KB-205: Resolving 'Inference Latency Timeout' (AI Model Failure)

Applicable Product: Predictive AI Workflow Trigger (PAWT)
Target Audience: Workflow Creators / L2 Technical Specialists
Last Updated: 2025-11-05

Problem Description

You have received a workflow execution failure with the error message: [ERROR_CODE: AI-LT-504] Inference Latency Timeout: Model NexaCore-TSM v1.2 exceeded P99 threshold of 500ms.

This means the AI Model responsible for generating a prediction took too long to return a result, causing the downstream workflow step to fail.

Possible Causes and Resolution

Cause 1: High Input Payload Volume (Most Common)

If your workflow is passing extremely large data packets (e.g., payloads) to the AI service, it can strain the inference engine, leading to latency.

- **Resolution:** Implement a pre-processing step using the **Data Transformer** block. Summarize or extract only the essential features required for the AI model *before* passing the data to the prediction block.

Cause 2: Regional Network Congestion

If the client's API Gateway is geographically distant from the AI inference cluster (e.g., accessing the Dubai AI cluster from an African regional node), network latency can contribute to the timeout.

- **Resolution (L2 Only):** Check the current **System Health Dashboard** for high-latency alerts between the client's region and the main AI cluster. If congestion is confirmed, escalate as a P1 incident to SRE for potential traffic routing adjustments.

Cause 3: Model Degradation / Out-of-Date Cache

The model cache might be stale, forcing a slow disk read, or the model itself may be degrading under load.

- **Resolution (L3 R&D Only):** L3 Engineering must check the **AI Model Validation Report** for the latest performance metrics. If performance is below the target, L3 must execute a manual model cache refresh or a full retraining cycle.