# White Paper: Generative AI for Infrastructure Load Prediction & Utilization Forecasting

## 1. Introduction

Traditional time-series forecasting models rely on deep learning and machine learning techniques such as LSTMs, ARIMA, and Transformer-based models. However, these approaches primarily focus on extrapolating past trends to predict future utilization. This paper explores a Generative AI-based approach to predicting infrastructure resource load and utilization by simulating realistic demand scenarios rather than deriving utilization trends from past data.

## 2. Problem Statement

Existing methods for infrastructure resource utilization forecasting extrapolate trends from historical data. This leads to challenges such as:

• Inability to model unseen scenarios (e.g., sudden demand surges, market fluctuations).

• Limited flexibility in business-driven forecasting where external drivers impact infrastructure demand.

• Overfitting to past utilization patterns rather than generating realistic future workloads.

## 3. Generative AI for Load Prediction

Instead of predicting utilization as a function of past values, a Generative AI model can simulate infrastructure load based on external business drivers and operational conditions. This involves:

• \*\*Generating synthetic workloads\*\* instead of forecasting point estimates.

• \*\*Creating future resource demand scenarios\*\* under hypothetical conditions.

• \*\*Adapting to dynamic infrastructure changes\*\*, including cloud migrations and elasticity.

## 4. Key Generative AI Models for Infrastructure Load Generation

### Diffusion Models for Workload Generation

Diffusion models can learn the complex data distribution of infrastructure workloads and generate realistic future resource usage under various conditions.

### Variational Autoencoders (VAEs) for Infrastructure Simulation

VAEs can capture latent patterns in business drivers and infrastructure load, allowing realistic scenario generation.

### GAN-based Infrastructure Demand Simulation

Generative Adversarial Networks (GANs) can simulate infrastructure usage variations, ensuring generated workloads match real-world behaviors.

### Transformer-based Generative Models

Large-scale transformer models can synthesize future resource usage based on historical patterns and hypothetical business conditions.