Panadero AI Engine Development Plan

Q4 2025 Internal Roadmap

Executive Summary

Panadero Services is embarking on the development of a proprietary AI engine to power our next-generation applications. This comprehensive 12-week development plan outlines the complete roadmap from foundational tensor operations to production-ready AI inference capabilities.

Phase 1: Foundation - Weeks 1-2

- Custom JavaScript tensor operations library
- Memory-efficient data structures
- Matrix and vector operations
- Comprehensive testing framework
- Build and development pipeline

Phase 2: Neural Core - Weeks 3-4

- Linear transformation layers
- Layer normalization (RMSNorm)
- Multi-head attention mechanism
- Group Query Attention (GQA)
- RoPE positional encoding

Phase 3: Architecture - Weeks 5-6

- Stacked transformer blocks
- Language modeling head
- Memory optimization techniques
- Model compression strategies

Phase 4: Training - Weeks 7-8

- Cross-entropy loss implementation
- AdamW optimizer with weight decay
- Learning rate scheduling
- Data loading and preprocessing
- Model checkpointing and recovery

Phase 5: Inference - Weeks 9-10

- Autoregressive text generation
- Beam search and sampling strategies
- RESTful API for text generation
- Real-time streaming output

Phase 6: Production - Weeks 11-12