$  (3/5 seeds  $\geq 99\%$ ), and THEIA  $99.96\% \pm 0.04\%$  (5/5 seeds  $\geq 99.9\%$ , one after a plateau-restart). *Falsified by:* a flat MLP or ResMLP with  $\geq 99\%$  accuracy on a majority of the strict-aggregate seeds under any Phase 3 learning rate, depth, width or expansion beyond the grid.

**H2 (uncertainty-verdict asymmetric propagation, with an implementation check by activation patching).** (a) Upstream, a 2-layer MLP probe does not decode the 3-class verdict above the 73.4% oracle reference. (b) HU is decodable at every upstream boundary at  $\geq 75\%$  (majority  $\approx 52\%$ ). (c) Patching the Order output  $\mathbf{v}_{\text{ord}}$  flips matched non-absorbent pairs. By Eq. 4 of the main article, the Logic Engine receives only the Order output, the Set output and the logic operator, and the two inputs of a pair differ only in the Order Engine input. So a correct implementation flips every eligible pair, and part (c) tests the implementation, not a learned mechanism. Part (a) is partly architectural (Section D). *Falsified by* any of: (i) an upstream MLP probe above 73.4% verdict accuracy, (ii) HU accuracy below 75% at any upstream boundary, or (iii) patching of  $\mathbf{v}_{\text{ord}}$  failing to flip a non-trivial fraction of matched  $T \vee F / U \vee F$  pairs.

**H3 (convergence speed under shared optimizer settings).** Under the *matched protocol*, THEIA and the Transformer share THEIA’s development default: AdamW with  $\text{lr } 10^{-3}$ , cosine schedule, batch 4096 and no warmup. “12/12” means that all 12 targeted rules (main Table 2) are above 99%. The Kleene-aware stopping criterion requires overall accuracy above 99.9% and all 12 rules above 99% on two consecutive checkpoints. H3 states that, under the matched protocol and a 150-epoch budget, THEIA reaches 12/12 in less wall-clock time than a somewhat larger 8-layer Transformer (3.64M vs. 2.75M parameters). *Falsified by:* a Transformer baseline that matches THEIA’s  $7.93 \pm 1.40$  min Kleene-aware convergence ( $n = 5$ ) on a majority of seeds under the matched protocol. With  $n = 8$  Transformer seeds, 7/8 converge in  $51.5 \pm 11.0$  min, and the  $6.5\times$  gap remains. With the Transformer-standard recipe for the Transformer only ( $n = 3$ ), the gap at its first 12/12 crossing narrows to  $\sim 3.6\times$  (Section G).

**Analysis choices fixed after seeing the data.** We fixed four analysis choices after we saw the data. First, we chose the thresholds of the Kleene-aware stopping criterion by looking at THEIA’s convergence trajectory. Section G repeats the H3 comparison with a criterion that uses only overall accuracy. Second, the first probe protocol (the 70/30 protocol) used a 70/30 train/test split and chose the MLP epoch by test accuracy. The clean-split protocol came later. It uses a

60/20/20 train/validation/test split, chooses the MLP epoch and the SVM  $C$  on the validation split, and evaluates on the test split once. Main Table 5 still uses the 70/30 protocol. Third, besides the 73.4% threshold of H2(i), Section D also reports margins to a stricter 72% value. Every upstream verdict probe value is below both. Fourth, the  $\geq 99.9\%$  threshold in the chain comparison is not part of H1, so that comparison is exploratory. These criteria were not deposited in an external registry.

## B K3 Diagnostic: Construction and Rule Frequencies

**Construction.** The Arithmetic Engine computes  $c = a \oplus b$ . The Order test compares  $c$  with  $d$  under relation  $R$ , and the Set test checks whether  $c \in S$ . Their true K3 values,  $\text{val}_{\text{ord}}$  and  $\text{val}_{\text{set}}$ , are the two operands of the final logic operator. A targeted test makes one of them Unknown and keeps the other definite. Two construction errors can silently corrupt such a test, and they affected some of our earlier runs.

*Pitfall 1: cross-domain Unknown contamination.* A simple way to make  $\text{val}_{\text{ord}}$  Unknown is to set  $a$  to Unknown. This also affects  $c$ , because the natural rule is that  $c$  is Unknown if  $a$  or  $b$  is Unknown. The Set Engine takes  $c$  as input, so  $\text{val}_{\text{set}}$  also becomes Unknown, whatever  $S$  is. The test then checks (Unknown, Unknown) in place of (Unknown, definite), and a different K3 rule decides the verdict. *Fix:* we inject Unknown through  $d$  and keep  $a$  and  $b$  definite.

*Pitfall 2: edge case in relation construction.* To force  $\text{val}_{\text{ord}} = \text{True}$  under “strictly greater than”, one may set  $d = \max(0, c - 1)$ . This works when  $c \geq 1$ . When  $c = 0$ , however,  $d = 0$  and  $0 > 0$  is False. About 5% of samples have  $c = 0$  (subtraction with  $a = b$ , or modulo when  $b$  divides  $a$ ). For them  $\text{val}_{\text{ord}}$  is False, which corrupts every rule that needs  $\text{val}_{\text{ord}} = \text{True}$ . The per-rule accuracy still looks plausible, but it is biased downward by the fraction of  $c = 0$  samples. *Fix:* we use “greater than or equal” with the same  $d$ , which is always True because  $c \geq \max(0, c - 1)$  for all  $c \geq 0$ .

We call the construction with both fixes the *doubly-fixed diagnostic protocol*. It produces all per-rule accuracies in the main article. We re-ran existing checkpoints under it, and the new per-rule numbers fall within the noise band of the multi-seed sweep.

**Two checkpoints for seed 42.** One consistency check re-trained seed 42 under the doubly-fixed protocol. This gave a second seed-42 checkpoint. The retrain first reached 12/12 at epoch 80 (7.31 min) and converged at epoch 100 (9.24 min). The wall-clock entry for seed 42 in main Table 4 uses this retrain, so the 5-seed mean of  $7.93 \pm 1.40$  min includes it. The earlier seed-42 checkpoint was saved under a 200-epoch budget. Under the doubly-fixed protocol it also passes all 12 targeted rules and all 39 rules at  $> 99\%$ . All other 5-seed analyses that include seed 42 use the earlier checkpoint: the probes, the centroid geometry, activation patching and the 39-rule diagnostic. With the retrain checkpoint, activation patching has a slightly larger set of eligible pairs: 992/992 OR pairs on seed 42 instead of 986/986, and 4,904/4,904 over all seeds instead of 4,898/4,898. The flip rate is 100% in both cases.

**Seeds of the diagnostic data.** For the reported per-rule accuracies, each rule used a data seed computed from Python’s string hash of the rule name. This hash is salted when the interpreter starts, so the seed changes between runs. The released code uses a deterministic seed computed from the rule index. We re-ran the full  $5 \times 39$ -rule diagnostic with this seed. All 195 rule-seed combinations stay above 99%. The worst rule reaches 99.31%, compared with the reported 99.15%. This lies within the re-seeding noise of  $\lesssim 0.2\text{pp}$  at 10,000 samples per rule. The pass/fail result of every rule stays the same.

**Natural per-rule frequency.** The diagnostic uses 10,000 samples for every rule, but the training distribution is not balanced across rules. Each training sample exercises one of the 39 rules:  $(\odot, \text{val}_{\text{ord}}, \text{val}_{\text{set}})$  for a binary operator and  $(\neg, \text{val}_{\text{ord}})$  for negation. Table S1 gives each rule’s frequency, counted by re-running the released data generator with the five training seeds at 2M samples per seed ( $N = 10,000,000$ ). An operand is Unknown when any input it depends on is Unknown, so with  $P_{\text{unk}} = 0.15$  it is Unknown with probability  $\approx 0.39$ , and False or True each with probability  $\approx 0.31$ . Three patterns stand out. First, the 16 mixed definite/Unknown rules are the rarest, at  $\approx 0.92\%$  each. This group contains all six absorption rules. Second, the 16 definite-definite rules, including  $F \vee F$ , occur at  $\approx 2.61\%$  each. Third, the Unknown-heavy rules are the most common: the  $U \odot U$  rules are at  $\approx 5.88\%$ , and the three NOT rules at

Table S1: Natural frequency of each of the 39 K3 rules in the four-domain training distribution, from raw counts over  $N = 10,000,000$  samples (5 training seeds  $\times$  2M). “Abs.” marks the six absorption rules.

| Rule                                     | Operands | Abs. | Freq. (%) | Rule                                | Operands | Abs. | Freq. (%) |
|------------------------------------------|----------|------|-----------|-------------------------------------|----------|------|-----------|
| <i>Mixed definite/Unknown (16 rules)</i> |          |      |           | <i>Definite-definite (16 rules)</i> |          |      |           |
| $U \wedge F$                             | U, F     | ✓    | 0.919     | $F \vee T$                          | F, T     | –    | 2.601     |
| $U \vee F$                               | U, F     | –    | 0.919     | $T \wedge F$                        | T, F     | –    | 2.602     |
| $T \vee U$                               | T, U     | ✓    | 0.920     | $T \rightarrow F$                   | T, F     | –    | 2.604     |
| $U \wedge T$                             | U, T     | –    | 0.920     | $F \wedge T$                        | F, T     | –    | 2.606     |
| $U \vee T$                               | U, T     | ✓    | 0.921     | $F \rightarrow F$                   | F, F     | –    | 2.606     |
| $F \vee U$                               | F, U     | –    | 0.921     | $T \rightarrow T$                   | T, T     | –    | 2.606     |
| $U \rightarrow T$                        | U, T     | ✓    | 0.921     | $F \rightarrow T$                   | F, T     | –    | 2.608     |
| $F \wedge U$                             | F, U     | ✓    | 0.922     | $F \wedge F$                        | F, F     | –    | 2.609     |
| $U \leftrightarrow F$                    | U, F     | –    | 0.922     | $F \vee F$                          | F, F     | –    | 2.609     |
| $T \leftrightarrow U$                    | T, U     | –    | 0.922     | $T \vee T$                          | T, T     | –    | 2.610     |
| $U \rightarrow F$                        | U, F     | –    | 0.922     | $T \leftrightarrow T$               | T, T     | –    | 2.611     |
| $F \rightarrow U$                        | F, U     | ✓    | 0.924     | $F \leftrightarrow T$               | F, T     | –    | 2.611     |
| $T \wedge U$                             | T, U     | –    | 0.924     | $T \leftrightarrow F$               | T, F     | –    | 2.613     |
| $F \leftrightarrow U$                    | F, U     | –    | 0.924     | $T \vee F$                          | T, F     | –    | 2.619     |
| $U \leftrightarrow T$                    | U, T     | –    | 0.925     | $T \wedge T$                        | T, T     | –    | 2.620     |
| $T \rightarrow U$                        | T, U     | –    | 0.927     | $F \leftrightarrow F$               | F, F     | –    | 2.625     |
| <i>Unknown-Unknown (4 rules)</i>         |          |      |           | <i>Unary NOT (3 rules)</i>          |          |      |           |
| $U \wedge U$                             | U, U     | –    | 5.860     | $\neg T$                            | unary    | –    | 6.137     |
| $U \leftrightarrow U$                    | U, U     | –    | 5.873     | $\neg F$                            | unary    | –    | 6.139     |
| $U \rightarrow U$                        | U, U     | –    | 5.878     | $\neg U$                            | unary    | –    | 7.722     |
| $U \vee U$                               | U, U     | –    | 5.880     |                                     |          |      |           |

6.1–7.7%. So overall accuracy can be high while the absorption rules still fail, and this is the main reason we report per-rule accuracy.

## C Architecture and Training Details

**Engines.** The Arithmetic, Order, Set and Logic Engines have 0.40M, 1.18M, 0.21M and 0.91M parameters. With the bridges (33K) and the output head with prototypes (17K), the total is 2.75M. The integers  $a$  and  $b$  take values from 1 to 20, and  $d$  from 0 to 20. The Arithmetic Engine fuses a number encoder with an operator embedding. The Order Engine has three parallel sub-MLPs (Global, Local and Event subspaces), and the Logic Engine has three (Conjunctive, Disjunctive and Implicative subspaces). In both, each pair of sub-MLP outputs is fused. The Set Engine encodes  $S$  and has its own Unknown embedding.

**Output head.** The prototypes are learned parameters, so they drift from the orthonormal initialization. At convergence, the F–T centroid cosine is  $-0.402 \pm 0.107$  at the Logic boundary. At inference, the verdict is  $\arg \max_v \cos(\mathbf{o}, \mathbf{p}_v)$ , and training uses dot-product logits. The two rules pick the same verdict whenever the three prototypes have equal norm. For the graph experiments, a *GNN extension* replaces the single-step Order Engine with a bidirectional message-passing network with shared MLPs. Its message-passing depth is set independently of chain length.

**Transformer baselines.** Each Transformer has 8 layers, 8 heads, width  $d=192$  and 3.58–3.64M parameters. The post-LN Transformer (3,641,859 parameters) is used for the K3 and probing comparisons. The pre-LN Transformer (3,582,147 parameters) is used for the chain pipeline, because the post-LN Transformer does not converge under the matched learning rate in the three-phase Gumbel-softmax training. Its Phase 1 accuracy stays at the class-prior level ( $\sim 43\%$ ) over several restarts. The pre-LN Transformer uses the same matched learning rate as the other step modules, not the tuned recipe of Section G. In the four-domain comparison, plateau-restart is disabled for both THEIA and the Transformer, so the Transformer with seed 123 fails because it uses up its 150-epoch budget. In the chain pipeline,

plateau-restart is enabled for all step modules through a shared Phase 1 implementation.

**Chain pipeline.** THEIA’s chain step module keeps all four engines, but it has a simpler output head and one Logic subspace. It has 1,508,096 parameters. The transition network has 4,803 parameters. In Phase 1, we train the step module alone on 2M single-step samples until local accuracy is  $\geq 99.9\%$ . A plateau-restart rule reinitializes the parameters when training stalls: accuracy  $< 90\%$  after 40 epochs, or no improvement for 30 epochs after reaching 90%. Seeds 42, 123, 256 and 777 reached the Phase 1 criterion on the first attempt, at epochs  $\{48, 44, 50, 47\}$ . Seed 999 triggered the plateau-restart and converged at epoch 49 of the second attempt, so it used about  $2\times$  the Phase 1 compute. In Phase 2, the step module is frozen, and the transition network is trained with teacher forcing. In 2 epochs it reaches 100% and learns the complete mod-3 addition table (9/9). In Phase 3, all parameters are unfrozen and trained end to end on 5-step chains ( $\sim 30$  epochs to 100%). The Gumbel temperature follows  $\tau = \max(0.1, 0.5 - 0.01 \text{ ep})$ . In the ResMLP  $4b \times 4d$  runs the only restarted seed is again seed 999. We take this shared seed as a coincidence.

**Chain head variant.** The four-domain head is Linear  $\rightarrow$  GELU  $\rightarrow$  Dropout  $\rightarrow$  LayerNorm, with dot-product logits in training and cosine similarity at inference. The chain head is Linear  $\rightarrow$  GELU  $\rightarrow$  Linear, followed by  $L_2$ -normalization and a scaled cosine similarity (temperature 10) to an axis-aligned prototype matrix. We made this change because the LayerNorm-then-cosine head gave less stable Gumbel-softmax gradients in preliminary Phase 3 runs. All chain step modules (flat MLP, ResMLP, pre-LN Transformer) use the same head. We chose this head, the Phase-3 learning rate and the temperature schedule in THEIA runs, and we did not tune them per baseline.

**Component ablations (main Section 5.1).** *Subspaces:* the single-MLP Logic Engine (2.49M parameters in total) reaches 99.97% overall accuracy and passes 12/12, with worst rule  $F \vee U$  at 99.73%. *Bridges:* the 2.72M model without bridges passes 12/12, with worst rule  $T \wedge U$  at 99.94%. *Prototype initialization:* random-normal initialization on seed 42 gives overall accuracy 99.97% (vs. 99.98% orthogonal). The first epoch with 12/12 (60) and the convergence epoch (100) are unchanged, and wall-clock time is  $\sim 15\%$  longer (10.68 vs. 9.24 min). All 12 rules reach  $\geq 99.89\%$ . *Class weights:* on seed 42, uniform weights (1, 1, 1) give overall accuracy 99.9457% (vs. 99.9427% with the default (1, 1, 2)), the same U-class accuracy (99.98%), and faster training in this run (6.73 vs. 10.53 min, timed in a separate paired run). The chain-pipeline weights (1, 2, 1) also give 12/12 (99.9102% overall, U-class 99.93%). *Unknown-rate shift:* trained at  $P_{\text{unk}}=0.05$  and tested at 0.50, the model reaches 99.9975% overall and passes 12/12. *Transitive chains:* the GNN extension reaches 99.99% from 5 to 50 hops, but 99.6% of chains are decided by the first non-identity edge (mean position about 0.5). Modular arithmetic on graphs (mod 5) needs global aggregation and could not be learned.

## D Additional Probing Results

**Extraction.** We take the engine outputs after all their internal computation. All probing uses the same 5 checkpoints, with 50K samples per checkpoint (data seed 999). The linear probes of main Table 5 use a linear SVM with feature standardization,  $C=1.0$  and a 70/30 split.

**HU baseline.** The four Unknown flags are independent with  $P_{\text{unk}} = 0.15$ , so the class “no input Unknown” has theoretical frequency  $0.85^4 \approx 0.522$ , and the majority baseline is 0.5220. The empirical frequency of this class is 0.5254. The empirical rate of true-U verdicts,  $P(VU=U) \approx 0.4027$ , stays below the bound  $P(VU=U) \leq P(HU=1) = 0.4780$ , because an absorption rule can give a definite verdict when an input is Unknown.

**Oracle reference.** The oracle reference is the accuracy of an ideal predictor that knows only whether the true verdict is  $U$ . Its value is  $0.4027 \times 1.0 + 0.5973 \times 0.5550 = 0.734$ , where 0.5550 is the True-class fraction among non- $U$  samples ( $N=50,000$ , data seed 999). No classifier can beat this value if its only verdict information is the  $U$ /non- $U$  split. It uses the full input, including the operator, so it is a benchmark and not a per-boundary ceiling.

**Nonlinear probe protocol.** Each MLP probe is  $128 \rightarrow 256 \rightarrow [256 \rightarrow 256]^{D-1} \rightarrow 3$ , with GELU layers and dropout 0.1. We train with AdamW ( $\text{lr } 10^{-3}$ , weight decay 0.01, cosine annealing), batch size 2048, for 40 epochs, on

Table S2: Verdict decodability and F–T centroid distance at each boundary (Euclidean view). Mean  $\pm$  std over 5 checkpoints, 50K samples each.

| Boundary     | SVM Acc.                            | F–T Distance                           | Sep. Ratio                      |
|--------------|-------------------------------------|----------------------------------------|---------------------------------|
| Arithmetic   | $0.609 \pm 0.001$                   | $0.146 \pm 0.015$                      | –                               |
| Order        | $0.672 \pm 0.002$                   | $3.130 \pm 0.302$                      | –                               |
| Set          | $0.697 \pm 0.002$                   | $5.967 \pm 0.643$                      | –                               |
| <b>Logic</b> | <b><math>0.999 \pm 0.000</math></b> | <b><math>272.738 \pm 37.070</math></b> | <b><math>1,898\times</math></b> |

a random 60/20/20 split (split seed 0). In each (seed, boundary, depth) cell, we pick the epoch with the best validation accuracy and evaluate on the test set once. For the Set MLP probe at  $D=2$ , verdict decodability is 70.4% under the 70/30 protocol and 69.8% under validation-based selection. The largest single cell in the seed  $\times$  boundary  $\times$  depth grid is 70.09% (Set,  $d=4$ , seed 999), 1.91pp below the stricter 72% threshold and 3.32pp below the oracle reference. The largest upstream 5-seed mean is  $69.92 \pm 0.12\%$  (Set,  $d=6$ ), 3.49pp below the oracle reference and 2.08pp below 72%. From  $d=2$  to  $d=6$  at width 256, the change is Arith +0.06pp, Order  $-0.02$ pp, Set +0.12pp and Logic  $-0.01$ pp.

**Width-512 audit.** A separate run with hidden width 512 and a stratified 60/20/20 split tests whether a wider probe finds more verdict information upstream. The highest upstream 5-seed mean is  $70.17 \pm 0.21\%$  (Set,  $d=6$ ). Going from width 256 to 512 adds about +0.22pp at the Set boundary at  $d=2$ . With both kinds of extra capacity, a 3.24pp margin to the 73.4% reference remains. The HU values of main Table 7 come from this run.

**HU ceilings.** At the Arithmetic and Set boundaries, the cross-seed std of the HU probe is 0.00% under the 70/30 protocol and at most 0.01% for the width-512 probe. This is structural: with 50K samples, the probe reaches the same ceiling on every seed. Under the 70/30 protocol, the Order boundary is +0.21pp above its empirical ceiling (width 512: +0.24pp). As explained in main Section 6.3, this is a finite-sample effect. The Logic boundary is 0.24 to 0.35pp below its 100% ceiling.

**NU-subset probes.** On the NU subset (true verdict T or F; approximately 8,972 samples per seed), all probes use a linear SVM with standardization,  $C=1.0$  and a 70/30 split. *Probe A* (Set-boundary vector) reaches  $60.13 \pm 0.33\%$ . *Probe B* (5-dimensional one-hot operator only) reaches  $48.40 \pm 0.00\%$ , deterministic given the operator. *Probe C* (Set-boundary vector plus operator, 133 dimensions) reaches  $61.93 \pm 0.16\%$ , only +1.80pp above Probe A. *Probe D* (operator identity from the Set-boundary vector) reaches  $20.25 \pm 0.38\%$ , at chance. For the per-operator NU majority, if  $\text{val}_{\text{ord}}$  and  $\text{val}_{\text{set}}$  were independent with  $P(T|\text{NU}) \approx 0.54$ , the AND majority would be  $\approx 0.71$  and the OR majority  $\approx 0.79$ . The observed value is 0.79 for both, because keeping only samples with verdict  $\neq U$  removes different samples for each operator.

**Euclidean geometry.** Table S2 gives the Euclidean F–T centroid distance at each boundary and the *separation ratio*, the Logic/Arithmetic ratio of F–T distances computed per seed and then averaged. The distance grows by  $1,898\times$  (95% bootstrap CI [1626, 2255], 1000 resamples). The absolute Euclidean scale depends on where LayerNorm is placed, so raw distances cannot be compared across architectures. For this reason the main article uses rescaling-invariant measures.

**Cosine observations.** The standard deviation of the F–T cosine grows from 0.001 (Arithmetic) to 0.124 (Set) and is 0.107 at Logic, so the Set and Logic representations vary more across seeds, though the direction of the pattern holds. At Logic, all off-diagonal cosines are negative (F–T =  $-0.402 \pm 0.107$ , F–U =  $-0.198 \pm 0.145$ , T–U =  $-0.240 \pm 0.070$ ). This is the expected effect of cross-entropy training, which widens the margin by pushing the centroids further apart than the orthogonally initialized prototypes.

**t-SNE.** Figure S1 gives a qualitative view. At every upstream boundary, samples with verdict U form visually separate regions, but True and False overlap. At the Logic boundary, all three verdict classes form distinct clusters.

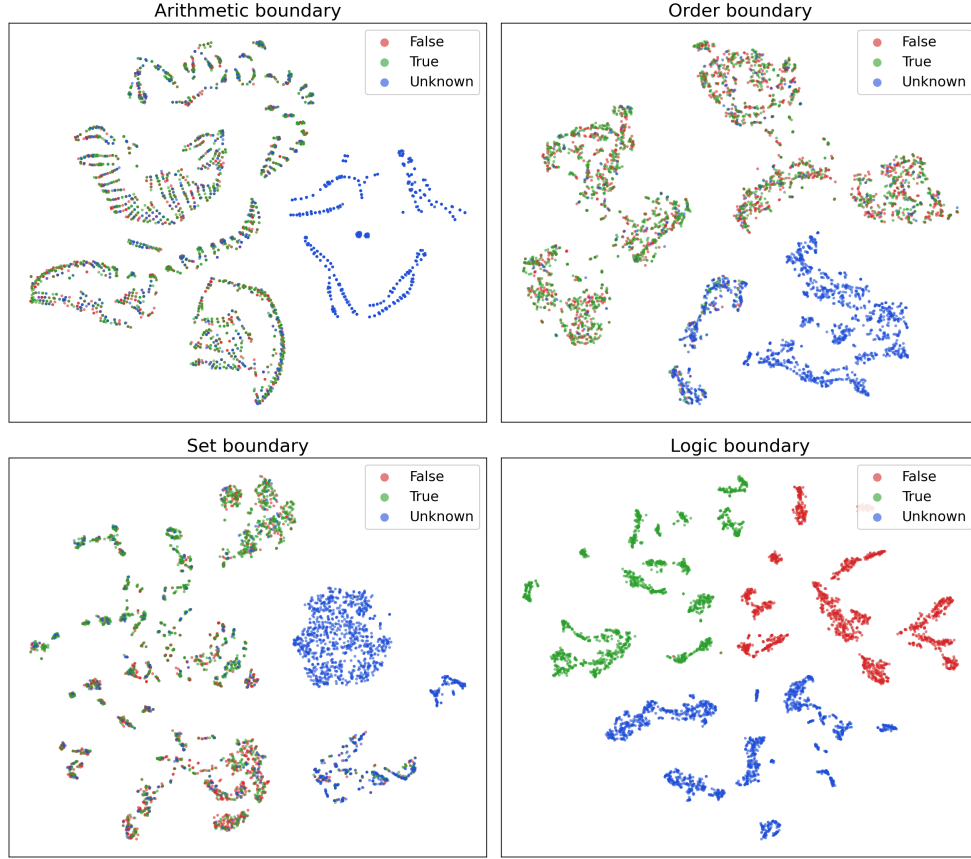


Figure S1: t-SNE projection of the boundary vectors of one checkpoint (seed 42), colored by the true verdict. t-SNE keeps local neighborhoods but not global distances, so cluster distances should not be read quantitatively. The Unknown class (blue) forms separate regions at every upstream boundary, while True (green) and False (red) overlap until the Logic boundary. The centroid analysis (main Figure 2) shows the same pattern without a nonlinear embedding.

**Transformer probing.** The layer-wise probes (main Table 11) use the post-LN Transformer trained under the matched protocol. Post-LN normalization may add to the contraction at layer 1.

## E Activation Patching Details

**OR pairs.** We use a converged seed-42 checkpoint and build 1,000 matched  $(T \vee F, U \vee F)$  pairs with data seed 12345. The two sides share  $a, b$ , the arithmetic operator, the set bits, the logic operator (OR) and the other Unknown flags (all False). The set bits give  $c \notin S$ , which fixes  $\text{val}_{\text{set}} = F$ , so  $c$ , both bridge outputs and  $\mathbf{v}_{\text{set}}$  are byte-identical on the two sides. The T-side sets  $d = \max(0, c-1)$  with “greater than or equal”, which gives  $\text{val}_{\text{ord}} = T$ . The U-side makes  $d$  Unknown, which gives  $\text{val}_{\text{ord}} = U$ . For each pair, we classify  $\text{OutHead}(\text{LogicEngine}(\mathbf{v}_{\text{ord}}^{(U)}, \mathbf{v}_{\text{set}}^{(T)}, \text{OR}))$ .

**Eligibility and results.** The unpatched baselines are correct on 1000/1000 T-side pairs and on 986/1000 U-side pairs. On the 986 pairs where both baselines are correct, the prediction flips from  $T$  to  $U$  in 986/986 cases. No pair stays at  $T$  and no pair goes to  $F$ . Over all 5 checkpoints, the aggregate flip rate is 4898/4898, and every seed has a flip rate of 100%. The smallest per-seed eligible count is 964, set by T-side baseline errors on that seed’s pairs.

Table S3: THEIA on the mod-3 chain task, trained on 5-step chains and tested on 5–500 steps ( $n = 5$  seeds).

| Chain Length | Accuracy (mean $\pm$ std)            | All $\geq 99\%$ |
|--------------|--------------------------------------|-----------------|
| 5 (in-dist.) | 100.00% $\pm$ 0.00%                  | yes             |
| 10           | 100.00% $\pm$ 0.00%                  | yes             |
| 50           | 100.00% $\pm$ 0.00%                  | yes             |
| 100          | 100.00% $\pm$ 0.00%                  | yes             |
| <b>500</b>   | <b>99.96% <math>\pm</math> 0.04%</b> | yes             |

**AND pairs.** We repeat the protocol with AND on matched  $(T \wedge T, U \wedge T)$  pairs, where  $c \in S$  fixes  $\text{val}_{\text{set}}=T$ . The prediction flips from  $T$  to  $U$  on 4719/4719 eligible pairs over the same 5 seeds. The smallest per-seed eligible count is 920, and the  $\text{v}_{\text{set}}$  byte-equality check passes on all pairs.

**Absorbent pairs.** For absorption rules such as  $T \vee U = T$ , the output does not depend on the absorbed operand, so patching  $\text{v}_{\text{ord}}$  gives no flip by definition. Studying how the network uses  $\text{v}_{\text{ord}}$  in these cases needs a different design, for example the Jacobian of the output with respect to  $\text{v}_{\text{ord}}$ . We leave this to future work.

## F Chain Pipeline Details

**Comparators.** We replace the step module and keep everything else identical. The comparators are (1) a flat MLP with hidden size 512, 3 layers and  $\sim 0.80\text{M}$  parameters, (2) a flat MLP with hidden size  $\sim 1024$ , 3 layers and  $\sim 2.75\text{M}$  parameters, (3) the ResMLP grid, and (4) the 8-layer pre-LN Transformer with 3,582,147 parameters.

**THEIA by chain length.** Table S3 gives THEIA’s accuracy for each chain length. All 5 seeds reach  $\geq 99.9\%$  at 500 steps: 99.90% (seed 42), 99.99% (123), 99.97% (256), 99.99% (777) and 99.96% (999). The 500-step std uses the sample standard deviation.

**Fisher tests.** We use a one-sided Fisher exact test of whether THEIA’s rate of seeds at  $\geq 99\%$  is higher than the comparator’s. We fixed this direction before the test. Pooled over all 20 ResMLP trials, the as-specified  $p$  is  $\approx 1.1 \times 10^{-3}$ , and the two-sided value is also  $1.1 \times 10^{-3}$ . The strict  $p$  is  $\approx 4.9 \times 10^{-4}$  with THEIA at 5/5, and  $\approx 1.7 \times 10^{-3}$  with THEIA’s matching strict count 4/4. Against each configuration ( $k/5$ ): for  $4\text{b} \times 4\text{d}$ ,  $8\text{b} \times 2\text{d}$  and  $8\text{b} \times 4\text{d}$  (1/5 each), the one-sided  $p$  is 0.024 and the two-sided  $p$  is 0.048; for  $4\text{b} \times 2\text{d}$  (0/5), 0.004 and 0.008.

**Depth and seed effects.** At a fixed budget, going from 4 to 8 blocks raises the mean 500-step accuracy by  $\approx 11\text{pp}$ . At fixed depth, changing the expansion has almost no effect ( $-1.34\text{pp}$  at depth 4,  $-0.02\text{pp}$  at depth 8). Seed 42 reaches  $\geq 99\%$  on both 8-block configurations (99.27% and 99.12%) but fails on both 4-block configurations (90.49% and 53.27%). So the ResMLP results depend on the combination of seed and depth.

### Worst-case per-seed accuracy at 500 steps.

- THEIA: min 99.90%; 5/5 seeds  $\geq 99.9\%$ .
- Pre-LN Transformer: min 98.85% (range 98.85–99.66%); 3/5 seeds  $\geq 99\%$ .
- ResMLP  $8\text{b} \times 2\text{d}$ : min 95.55% (range 95.55–99.27%); 1/5 seeds  $\geq 99\%$ .
- ResMLP  $8\text{b} \times 4\text{d}$ : min 94.34%; 1/5 seeds  $\geq 99\%$ .
- ResMLP  $4\text{b} \times 4\text{d}$ : min 75.27%; 1/5 seeds  $\geq 99\%$  (the passing seed used a plateau-restart).
- ResMLP  $4\text{b} \times 2\text{d}$ : min 53.27%; 0/5 seeds  $\geq 99\%$ .

Cross-seed std is a secondary metric, because means near 100% force a low std. The ResMLP std values and their ratios to THEIA’s exact std (0.0370%) are 1.43% ( $8\text{b} \times 2\text{d}$ ,  $39\times$ ), 2.03% ( $8\text{b} \times 4\text{d}$ ,  $55\times$ ), 9.17% ( $4\text{b} \times 4\text{d}$ ,  $248\times$ ) and 19.16% ( $4\text{b} \times 2\text{d}$ ,  $518\times$ ).

Table S4: Multi-hop SQL-NULL: per-class verdict accuracy and the two NULL-safety error types (mean over 6 depths  $\times$  3 seeds). F-abst. is false-abstain (definite  $\rightarrow$  U).

| Step module | False  | True   | Unknown | Unsafe | F-abst. |
|-------------|--------|--------|---------|--------|---------|
| THEIA       | 100.00 | 100.00 | 100.00  | 0.00   | 0.00    |
| Transformer | 99.97  | 99.99  | 99.99   | 0.01   | 0.01    |
| Flat MLP    | 89.42  | 87.58  | 94.49   | 5.51   | 3.13    |

Table S5: Tuned post-LN Transformer, seed 42: time to two milestones. THEIA ref is a 5-seed mean under the matched protocol. In the 12/12 row, THEIA ref is THEIA’s time to the full Kleene-aware criterion, and the Transformer value is its first 12/12 crossing.

| Milestone             | Tuned Transformer       | THEIA ref | Ratio            |
|-----------------------|-------------------------|-----------|------------------|
| Kleene 12/12          | <b>28.9 min</b> @ ep 60 | 7.93 min  | $\sim 3.6\times$ |
| Overall $\geq 99.9\%$ | 41.7 min @ ep 84        | 5.7 min   | $\sim 7.3\times$ |

**A hypothesis for the spectrum.** Self-attention in the Transformer and domain-separated subspaces in THEIA may keep gradient pathways apart during discretized end-to-end training. This would prevent interference that a residual-only structure reduces only in part, even when its depth is doubled. This is a hypothesis for future testing.

**SQL-NULL failure analysis.** Table S4 splits the SQL-NULL verdict error by class. The flat MLP is lowest on the definite verdicts (True 87.6%, False 89.4%). It gives an unsafe verdict on 5.5% of columns whose correct value is Unknown, and it answers U on 3.1% of definite columns (false-abstain). Its first per-hop error is at hop  $\approx 18$  on average, compared with hop  $\approx 77$  for the Transformer and  $\approx 93$  for THEIA. THEIA and the pre-LN Transformer keep every class at or above 99.97%.

## G Tuned Transformer and Wall-Clock Details

**Matched protocol.** The post-LN Transformer runs on 8 seeds {42, 123, 256, 777, 999, 31415, 27182, 14142} and reaches 12/12 on 7 of them in  $51.5 \pm 11.0$  min (range 34–63 min). THEIA needs 2.88 min per million parameters and the Transformer 14.1, a ratio of about  $4.9\times$ . We report the unnormalized  $6.5\times$  as the main number, because total wall-clock time is the time a user waits.

**Overall-accuracy criterion.** On the initial 5 seeds, THEIA reaches overall validation accuracy  $\geq 99.9\%$  in  $5.7 \pm 1.4$  min ( $n=5$ ) and the Transformer in  $39.5 \pm 7.1$  min ( $n=4$ ; seed 123 does not converge). The ratio is  $7.0\times$ , and it grows to  $9.8\times$  at  $\geq 99.5\%$  and  $12.2\times$  at  $\geq 99.0\%$ .

**Tuned recipe.** We rerun the same post-LN Transformer with the Transformer-standard recipe. Data generation, class weights, gradient clipping (1.0), CUDA determinism settings and the diagnostic code stay the same. Only the optimizer settings change: peak learning rate  $10^{-4}$  (vs.  $10^{-3}$ ), AdamW  $\beta = (0.9, 0.98)$  (vs.  $(0.9, 0.999)$ ), linear warmup over 5 epochs then cosine decay, weight decay 0.01 and batch size 2048. We treat this bundle as one variable.

**Seed 42.** The tuned run takes 42.4 min under the 150-epoch cap. The final per-class validation accuracies are 99.79% (False), 99.90% (True) and 99.99% (Unknown), and the final diagnostic passes 12/12 with worst rule  $U \vee T$  at 99.53%. Table S5 gives two milestones. At epoch 60, all 12 rules first exceed 99%:  $F \wedge U$  99.80,  $T \wedge U$  100.00,  $U \wedge F$  99.06 (worst),  $U \wedge T$  99.82,  $T \vee U$  100.00,  $F \vee U$  99.87,  $U \vee T$  99.33,  $U \vee F$  99.78,  $F \rightarrow U$  99.81,  $T \rightarrow U$  100.00,  $T \leftrightarrow U$  100.00,  $F \leftrightarrow U$  100.00.

**Seeds 123 and 256.** Both pass 12/12, with overall accuracy  $\geq 99.93\%$  and  $\geq 99.90\%$ . Seed 123 reaches overall  $\geq 99.9\%$  at epoch 73 (34.2 min), and seed 256 at epoch 98 (43.3 min). Over the three seeds, the time to overall 99.9%

is  $39.7 \pm 5.0$  min, and the time to 12/12 is  $28.8 \pm 1.7$  min. Seed 123, the only seed that fails under the matched protocol, converges under the tuned recipe.

**Full 39-rule coverage.** We use 10,000 samples per rule per seed and standard K3 ground truth ( $U \rightarrow T = T$ ,  $U \rightarrow F = U$ ), imported from the Transformer training script. All three tuned checkpoints pass 39/39 at  $> 99\%$  (117/117, grand mean 99.89%). The worst rules are  $F \leftrightarrow T$  on seed 42 (99.48%) and  $T \rightarrow F$  on seeds 123 and 256 (both 99.50%), and the 3 NOT rules are all at  $\geq 99.87\%$ .

**Order of milestones.** The tuned Transformer reaches 12/12 before overall 99.9% (epoch 60 vs. 84). The matched-protocol Transformer shows the opposite order. The most likely reason is the lower learning rate, which makes rare configurations stable earlier, while the last  $\sim 0.1\%$  of overall accuracy takes longer. The  $6.5\times$  gap uses the full criterion for both models, and the  $3.6\times$  gap uses the Transformer’s first 12/12 crossing, so part of the narrowing may come from the change of milestone.

**Limits.** The five settings of the recipe are changed together, and a stronger recipe could narrow the gap further.

**The recipe applied to THEIA.** We also apply the Transformer-standard recipe (with gradient clipping 1.0) to THEIA. THEIA then reaches 12/12 on all 5 seeds {42, 123, 256, 777, 999} in  $9.88 \pm 1.42$  min per seed (range [7.71, 11.40] min), compared with 7.93 min under its development default.

## H Reproducibility, Scope and Ethics

**Compute and settings.** All experiments use PyTorch with CUDA on a single NVIDIA RTX 5080. Seeds are {42, 123, 256, 777, 999}, and the Transformer baseline adds {31415, 27182, 14142}. The tuned-Transformer runs use seeds {42, 123, 256}. The matched protocol uses AdamW (lr  $10^{-3}$ , weight decay 0.01), cosine annealing and FP16 mixed precision. Four-domain training uses 2M samples, batch size 4096 and 80–120 epochs depending on the stopping criterion ( $\sim 8$  minutes per seed on average). Multi-hop training uses 1M samples, batch size 2048 and  $\sim 50$  epochs ( $\sim 15$  minutes). Chain training uses 2M single-step and 500K chain samples ( $\sim 16$  minutes per seed).

**Train/test split.** We draw the 2M four-domain samples at random and split them 80/20 (1.6M/400K) by a random permutation of samples, not of input configurations. The configuration space has on the order of  $3.4 \times 10^{13}$  elements, but a field marked Unknown is replaced by a learned Unknown vector, so the effective input space is smaller, and inputs with Unknown fields can occur in both sets. We did not measure how many test inputs also occur in the training set.

**Scope notes.** (i) The mod-3 chain uses a minimal 3-state space. Richer state spaces (mod- $k$  for  $k > 3$ , automaton traces, larger learned codebooks) are future work. (ii) We did not run the 12-rule diagnostic on the flat MLPs of the chain pipeline; their Phase 1 match with THEIA is a training-accuracy match. (iii) The four benchmark domains map one-to-one onto the four engines by design. The misalignment test (main Section 8.3) covers four-domain accuracy and rule coverage, but not chain reliability, and overlapping or continuous domain boundaries remain untested. (iv) The chain ablations change the step module but not the domain–engine map.

**Future work.** (a) Deeper nonlinear probes, and layer-wise probes of the tuned Transformer. (b) Richer chain states. (c) First-order and other non-classical logics (fuzzy, paraconsistent). (d) A tuned-THEIA versus tuned-Transformer comparison, a larger ResMLP grid ( $\geq 12$  blocks, other normalization placements), and a post-LN Transformer in the chain pipeline. (e) Real database benchmarks and medical or legal reasoning where Unknown values arise naturally.

**Ethics and broader impact.** This work uses only synthetic data. It does not involve human subjects, private data, or deployment in safety-critical settings. A possible negative impact is that readers take synthetic K3 performance as evidence that a neural system is safe for medical, legal or database decisions without domain validation. We state the benchmark scope and report failure modes (global graph computation, numeric extrapolation) to reduce this risk.