AN1 White Paper: Meaning-First Compute and Field-Based Inference

Anima Core Inc.

Version 1.0

1. Introduction

Transformer-based inference dominates enterprise AI costs. While training deep models is expensive, it is predictable. Inference, however, grows with user volume, request frequency, and operational scale. Modern LLMs require large GPU fleets to serve production traffic, creating cost barriers and increasing carbon impact.

AN1 proposes a meaning-first compute framework. Rather than executing the entire transformer stack, AN1 extracts semantic fields from early layers and produces predictions directly in a compact representation. This reduces compute while preserving teacher-level structure and alignment.

2. The Problem: Transformer Inference at Scale

Transformer inference requires sequential computation through many layers, including attention blocks and feedforward networks. Each layer contributes significant FLOPs, making costs scale poorly.

A scalable alternative must:

- preserve model behavior
- require no retraining
- reduce FLOPs by an order of magnitude
- integrate with existing infrastructure

AN1 achieves these goals by bypassing deep-layer computation.

3. AN1 Architecture Overview

AN1 consists of three components:

1. Field Extraction

Activations from the first transformer layer (or anchor layers) are collected. These activations encode strong semantic signal before deeper combinatorial mixing.

2. Field Compression

Using PCA/SVD-style projection, activations are compressed by 40×–224×. Compression reveals that early semantic fields lie in stable low-dimensional manifolds.

3. Symbolic Inference Head

Predictions are generated from compressed fields using a lightweight head. No transformer blocks are executed.

This removes 90–99 percent of compute for common tasks.

4. Field Extraction Pipeline

The AN1 pipeline includes:

- Token embeddings
- First-layer attention activations
- Multi-anchor stabilization
- Field normalization
- Low-dimensional projection

Multi-anchor extraction increases robustness and reduces drift across prompts.

5. Compression Results: 40×–224× Reduction

Field compression experiments show:

- 40×–224× dimensionality reduction
- 0.88–0.92 logit rank correlation with teacher model
- 61–68 percent top-1 agreement across tasks
- High stability across seeds

These results demonstrate that semantic content is concentrated in narrow activation subspaces.

6. Benchmark Performance

Benchmarks evaluated:

- RTE
- HellaSwag
- SST-2
- ARC-AE
- Enterprise routing tasks

Results:

- 10–100× FLOP reduction
- Monotonic logit alignment
- Predictive stability at high compression

Meaning-first compute generalizes across teacher scales.

7. Technical Implementation Details

AN1 integrates as a flexible architecture:

- Containerized microservice
- CPU or GPU execution
- Deterministic compute graph
- Pluggable inference head
- No fine-tuning required

AN1 only needs early-layer activations, not full model execution.

8. Enterprise Deployment Model

Deployment options:

- API proxy
- On-prem cluster
- Secure VPC setup
- Hybrid teacher–monitor mode

AN1 reduces GPU load from day one without requiring architecture modification.

9. Pilot Program

The pilot includes:

- Workload and cost analysis
- Field extraction configuration
- AN1 deployment and monitoring
- Accuracy and agreement metrics
- Weekly engineering updates

Typical duration: 7–30 days.

10. Conclusion

Meaning-first compute reframes how inference is performed. By relying on compressed semantic fields rather than full transformer computation, AN1 dramatically reduces cost while preserving alignment.

This architecture enables sovereign, efficient AI infrastructure at global scale.

Contact: partner@animacore.ai