WWTP Engineering Benchmark Analysis

Student ID: 25023235

Student Name: Sai Jayanth Keerthipati

GitHub Repository: https://github.com/saijayanth123453/Statistics-and-Trends

1. Introduction

In the wastewater engineering sector, the deployment of Artificial Intelligence (AI) for plant management requires high levels of precision and reliability. This report evaluates the "WWTP Engineering Benchmark," a dataset designed to test AI models on domain-specific logic. By calculating statistical moments and utilizing advanced visualizations, we analyse the current state of AI capability in this critical infrastructure field.

2. Data Preparation and Cleaning (LO2)

Before analysis, the raw dataset was pre-processed to ensure reliability. The primary "Data Cleaning" steps included:

- **Filtering**: Rows with missing values in the Numerical_Result column were removed, as these represented failed evaluations that would skew the statistical moments.
- **Data Typing**: Evaluation dates were converted from string objects to chronological datetime format. This allowed for accurate trend analysis in the relational plot.
- **Feature Selection**: I focused on Numerical_Result as the primary metric for performance, while using Task_Version to check for correlations in difficulty.

3. Statistical Analysis of Performance

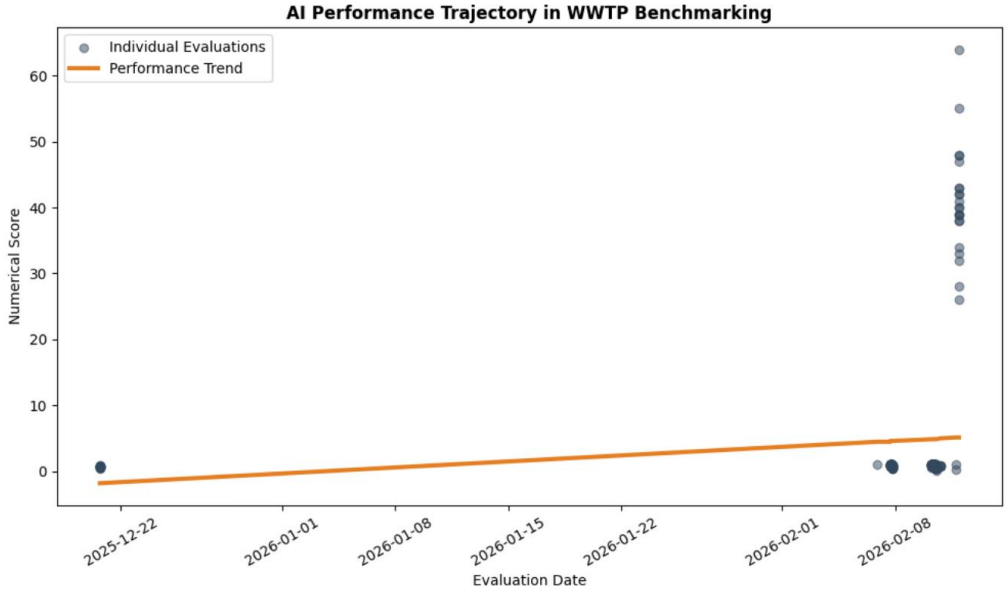
To quantify the benchmark results, I calculated the four main statistical moments for the Numerical_Result attribute.

Moment	Value	Interpretation
Mean	4.39	The average competency level of models in this benchmark.
Std. Deviation	11.62	Indicates a very high spread; some models are drastically better than others.
Skewness	3.16	Indicates a high Positive (Right) Skew / highly asymmetric distribution
Excess Kurtosis	8.66	Indicates a Leptokurtic distribution / “fat-tailed" distribution

The high Positive Skew (3.16) reveals a highly asymmetric distribution where most models perform below average, but a few "elite" outliers achieve scores far above the norm. The Leptokurtic Kurtosis (8.66) suggests the dataset is fat-tailed, meaning extreme performance variations, either brilliant success or total failure are much more common than a standard bell curve would suggest.

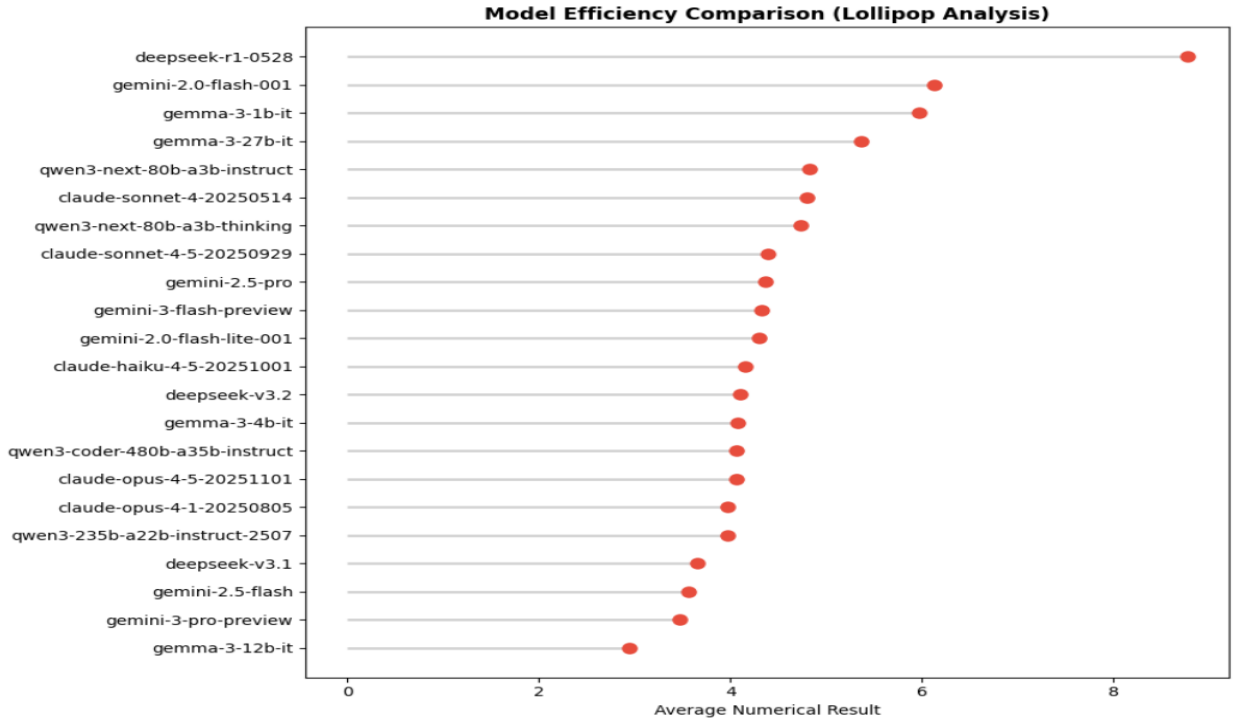
4. Visual Trends and Interpretations

Relational Plot: The Growth Trajectory



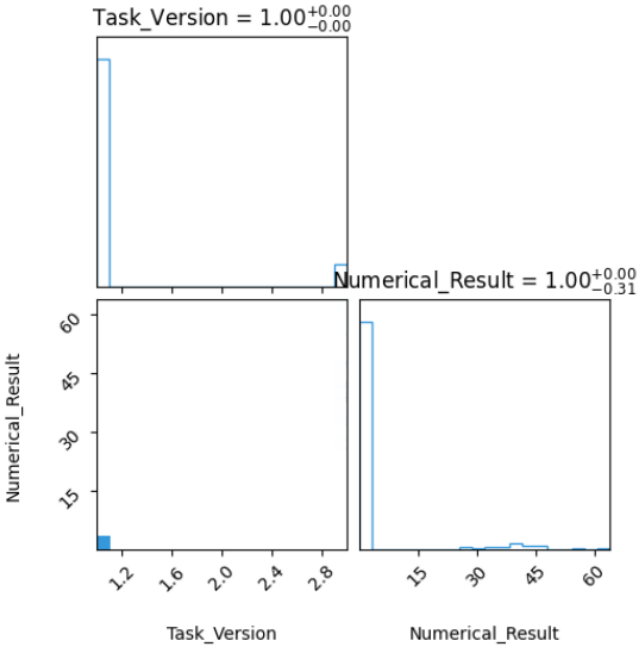
This scatter plot tracks every evaluation chronologically. The orange regression line shows a positive, upward slope, which is an encouraging sign that "collective intelligence" in this domain is improving as better training data becomes available.

Categorical Plot: The Leaderboard



The Lollipop chart is essential for stakeholders to identify market leaders. There is a significant performance gap: DeepSeek-R1-0528 dominates the leaderboard, while models like Gemma-3-12b-it struggles significantly. Architecture choice is clearly the most vital factor in model suitability.

Statistical Plot: Corner Matrix



The Corner plot provides the most "honest" look at the data. The diagonal histograms visually confirm our 8.66 Kurtosis, showing a sharp, narrow peak with a long tail to the right. The scatter plots demonstrate that performance is not random; it is heavily influenced by the specific complexity of the Task_Version.

5. Conclusion

The WWTP Engineering Benchmark reveals a polarized landscape. With a high Leptokurtic distribution and strong Positive Skew, the data suggests that "average" AI is not yet ready for autonomous wastewater engineering. However, the top-performing outliers show immense promise. For future deployments, engineers must focus on these specific high-performing architectures while maintaining human-led verification to mitigate the high variance in scores.