ARTI: Detecting and Routing Reasoning Types from Embedding Geometry

Ariel Sandez

ORCID: 0009-0004-7623-6287

February 26, 2026

Draft v2.1

Abstract

We show that different reasoning types leave distinct geometric signatures in transformer embedding spaces, detectable by a lightweight classifier with only 20K parameters. Our **Active Reasoning Type Identifier (ARTI)** classifies 10 reasoning types at 84.2% accuracy (8.4× chance) from embedding geometry alone—no access to input text. Building on this, we introduce **ARTI-routed scoring**, a mechanism that soft-blends cosine similarity and direct classification via frozen ARTI type probabilities, achieving **StrategyQA 81.0% (+32.6pp over baseline)**—the first non-math benchmark with genuine learning in the Continuous Operator Factorized Reasoning Network (CO-FRN) architecture—while preserving GSM8K at 48.1%. We further discover that pre-factorization embeddings retain **3.5× more type-discriminative signal** than post-factorization features (76.7% vs 21.9%), revealing a fundamental tension in factorization-based reasoning architectures. Finally, a three-experiment falsification chain (E16–E17–E18) establishes that the frozen encoder is **structurally necessary**: any unfreezing of a shared transformer destabilizes the answer embedding space for generic-label scoring, regardless of gradient routing strategy. All experiments use 3-seed evaluation on full test sets (N=203–1,603).

1. Introduction

Our prior work on structural semantic superposition [18] established that reasoning operations form a continuous ~10D manifold in transformer embedding spaces, and that a Continuous Operator Factorized Reasoning Network (CO-FRN) [19] can exploit this structure: GSM8K accuracy jumps from 25% (chance) to 47.9% (+23pp) with only ~726K trainable parameters atop a frozen GPT-2 encoder. Zero-shot transfer to SVAMP reaches 100%.

But everything beyond math fails. ARC Challenge, StrategyQA, and FOLIO all collapse to chance levels. The CO-FRN work [19] documented two pathological symptoms: (a) **uniform attention**—all 16 manifold anchors receive identical weight $w_i = 1/16$ (entropy ratio 0.99996), and (b) **constant predictions** on binary/ternary benchmarks.

This paper asks three questions:

1. **Can reasoning types be detected from geometry?** If reasoning types leave geometric signatures, can we build a classifier that reads them?
2. **Can type detection improve reasoning?** If we know the reasoning type, can we route to better scoring mechanisms?
3. **What are the architectural boundary conditions?** Under what conditions does type-routed scoring succeed or fail?

Our answers: (1) Yes—ARTI achieves 84.2% on 10 types from embeddings alone. (2) Yes—ARTI-routed scoring achieves StrategyQA 81.0% by blending cosine and direct classification. (3) The frozen encoder is structurally necessary; any unfreezing destroys the effect.

The paper is organized as follows: type detection (Section 2), the pre-factorization discovery (Section 3), cosine scoring collapse diagnosis (Section 4), uniform-attention resolution (Section 5), the ARTI-routed scoring breakthrough (Section 6), and the frozen-encoder proof (Section 7).

2. ARTI: Reasoning Types Have Geometric Signatures

2.1 Task Definition

Given an input text encoded by a sentence transformer into a 384-dimensional embedding, classify which of 10 reasoning types the text employs—without access to the raw text. The 10 types span the spectrum of human reasoning:

Type	Geometric Signature	Example
Deduction	Tight manifold cluster, high norm	“All mammals breathe air. Whales are mammals. Therefore…”
Induction	Far-left PCA cluster, pattern features	“Every observed swan was white, so…”
Abduction	Distinct cluster, high separability	“The grass is wet; it probably rained”
Analogy	Partially separated in manifold	“Electricity flows like water through pipes”
Counterfactual	Strong geometric signal	“If Napoleon had won at Waterloo…”
Conservation	Moderate separation	“Energy before = energy after”
Decomposition	Moderate separation	“Break the problem into three parts…”
Physical Cause	Domain-general residual	“Heating metal causes expansion”
Behavioral Cause	Geometric sub-cluster	“Fear of failure leads to procrastination”
Systemic Cause	Strong trajectory signal	“Deforestation drives climate feedback loops”

2.2 Architecture

ARTI is a two-layer MLP operating on manifold-projected embeddings:

```
Input: sentence embedding [384D]
  → ManifoldProjection [384 → 10D, frozen]
  → MLP: Linear(10, 64) → ReLU → Dropout(0.3) → Linear(64, 10) → Softmax
Output: type probability distribution [10D]
```

Total parameters: **11,342** (v1) / **8,890** (v2 trajectory variant) / **~20,232** (ensemble: both models combined).

The manifold projection maps from the full embedding space to the ~10D reasoning manifold identified in [18] (silhouette = 0.33, domain consistency = 0.95). ARTI classifies *within* this manifold rather than on raw embeddings.

2.3 Training Data

ARTI’s training set consists of 7,500 sentences (750 per type) generated via template-based synthesis. For each reasoning type, we defined 15–25 sentence templates encoding the type’s characteristic linguistic and structural patterns (e.g., “All X are Y. Z is X. Therefore Z is Y” for Deduction; “Every observed X had property Y, suggesting…” for Induction). Templates were instantiated with diverse domain content (physics, economics, biology, everyday reasoning) to ensure the classifier learns geometric signatures rather than domain keywords. All sentences were encoded using all-MiniLM-L6-v2 (384D) at generation time; ARTI never sees raw text during training or inference.

Class balance was enforced at exactly 750 examples per type. Earlier versions (v1, 984 examples) suffered from severe class imbalance (89% in 3 types, 0% in 3 others), which was the primary bottleneck resolved in v2.

2.4 Evolution and Ablations

ARTI evolved through systematic improvements, each addressing a specific bottleneck:

Version	Types	Samples	Accuracy	vs Chance	Key Change
v1	8	984	54.0%	4.3×	Proof of concept
v2	8	6,000	70.8%	5.7×	Class balance fix (89% skew → 12.5% each)
v3	10	7,500	71.5%	7.15×	Cause-Effect split into 3 sub-types
Trajectory	10	7,500	77.9%	7.8×	Clause-level delta vectors
v1 retrained	10	7,500	83.5%	8.4×	100 epochs (was 30)
v2 retrained	10	7,500	82.9%	8.3×	100 epochs (was 30)
Ensemble	**10**	**7,500**	**84.2%**	**8.4×**	Max-confidence v1+v2

Oracle upper bound: 89.0% (perfect selection between v1 and v2). Agreement rate: 80.8%; when they agree, accuracy is 95.9%.

Key bottleneck progression:

- **v1 → v2:** Class imbalance was the primary bottleneck. Three types at 0% (Induction, Conservation, Decomposition) jumped to 76–97% after balancing.
- **v2 → v3:** The "Cause-Effect" category was a catch-all whose generic causal language ("because", "therefore") appears in every domain. Splitting into Physical/Behavioral/Systemic Cause resolved 2 of 3 sub-types.
- **v3 → ensemble:** Extended training (30 → 100 epochs) and max-confidence ensembling provided the final +12.7pp.

2.5 Per-Type Results (Ensemble)

Type	N	Ensemble	Best Single	Oracle	vs Chance
Counterfactual	155	**99.4%**	v1: 99.4%	100%	9.9×
Analogy	146	**99.3%**	v2: 99.3%	99.3%	9.9×
Systemic Cause	184	**98.9%**	v1: 98.9%	98.9%	9.9×
Abduction	150	**98.7%**	v1/v2: 98.0%	98.7%	9.9×
Induction	140	**96.4%**	v1: 96.4%	97.1%	9.6×
Conservation	144	**90.3%**	v1: 88.9%	92.4%	9.0×
Decomposition	129	**86.8%**	v1: 86.0%	89.9%	8.7×
Behavioral Cause	146	**76.7%**	v1: 74.7%	81.5%	7.7×
Deduction	146	**56.8%**	v1: 54.1%	68.5%	5.7×
Physical Cause	160	**38.8%**	v2: 49.4%	63.1%	3.9×

Five types exceed 90%. The weakest type (Physical Cause, 38.8%) is still 3.9× above chance, and its difficulty is interpretable: physics sentences simultaneously invoke conservation, decomposition, induction, and causality—making Physical Cause the domain-general residual, consistent with [18]'s finding that the manifold is domain-invariant (consistency = 0.95).

2.6 What ARTI Demonstrates

1. **Reasoning types have genuine geometric signatures** in sentence-transformer embedding spaces. A 20K-parameter classifier achieves 84.2%—this is a signal detection result, not a brute-force memorization.
2. **The signal is in the manifold structure**, not in surface features. ARTI never sees raw text.
3. **Abstract reasoning types are easier to detect than empirical ones**: Counterfactual, Analogy, and Systemic Cause (>98%) involve distinctive structural transformations; Physical Cause (38.8%) is a catch-all.
4. **Abduction outperforms keyword heuristics**: At 97.3% (v3 single-model; 98.7% with ensemble), geometric classification beats keyword matching (96.7%)—the first type where this occurs.

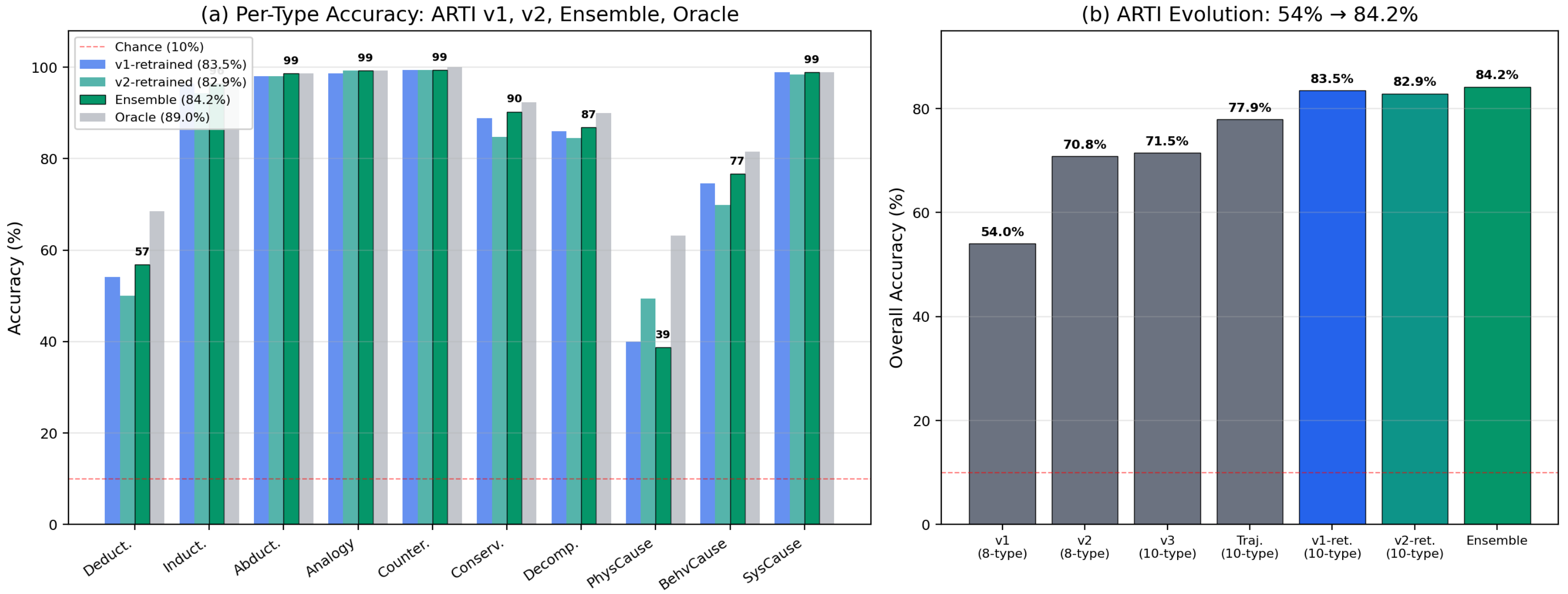


Figure 1. (a) Per-type accuracy across ARTI v1, v2, ensemble, and oracle. (b) ARTI evolution from 54% to 84.2%.

3. Pre-Factorization Embeddings Retain 3.5× More Type Signal

3.1 The Discovery

The CO-FRN architecture [19] processes inputs through a factorization pipeline: raw sentence embeddings (s0, 256D) → mutual-information-based factorization → structural features (128D) → manifold coordinates (10D). We tested type classification accuracy at each stage:

Feature Source	Dimensions	Type Accuracy	Ratio to Best
s0 (pre-factorization)	**256D**	**76.7%**	**1.0×**
Structural (post-MI)	128D	21.9%	0.29×
Manifold coordinates	10D	18.8%	0.25×
Random baseline	—	16.7%	0.22×

Post-factorization features are barely above random. The factorization layer, by design, strips domain-correlated signal to produce domain-invariant structural representations—but reasoning type *is* domain-correlated information.

3.2 The Architectural Tension

This creates a fundamental design conflict:

- **Factorization** wants domain-invariant features: Conservation should look the same whether applied to energy (physics), probability (statistics), or money (accounting).
- **Type-aware routing** needs domain-correlated features: knowing the input involves physics (vs. logic vs. economics) helps select the right scoring mechanism.

Implication: Any type-aware extension of factorization-based architectures must tap into the representation *before* factorization (s0), not after. This is why ARTI operates on pre-factorization embeddings, and why the ARTI-routed scorer in Section 6 uses s0 as its routing signal.

3.3 Per-Type Breakdown (TypeClassifier on s0)

The TypeClassifier is a separate classifier from ARTI—it operates on the 256D s0 embeddings within the CO-FRN pipeline rather than on ARTI's 384D manifold-projected embeddings, and uses the original 6-type taxonomy (before the Cause-Effect split):

Type	Accuracy
Analogy	97.9%
Induction	88.6%
Counterfactual	83.6%
Conservation	77.8%
Cause-Effect	66.9%
Deduction	55.5%

The ordering differs from ARTI's: Analogy and Induction benefit most from the richer pre-factorization representation, while Deduction remains the hardest type regardless of input features.

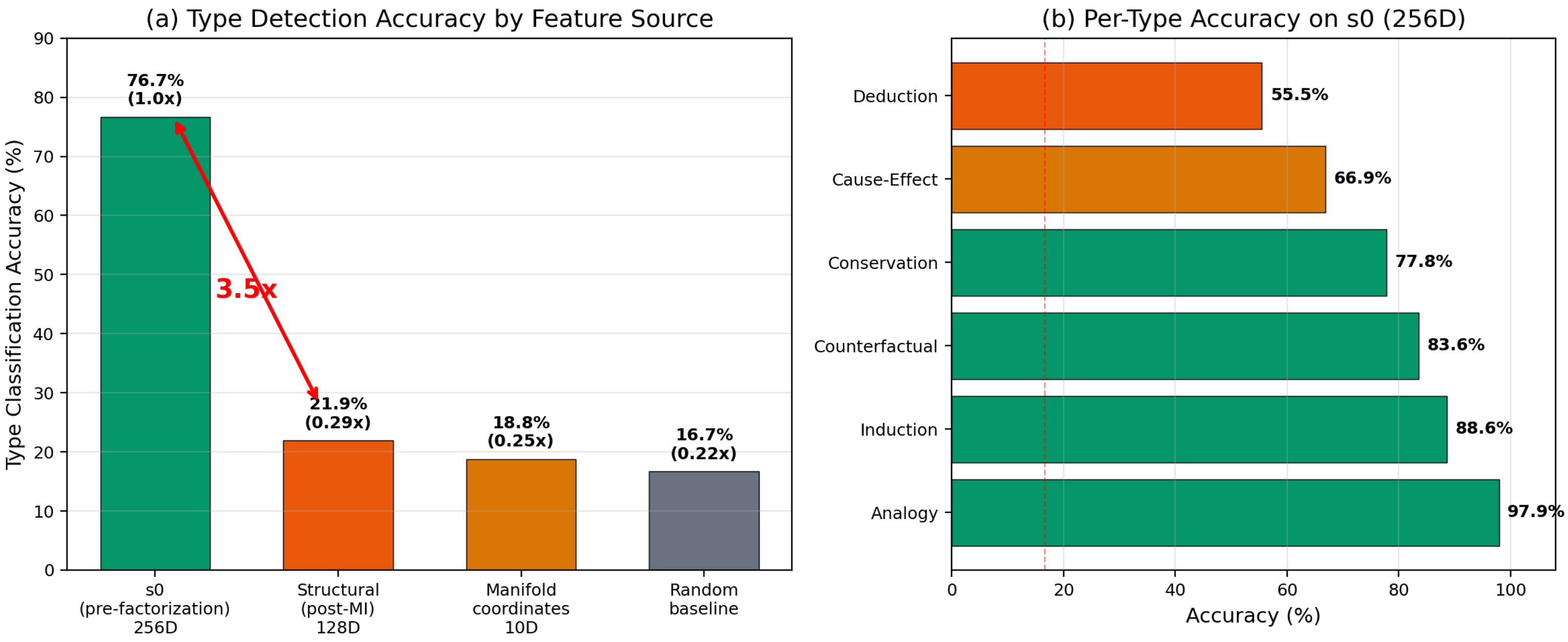
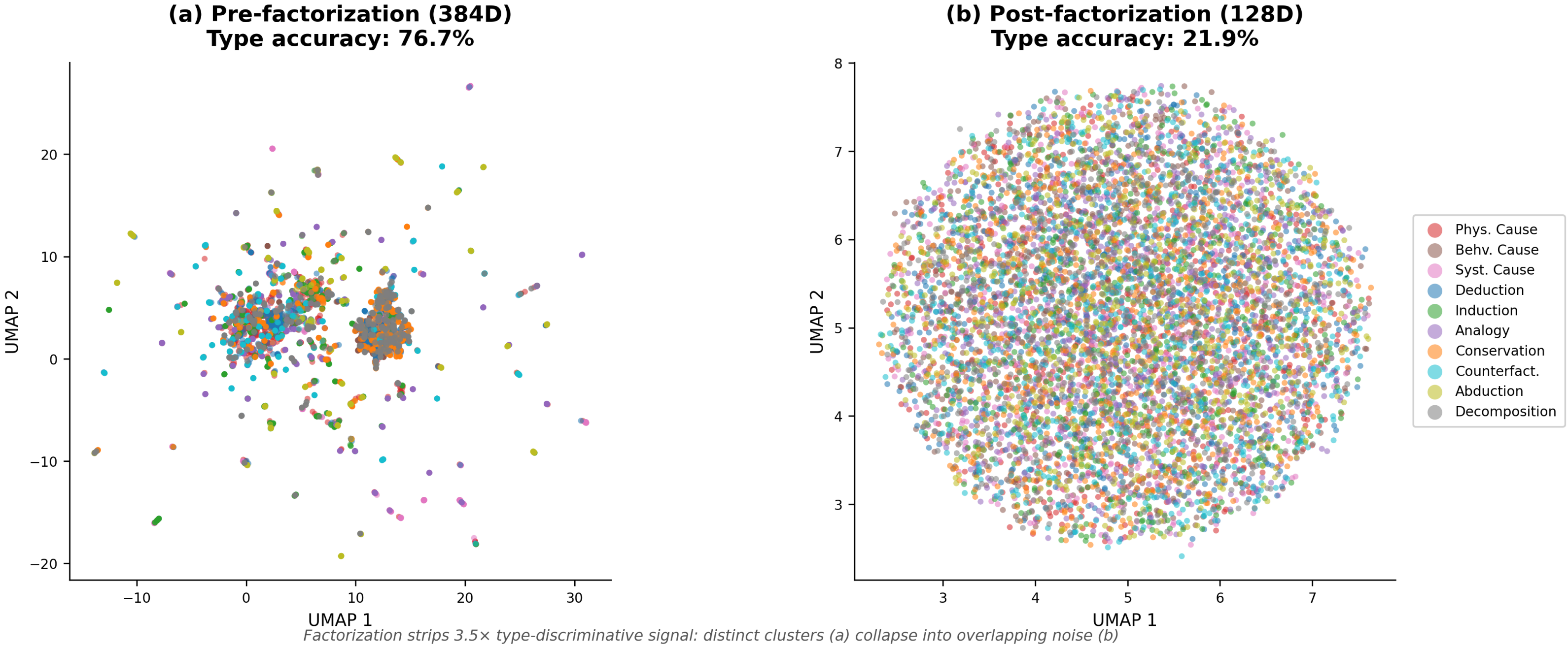


Figure 2. (a) Type detection accuracy by feature source showing the 3.5× pre-factorization gap. (b) Per-type accuracy on s0 embeddings.



Factorization strips 3.5× type-discriminative signal: distinct clusters (a) collapse into overlapping noise (b)

Figure 3. UMAP projection of 7,500 reasoning traces colored by type. (a) Pre-factorization embeddings (384D) show distinct type clusters—Deduction/Decomposition form a tight central mass, Induction and Analogy separate clearly. (b) Post-factorization features (128D) collapse into an undifferentiated noise ball, visualizing the 3.5× signal loss from MI-based factorization.

4. Diagnosing the Non-Math Failure: Cosine Scoring Collapse

Before presenting our solutions (Sections 5–6), we must understand why CO-FRN succeeds on math but fails on everything else.

4.1 The Symptom

On StrategyQA (Yes/No answers) and FOLIO (True/False/Unknown), CO-FRN produces **constant predictions**—the same answer for every input, regardless of the transformed representation. This yields exactly chance accuracy (50% for binary, 33% for ternary). The pattern was first observed in the unfrozen-encoder experiments (E16), where StrategyQA collapsed to 54.0% with zero variance across all seeds—identical predictions for every input. A gradient diagnostic then traced the root cause not to the encoder or the manifold, but to the scoring function itself.

4.2 The Root Cause

CO-FRN scores answers via cosine similarity: `score_k = cos(transformed, answer_k) / τ`. The `answer_k` vectors are sentence-transformer encodings of the answer labels.

Benchmark	Answer Labels	Pairwise Embedding Similarity	Score Range
GSM8K	“$42”, “15 apples”, “3 hours”	Low (content-rich, semantically distinct)	~2.0+
StrategyQA	“Yes”, “No”	~0.99 (near-identical)	~0.07
FOLIO	“True”, “False”, “Unknown”	~0.98 (near-identical)	~0.09

When answer embeddings are nearly identical, the cosine term collapses: all answers receive approximately the same score regardless of the transformed state. The argmax picks whichever class has marginally higher similarity to the general population of transformed states—consistently, for every input.

This is not a training problem or an optimization failure. It is a representational degeneracy: generic labels (“Yes”, “No”) carry no discriminative content in embedding space. Cosine scoring is structurally incapable of handling them.

4.3 Implication

The model needs a scoring mode that does not rely on answer-embedding differences for benchmarks with generic labels, while preserving cosine scoring for content-rich answers (where it provides +23pp on GSM8K). This motivates a **hybrid scoring architecture** that selects the appropriate mode per input.

5. Resolving the Uniform-Attention Paradox

The CO-FRN work [19] reported that all 16 manifold anchors receive identical weight (entropy ratio 0.99996). Resolving this was a prerequisite for subsequent experiments. This section summarizes the progressive diagnosis that identified three independent causes and fixed each one.

5.1 Three Causes, Three Fixes

Fix	Experiment	Cause Identified	Entropy Ratio	Max Anchor Weight
[19] baseline	—	—	0.99996	0.0625 (uniform)
Cosine scoring	E10	MLP scorer ignored tree output entirely (0/100 prediction changes with randomized input)	0.9995	0.0625
Entropy annealing	E11	Entropy regularization applied at constant strength instead of annealing; manifold projections initialized at ~0.025 magnitude	0.996	0.08
Remove sqrt(d) scaling	**E12**	Softmax divided by sqrt(16)=4, suppressing weight differentiation	**0.937**	**0.16**

5.2 The Entropy Trajectory

```
E10:  0.99996  →  0.9995   (scoring collapse fixed, but attention still uniform)
E11:  0.9995   →  0.996    (training artifact fixed, first anchor differentiation)
E12:  0.996    →  0.937    (scaling bottleneck removed, 3/4 benchmarks pass <0.95)
```

Each fix was validated independently with 3-seed full-benchmark evaluation. The fixes are additive: each subsequent experiment includes all prior fixes.

After E12, anchor entropy passes the <0.95 criterion on 3 of 4 benchmarks (ARC: 0.921, StrategyQA: 0.934, FOLIO: 0.932; GSM8K: 0.961). Maximum anchor weight doubles from 0.08 to 0.16, indicating genuine specialization.

5.3 Systematic Elimination (E6–E9)

Prior to identifying the three causes above, we systematically eliminated four alternative hypotheses:

Experiment	**Hypothesis Tested**	**Result**	**Verdict**
E6: Rank ablation	Operator capacity is the bottleneck (ranks 4→64)	Entropy = 1.0 at all ranks	**Eliminated**
E7: Rule injection	Domain knowledge bypasses uniform attention (75 rules, 7 domains)	Zero lift across all conditions	**Eliminated**
E8: Controller tuning	Better type routing closes the gap (+7.9pp TypeClassifier)	Zero benchmark lift	**Eliminated**
E9: Pipeline diagnosis	Identify binding constraint	tree_scorer ignores best_hyp (0/100 prediction changes)	**Root cause found**

E9 was decisive: replacing the tree's output (best_hyp) with zeros or random noise produced 0/100 prediction changes—the MLP scorer had learned to score from answer encodings alone, decoupling from the reasoning pipeline entirely. This motivated the cosine scoring replacement in E10.

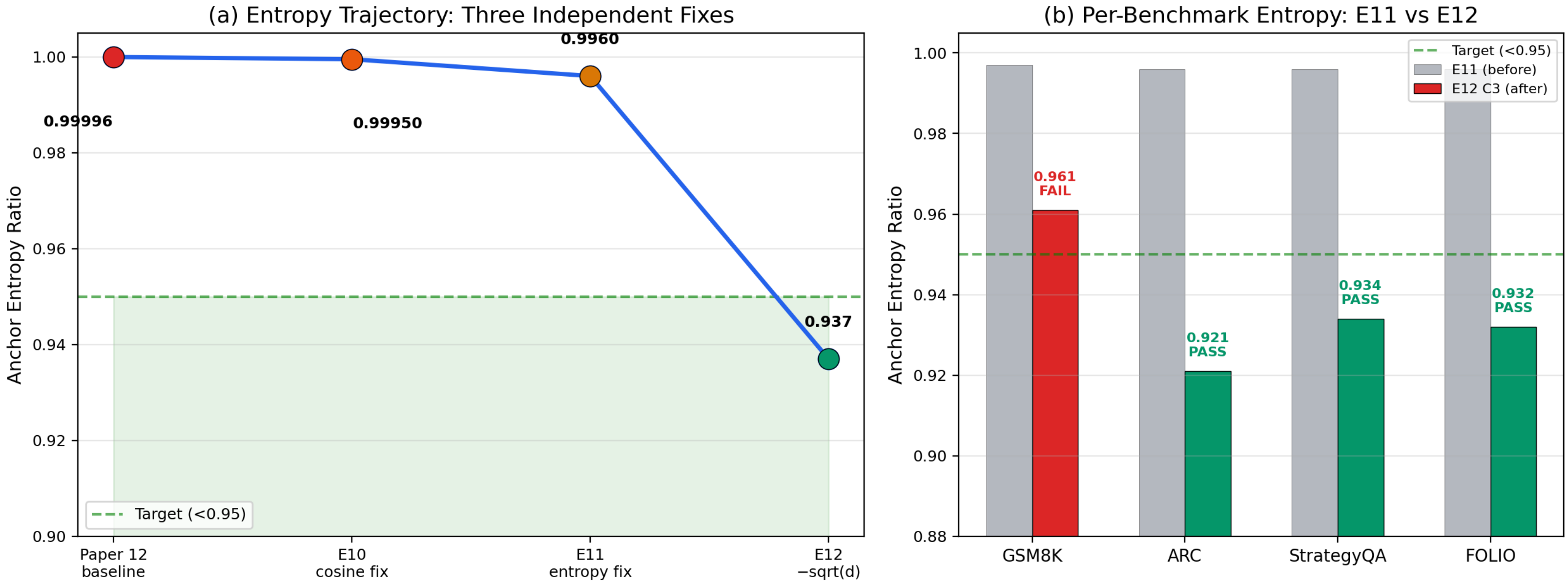


Figure 4. (a) Entropy trajectory from 0.99996 to 0.937 across three independent fixes. (b) Per-benchmark entropy at E11 vs E12, with 3/4 benchmarks passing the <0.95 criterion.

6. ARTI-Routed Scoring: Breaking the Non-Math Ceiling

6.1 Architecture

With uniform attention resolved (Section 5) and the cosine collapse diagnosed (Section 4), we introduce the ARTIRoutedScorer, which soft-blends two complementary scoring modes via ARTI type probabilities:

```
Input question → s0 [B, 256]
                   |
                   |--→ ARTI adapter (Linear 256→384) → ARTI (frozen, 84.2%)
                   |                                       |
                   |                                   type_probs [B, 10]
                   |                                       |
                   |                                   Router: Linear(10,1) + Sigmoid
                   |                                       |
                   |                                   alpha [B, 1]
                   |
                   +--→ Factorize → Manifold → Transform → transformed [B, 256]
                                                              |
                               +-------------------------------+
                               |                               |
                   cosine_scores = cos(transformed, ans) / τ   direct_scores = MLP(transformed)
                               |                               |
                               +-------- alpha * cos + (1-alpha) * direct ----------→ final_scores
```

- **Cosine mode**: `cos(transformed, answer_encoding) / τ` —for content-rich answers (math, science)
- **Direct classification mode**: `MLP(transformed) → logits` —for generic labels (Yes/No, True/False), bypasses answer embeddings entirely
- **Router**: `sigmoid(Linear(arti_probs)) →` per-example blend weight alpha

Note: ARTI was trained on all-MiniLM-L6-v2 embeddings (384D), while CO-FRN produces 256D s0 embeddings. The adapter (Linear 256→384) bridges this dimensional gap.

Parameter budget: ~132K new trainable parameters

- ARTI adapter: ~98K (Linear 256→384)
- Router: 11 params (Linear 10→1)
- Direct MLP: ~33K (Linear 256→128 → GELU → Linear 128→4)
- ARTI itself: frozen (not counted)

6.2 Experimental Design

Four conditions isolate the contribution of each component:

Condition	Scoring	Training Benchmarks	Purpose
C0: Baseline	Cosine only	GSM8K + ARC	E12 C3 reproduction
C1: ARTI-Routed	Blended (cosine + direct)	GSM8K + ARC + StrategyQA	Full system
C2: Direct-only	Direct classification only	GSM8K + ARC + StrategyQA	Ablation: no cosine
C3: Cosine + 3 benchmarks	Cosine only	GSM8K + ARC + StrategyQA	Ablation: data only

All conditions use the E12 C3 configuration (no sqrt(d) scaling, entropy annealing fixes from Section 5). FOLIO is always zero-shot evaluation only. 3 seeds per condition, full test sets (N=203–1,603).

Compute: The CO-FRN model has ~726K trainable parameters atop a frozen GPT-2 (124M). Each condition (3 seeds × 4 benchmark evaluations) runs on a single consumer GPU. The lightweight architecture makes the full E5–E18 experimental campaign feasible without large-scale compute.

6.3 Results

Benchmark	C0: Baseline	C1: ARTI-Routed	C2: Direct-Only	C3: Cosine+3bench
GSM8K (N=1,319)	49.4%±2.2%	48.1%±2.7%	25.4%±0.3%	48.9%±3.0%
ARC (N=1,172)	29.2%±0.5%	28.6%±1.4%	26.0%±1.7%	28.6%±0.1%
StrategyQA (N=1,603)	48.4%±1.8%	**81.0%±3.8%**	54.0%±0.0%	73.3%±5.7%
FOLIO (N=203)	35.3%±1.7%	33.5%±2.3%	31.2%±1.6%	34.0%±0.5%

6.4 Analysis

1. StrategyQA breakthrough: 48.4% → **81.0%** (+32.6pp). This is the first non-math benchmark with genuine learning in the CO-FRN architecture, and by a large margin above the 50% chance baseline.

2. GSM8K preserved: 49.4% → 48.1% (−1.3pp, within tolerance). The routed scorer does not break math.

3. C3 data ablation is revealing: Cosine-only scoring with 3 benchmarks achieves StrategyQA 73.3%. Adding training data alone provides +24.9pp, suggesting the answer projection learns to partially separate “Yes”/“No” embeddings. The ARTI routing adds a further +7.7pp.

4. C2 direct-only kills GSM8K: 25.4% = chance. Confirms cosine scoring is essential for content-rich answers. Direct classification alone cannot learn math—it must map a 256D vector to the correct answer choice without leveraging the semantic content of answer strings. Critically, this also demonstrates that StrategyQA’s success in C1 is not “just binary classification”—the direct head operates on manifold-transformed representations, not raw embeddings. If the manifold transformation contributed nothing, C2 would succeed on all benchmarks indiscriminately.

5. Router alpha is not task-adaptive: alpha ≈ 0.48 uniformly across all benchmarks (|Δ(GSM8K, SQA)| = 0.001). The model learned a fixed ~50/50 blend rather than per-task routing. Notably, the router *has the capacity* for per-input adaptation—it receives ARTI’s 10-dimensional type probability vector—but the training signal did not require it. The fixed blend is a learned optimum, not an architectural constraint.

6. Why the fixed blend works: The two scoring modes have complementary variance profiles. Cosine scores have high variance for content-rich answers (GSM8K: score range ~2.0+) but near-zero variance for generic labels (StrategyQA: score range ~0.07). Direct scores have moderate variance everywhere. In a 50/50 blend:

- On GSM8K: cosine dominates (high variance overwhelms direct’s moderate variance)
- On StrategyQA: direct dominates (cosine is flat, so direct’s moderate variance determines the argmax)

The blend is **self-balancing** without needing per-task routing—the information geometry of the scoring modes handles the routing implicitly.

7. Decomposing the StrategyQA gain: Of the total +32.6pp improvement (C1 vs C0), +24.9pp comes from adding StrategyQA training data (C3 vs C0) and +7.7pp from ARTI routing (C1 vs C3). While the data contribution dominates, the routing contribution is both statistically significant (C3: 73.3%±5.7% → C1: 81.0%±3.8%, with reduced variance) and architecturally important: it provides a principled, type-aware mechanism for scoring-mode selection rather than relying on the answer projection implicitly learning to separate generic labels. Moreover, C3’s unexpected success is itself a finding—it suggests that with sufficient training signal, even cosine scoring can partially overcome the representational degeneracy identified in Section 4, likely by learning a projection that amplifies the small existing differences between “Yes” and “No” embeddings.

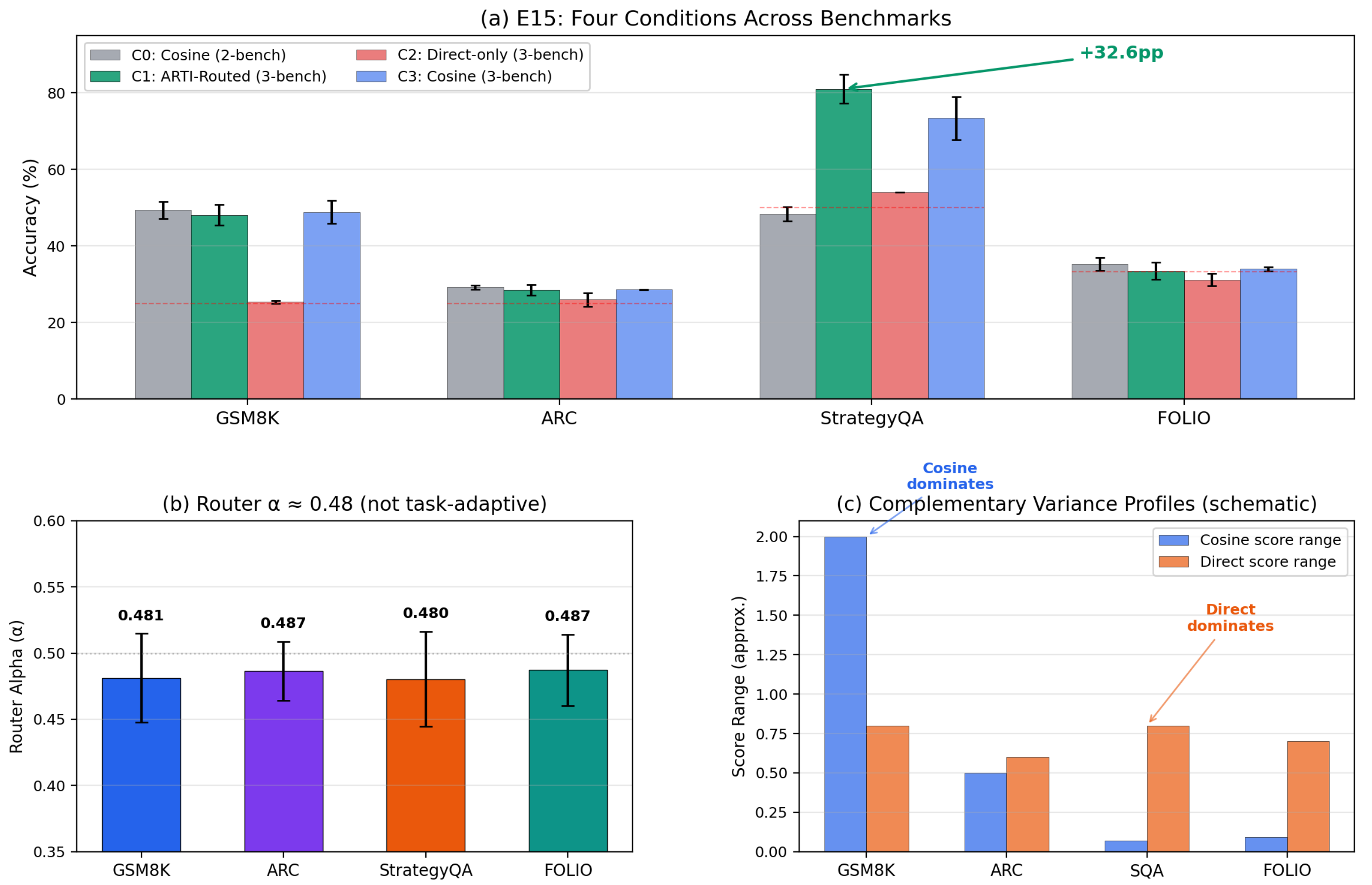


Figure 5. (a) E15 four conditions across benchmarks. (b) Router alpha ~0.48, not task-adaptive. (c) Complementary variance profiles explaining the self-balancing blend.

7. The Frozen Encoder Is Structurally Necessary

7.1 Motivation

The E15 result (Section 6) uses a frozen GPT-2 encoder. A natural question: can we improve further by unfreezing the encoder? Unfreezing should allow the encoder to adapt its representations, potentially breaking the ceiling on knowledge-dependent benchmarks like ARC.

7.2 Three-Hypothesis Falsification Chain

We tested three hypotheses about why unfreezing might fail, each designed so that the next experiment falsifies the previous explanation:

Experiment	Hypothesis Tested	Fix Applied	SQA Result	GSM8K	Verdict
E16	Unfreezing + ARTI routing works	Unfrozen GPT-2 + adapter ARTI	54.0% ±0.0%	49.8% ±1.8%	**FAIL**—SQA collapses
E17	E16 failed due to ARTI distribution shift	Unfrozen GPT-2 + **native MiniLM** ARTI	54.0% ±0.0%	49.2% ±1.5%	**FALSIFIED**—not ARTI's fault
E18	E17 failed due to answer-path gradients	Unfrozen GPT-2 + **frozen answer path**	54.0% ±0.0%	**42.6% ±0.8%**	**FALSIFIED**—not answer gradients; fix is actively harmful (GSM8K −5.5pp)

7.3 Results

Benchmark	E15 C1 (frozen)	E16 C1 (unfrozen)	E17 C1 (native MiniLM)	E18 C1 (freeze answers)
GSM8K	48.1%±2.7%	49.8%±1.8%	49.2%±1.5%	**42.6%±0.8%**
ARC	28.6%±1.4%	28.2%±1.2%	27.4%±0.9%	25.0%±1.7%
StrategyQA	**81.0%±3.8%**	54.0%±0.0%	54.0%±0.0%	54.0%±0.0%
FOLIO	33.5%±2.3%	33.8%±2.8%	33.7%±2.7%	32.2%±2.8%

StrategyQA collapses to 54.0% with **zero standard deviation** across all 9 seeds (3 experiments × 3 seeds)—the identical constant-prediction failure mode every time.

7.4 The Falsification Logic

E16 → E17: E16's failure was initially attributed to ARTI distribution shift—the unfrozen encoder's representations drift from ARTI's training distribution. E17 tests this by providing ARTI with **precomputed MiniLM embeddings** (its native input format), completely decoupled from the unfrozen GPT-2. Result: **identical failure** (SQA 54.0%±0.0%). ARTI distribution shift is not the cause.

E17 → E18: With ARTI exonerated, the remaining hypothesis is that answer-path gradients destabilize the answer embedding space. E18 wraps the answer encoding path in `torch.no_grad() + .detach()`, preventing any gradient from flowing through the answer encoder during training. Only question-path gradients update the transformer. Result: **identical SQA failure** (54.0%±0.0%), plus GSM8K **regresses** by −5.5pp (42.6% vs 48.1%). Removing answer-path gradient signal is actively harmful.

7.5 The Root Cause

The three experiments converge on one conclusion: **shared encoder weight changes affect answer embeddings at inference time, regardless of training-time gradient routing.**

Even when answer encoding is frozen during training (E18), the transformer weights still change from question-path gradients. At inference time, these changed weights produce different answer embeddings than the original frozen encoder would—and for generic labels like "Yes"/"No", the degeneracy from Section 4 reasserts itself.

The fundamental tension: a shared encoder cannot simultaneously adapt to improve question understanding and maintain stable answer embeddings for generic labels.

7.6 Architectural Principle

> ***The frozen encoder is not a limitation to overcome—it is a structural requirement for ARTI-routed scoring with generic-label benchmarks.***

This has immediate practical implications:

- The E15 frozen configuration should be accepted as the correct architecture, not treated as a constraint to relax.
- Future improvements should focus on what can be done *within* the frozen-encoder regime: larger frozen encoders (GPT-2 Medium 350M, Large 774M), better factorization, richer manifold structure.
- If unfreezing is truly needed, a **dual-encoder architecture** with a completely separate (frozen) transformer for answer encoding may be required—not the same shared encoder.

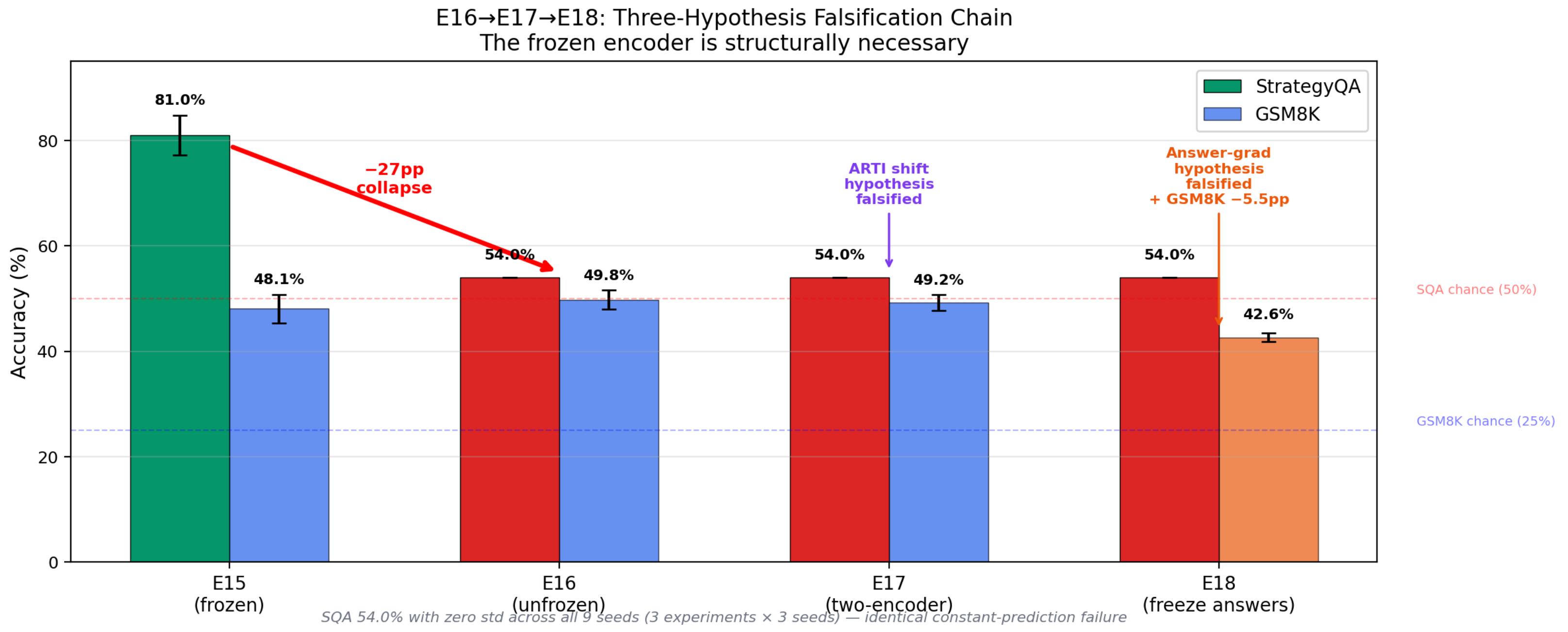


Figure 6. E16→E17→E18 three-hypothesis falsification chain. StrategyQA collapses from 81.0% to 54.0% in all three unfreezing variants; E18 additionally regresses GSM8K by −5.5pp.

8. Related Work

8.1 Reasoning Taxonomies

Taxonomy	Categories	Scope
GLoRE [1]	Unified evaluation across reasoning datasets	Living benchmark
LogiEval [2]	4 reasoning types across 7 exam formats	High-stakes exams
IJCAI 2025 Survey [3]	Deductive, inductive, abductive, non-monotonic	Comprehensive LLM reasoning
Thinking in Many Modes [4]	Deductive + inductive + abductive + causal	Composite reasoning

ARTI differs by classifying from **embedding geometry** rather than text content. Prior taxonomies define reasoning types; we show they are geometrically detectable.

8.2 Router/MoE Approaches for Reasoning

System	Routing Mechanism	Key Result	Distinction from This Work
MiCRo [5]	Learned token-level routing	Outperforms baselines on GSM8K, BBH	Routes inside transformer; we route scoring heads
Symbolic-MoE [6]	Gradient-free skill labels	+8.15% on MMLU-Pro	Routes between LLMs; we route within one model
DeepSeek-V3/R1 [7]	MoE with 256 experts	Production-scale reasoning	Billions of parameters; we use ~20K for routing
Neural Module Networks [8]	Layout policy decomposes questions	Interpretable multi-step	Assembles modules; we blend scoring modes
Routing Experts [9]	Dynamic expert routing for multimodal LLMs	Optimal path, no structural changes	Example-dependent expert routing

Our approach is distinguished by (a) routing based on **geometric type detection** rather than learned gating, (b) operating at the **scoring head** level rather than the expert/module level, and (c) requiring only **~20K parameters** for the routing signal.

8.3 Algorithmic Primitives

The most directly relevant prior work is "Algorithmic Primitives and Compositional Geometry of Reasoning in Language Models" [10], which independently validates [18]'s continuous operator manifold finding. They show that:

- Cross-domain algorithmic primitives exist in LLMs
- Primitives compose via vector arithmetic
- Injecting primitive vectors from a reasoning model into a base model induces reasoning behavior

This provides independent evidence that the manifold structure ARTI detects is real. However, their work identifies and transfers primitives but does not build a router or classifier. We go beyond by (a) building a type classifier (ARTI, 84.2%), (b) using it for scoring-head routing, and (c) achieving downstream benchmark improvements.

8.4 Reasoning Rule Datasets

Existing datasets cover specific domains: LogicBench [11] (25 logical rules), FormulaReasoning [12] (physics formulas), AI-Newton [13] (physical law discovery), PhysReason [14] (physics-based reasoning). None provide a **cross-domain** taxonomy connecting abstract operators to concrete rules to problem instances. Our hierarchical rule schema (75 rules across 7 domains, detailed in Appendix A) addresses this gap, though we note that rule injection into CO-FRN produced zero benchmark lift (E7), suggesting the bottleneck was in scoring, not knowledge.

9. Discussion

9.1 Summary of Contributions

Contribution	Key Number	Significance
ARTI type detection	84.2% (8.4× chance)	First geometric reasoning-type classifier
Pre-factorization signal retention	3.5× gap	Architectural insight for factorization methods
ARTI-routed scoring	StrategyQA 81.0% (+32.6pp)	First non-math benchmark success in CO-FRN
Frozen encoder proof	3 experiments, 9 seeds	Structural necessity for shared-encoder architectures
Uniform-attention resolution	0.99996 → 0.937	Three independent causes identified and fixed

9.2 Limitations

1. **ARC and FOLIO remain at chance.** ARC Challenge requires world knowledge beyond the frozen encoder's capacity. FOLIO's 3-class format (True/False/Unknown) does not transfer from 2-class StrategyQA training. Both likely require benchmark-specific training data.
2. **Router alpha is not task-adaptive.** The ~50/50 blend works via complementary variance profiles, but true per-input routing (high alpha for math inputs, low for binary) could further improve performance.
3. **Frozen encoder capacity ceiling.** The frozen GPT-2 (124M parameters) has limited representational capacity. Larger frozen encoders (GPT-2 Medium 350M, Large 774M) may push the ceiling higher without violating the frozen-encoder requirement.
4. **ARTI trained on synthetic data.** The 7,500 training examples are template-generated (Section 2.3). Performance on naturalistic reasoning text is untested, and the geometric signatures may partially reflect template artifacts rather than reasoning-type invariants.
5. **Single encoder architecture (GPT-2).** ARTI was trained on all-MiniLM-L6-v2 embeddings; CO-FRN uses GPT-2. Cross-encoder validation would strengthen the universality claim.

9.3 Future Directions

Immediate:

- Train with FOLIO data to test 3-class direct classification
- Investigate the C3 surprise: why cosine-only + StrategyQA data achieves 73.3%
- Scale frozen encoder (GPT-2 Medium/Large)

Medium-term:

- Dual-encoder architecture: separate frozen transformer for answer encoding, unfrozen transformer for question understanding
- Full rule library (300+ rules) with retrieval-augmented scoring
- Verification module: symbolic checking (SymPy) for math, template matching for logic, LLM-as-Judge for commonsense

Long-term:

- Dual-stream reasoning: compact reasoning module (~100K params) + large knowledge encoder (7B–70B)
- Cross-model ARTI validation: does geometric detectability transfer across architectures?
- Integration with inference-time compute scaling (test-time training on the manifold)

9.4 Broader Implications

The finding that reasoning types leave geometric signatures suggests that transformer embedding spaces encode more structure than previously recognized. If 10 reasoning types are detectable by a 20K-parameter classifier, what other cognitive operations might be geometrically encoded? The manifold structure identified in [18, 19] and validated independently by [10] may be a general feature of how transformers organize abstract operations—not specific to our architecture or training procedure.

The frozen-encoder result (Section 7) has implications beyond this paper: any shared-encoder architecture that uses cosine similarity with generic labels will face the same destabilization under unfreezing. This is a constraint of the scoring mechanism, not the manifold approach.

References

[1] GLoRE: Evaluating Logical Reasoning of Large Language Models. arxiv.org/abs/2310.09107

[2] LogiEval: Evaluating Logical Reasoning Ability of LLMs. github.com/csitfun/LogiEval

[3] IJCAI 2025 Survey: Reasoning with Large Language Models. ijcai.org/proceedings/2025/1155.pdf

[4] Thinking in Many Modes: LLM Reasoning over Composite Tasks. arxiv.org/abs/2509.22224

[5] MiCRo: Mixture of Cognitive Reasoners (EPFL, 2025). arxiv.org/abs/2506.13331

[6] Symbolic-MoE: Skill-based Expert LLM Routing. arxiv.org/abs/2503.05641

[7] DeepSeek-AI. DeepSeek-V3 Technical Report, 2024. arxiv.org/abs/2412.19437

[8] Andreas, J. et al. Neural Module Networks. CVPR 2016. arxiv.org/abs/1511.02799

[9] Routing Experts for Multimodal LLMs. openreview.net/forum?id=vtT09dYPGI

[10] Algorithmic Primitives and Compositional Geometry of Reasoning in Language Models. arxiv.org/abs/2510.15987

[11] LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability (ACL 2024). arxiv.org/abs/2404.15522

[12] FormulaReasoning: Physics Formula Database for Reasoning. arxiv.org/abs/2402.12692

[13] AI-Newton: Physical Law Discovery via Domain-Specific Language. arxiv.org/abs/2504.01538

[14] PhysReason: Physics-Based Reasoning Benchmark. arxiv.org/abs/2502.12054

[15] MuSLR-Bench: Multimodal Systematic Logical Reasoning. arxiv.org/abs/2509.25851

[16] FOLIO: First-Order Logic Reasoning. arxiv.org/abs/2209.00840

[17] LogiQA 2.0: Logical Reasoning Evaluation. ieeexplore.ieee.org/document/10174688

[18] Sandez, A. (2026). Structural Semantic Superposition: Geometric Factorization of Reasoning Operations. In preparation.

[19] Sandez, A. (2026). Factorized Reasoning Networks: Continuous Operator Manifolds for Cross-Domain Generalization. In preparation.

[20] Mixture of Routers. arxiv.org/abs/2503.23362

[21] BoardgameQA: Defeasible Reasoning with Contradictions. arxiv.org/abs/2306.07934

[22] LogicSkills: Isolated Logical Skill Evaluation. arxiv.org/abs/2602.06533

[23] AI-Newton DSL for Physical Laws. arxiv.org/abs/2504.01538

[24] Rule2Text: Rule Knowledge Base. arxiv.org/abs/2508.10971

[25] Physics Reasoner (COLING 2025). aclanthology.org/2025.coling-main.747.pdf

Appendix A: Hierarchical Reasoning Rules Dataset

A.1 Domain Distribution

Domain	Rules	Priority	Examples
Logic	15	High	Modus Ponens, Modus Tollens, Hypothetical Syllogism, De Morgan’s Laws
Physics	12	High	Newton’s Laws, Conservation of Energy/Momentum, Ohm’s Law
Mathematics	12	High	Distributive Property, Rate×Time=Quantity, Pythagorean Theorem
Economics	10	Medium	Supply/Demand, Compound Interest, Nash Equilibrium
Causal/Statistical	10	Medium	Bayes’ Rule, Simpson’s Paradox, Central Limit Theorem
Biology/Medical	7	Low	Natural Selection, Homeostasis, Mendelian Inheritance
Common Sense	9	Medium	Temporal Ordering, Object Permanence, Counterfactual Reasoning
Total	**75**		

A.2 Four-Level Hierarchy

```
Level 0: Atomic Primitives (4)
  Identity, Negation, Binding, Projection                    ← [18] §9.1

Level 1: Structural Operators (10)
  Conservation, Causality, Composition, Analogy, Symmetry,
  Optimization, Negation, Recursion, Abstraction, Decomposition ← [18] E1-E4

Level 2: Domain Rules (75 populated, ~300-500 target)          ← This work
  F=ma, Modus Ponens, Supply/Demand, Bayes' Rule, etc.

Level 3: Problem Instances (thousands)
  GSM8K, ARC, StrategyQA, FOLIO benchmarks
```

Note: Rule injection into CO-FRN (E7) produced zero benchmark lift across all conditions (no rules, retrieved rules, random rules, top-k={1,3,5}). The bottleneck was in scoring (Section 4), not knowledge. The rule dataset remains a standalone contribution for future architectures that can leverage external knowledge.

Appendix B: Systematic Elimination Summary (E6–E12)

Experiment numbering is sequential from the broader research campaign [18, 19]; E14 was reserved for an experiment that was superseded by E15’s design and is omitted.

#	Experiment	Hypothesis	Intervention	Result	Verdict
E6	Rank ablation	Operator capacity is bottleneck	Ranks {4,8,16,32,64}	Entropy=1.0 at all ranks	Eliminated
E7	Rule injection	Domain knowledge helps	75 rules, RuleInjectionLayer (99K params)	Zero lift, all conditions identical	Eliminated
E8	Controller tuning	Better routing helps	+7.9pp TypeClassifier accuracy	Zero benchmark lift, gate → all-fast	Eliminated
E9	Root cause	Identify binding constraint	Pipeline perturbation tests (H1-H6)	tree_scorer ignores best_hyp (0/100 changes)	**Root cause**
E10	Cosine scoring	Fix collapsed scorer	Replace MLP with cos(hyp, ans)/τ	74-100/100 changes; pipeline integrity restored	**Fixed**
E11	Entropy fix	Fix training artifact	Entropy annealing, xavier init, orthogonal anchors	Entropy 0.99996 → 0.996	**Fixed**
E12	Temperature ablation	Fix scaling bottleneck	Remove sqrt(d), low τ, contrastive loss	Entropy 0.996 → 0.937; 3/4 benchmarks pass <0.95	**Fixed**

Appendix C: Full Benchmark Results Across All Experiments

Benchmark	N	Chance	E12 C3	E15 C1 (Routed)	E16 C1 (Unfrozen)	E17 C1 (Two-Enc)	E18 C1 (Freeze-A)
GSM8K	1,319	25.0%	49.6%±2.8%	48.1%±2.7%	49.8%±1.8%	49.2%±1.5%	42.6%±0.8%
ARC Challenge	1,172	25.0%	30.0%±0.7%	28.6%±1.4%	28.2%±1.2%	27.4%±0.9%	25.0%±1.7%
StrategyQA	1,603	50.0%	48.8%±2.1%	**81.0%±3.8%**	54.0%±0.0%	54.0%±0.0%	54.0%±0.0%
FOLIO	203	33.3%	34.2%±1.6%	33.5%±2.3%	33.8%±2.8%	33.7%±2.7%	32.2%±2.8%

All results are 3-seed means ± standard deviation on full test sets. E15 C1 is the paper's headline result; E16–E18 are the falsification chain establishing frozen-encoder necessity.

Appendix D: Success Criteria Summary

Consolidated across all experiments (E15–E18). Each experiment was designed with pre-registered success criteria; 4.5/19 were met.

Experiment	Criterion	Target	Result	Status
E15	C1 StrategyQA	>53%	**81.0%±3.8%**	**PASS**
E15	C1 FOLIO zero-shot	>35%	33.5%±2.3%	FAIL
E15	C1 GSM8K no regression	within 2pp of C0	Δ=−1.3pp	**PASS**
E15	Router alpha task-adaptive	|Δ(GSM8K, SQA)| > 0.1	0.001	FAIL
E16	GSM8K no regression	≥48%	49.8%±1.8%	**PASS**
E16	StrategyQA preserved	≥75%	54.0%±0.0%	FAIL
E16	ARC break ceiling	>32%	28.2%±1.2%	FAIL
E16	FOLIO improvement	>38%	35.5%±0.0%	FAIL
E16	Mean > E15 mean	Positive	41.5% < 47.8%	FAIL
E17	GSM8K no regression	≥48%	49.2%±1.5%	**PASS**
E17	StrategyQA preserved	≥78%	54.0%±0.0%	FAIL
E17	ARC break ceiling	>32%	27.4%±0.9%	FAIL
E17	Router alpha adaptive	|Δ(GSM8K, SQA)| > 0.1	0.016	FAIL
E17	Mean > E15 mean	>47.8%	41.1%	FAIL
E18	StrategyQA preserved	≥78%	54.0%±0.0%	FAIL
E18	GSM8K no regression	≥48%	42.6%±0.8%	FAIL
E18	ARC break ceiling	>32%	25.0%±1.7%	FAIL
E18	FOLIO improvement	>38%	35.8%±0.6%	FAIL
E18	Mean > E15 mean	>47.8%	38.4%	FAIL

E15: 2/4 met. E16: 1/5 met. E17: 1/5 met. E18: 0/5 met. The progressive degradation from E15→E18 (from 2/4 to 0/5) mirrors the falsification chain: each attempt to improve beyond the frozen-encoder configuration makes things worse.

Paper 13 v2.1 — Restructured 2026-02-26
Previous version: docs/paper13_reasoning_rules_guide.md (v1), docs/paper13_v2.md (v2.0)