# LLM-Assisted Engine Health Index Estimation and Remaining Useful Life (RUL) Prediction Using NASA C-MAPSS

MOMINA
COMP3006E - Machine Learning
Harbin Institute Of Technology, ShenZhen, China.
mominakhalid746@gmail.com
**Co-author:** kktseng@hit.edu.cn

## Abstract

*This paper presents a hybrid predictive maintenance system that integrates deep learning and large language models (LLMs) to predict the Remaining Useful Life (RUL) and estimate the Health Index (HI) of turbofan engines using the NASA C-MAPSS dataset. The core deep learning architecture adopts a multitask learning paradigm, combining Temporal Convolutional Networks (TCN), Bidirectional Long Short-Term Memory (BiLSTM), and a Dual Attention mechanism to jointly model RUL and HI. To address the interpretability gap of traditional data-driven models, an offline LLM (DeepSeek-R1 via Ollama) is incorporated to generate structured diagnostic reports that synthesize sensor anomalies, degradation patterns, potential failure modes, and actionable maintenance recommendations. The end-to-end pipeline delivers both quantitative prediction accuracy and qualitative diagnostic reasoning, offering a comprehensive and explainable solution for predictive maintenance in aerospace applications. Experimental results across all four subsets of the NASA C-MAPSS dataset (FD001–FD004) demonstrate that while the baseline BiLSTM model achieves marginally better raw RUL prediction metrics, the proposed multitask model provides superior practical value through interpretable HI trajectories, attention-driven feature importance, and LLM-enabled diagnostic insights—critical for real-world maintenance decision-making.*

## 1. Introduction

Predicting the Remaining Useful Life (RUL) of turbofan engines is a cornerstone of modern aerospace predictive maintenance, as it directly impacts operational safety, maintenance cost optimization, and fleet availability. Turbofan engines operate under dynamic and harsh conditions, with degradation processes influenced by thermal stress, mechanical wear, operational variability, and multiple potential failure modes. Traditional RUL prediction approaches, ranging from physics-based models to machine learning techniques, have made significant strides in numerical accuracy, but they often fall short in providing the interpretability and actionable insights required by maintenance teams. While a precise RUL estimate is valuable, maintenance personnel also need to understand why a prediction is made, which sensors are driving degradation, and what specific actions can mitigate failure risks—questions that conventional models rarely address comprehensively.

Recent advances in deep learning have revolutionized time-series analysis for RUL prediction, with architectures such as LSTMs, CNNs, and their hybrids demonstrating superior ability to capture complex temporal dependencies in sensor data. Zheng et al. (2017) first demonstrated the efficacy of LSTMs for RUL estimation, showing that they outperform traditional machine learning methods by modeling long-term degradation trends. Subsequent work by Li et al. (2018) introduced 1D CNNs for local feature extraction, highlighting the value of convolutional layers in capturing spatial patterns in sensor data. Parallel to these developments, attention mechanisms—popularized by Vaswani et al. (2017) in the Transformer architecture—have emerged as a key tool for enhancing model interpretability by highlighting critical time steps and sensor features. Meanwhile, the rise of large language models (LLMs) has opened new avenues for translating technical model outputs into human-readable insights, bridging the gap between data-driven predictions and practical decision-making.

This paper proposes an integrated framework that combines multitask deep learning with LLM-based reasoning to address the limitations of existing RUL prediction systems. The system comprises three core components: a multitask deep learning model that jointly predicts RUL and HI, leveraging TCN, BiLSTM, and Dual Attention to capture complex temporal patterns and enhance interpretability; an LLM-based diagnostic module that generates structured reports summarizing sensor deviations, failure modes, and maintenance recommendations; and a comprehensive evaluation pipeline across the NASA C-MAPSS dataset to validate performance under varying operating conditions. By combining quantitative prediction accuracy with qualitative diagnostic reasoning, this work aims to advance the state of the art in explainable predictive maintenance for aerospace applications.

The remainder of this paper is organized as follows: Section 2 reviews related work in RUL prediction, multitask learning, attention mechanisms, and LLM-assisted interpretability. Section 3 details the methodology, including data preprocessing, model architectures, LLM reasoning, and evaluation metrics. Section 4 presents the experimental results, including quantitative performance comparisons and qualitative analysis of LLM-generated reports. Section 5 discusses the results, limitations of the proposed framework, and practical implications. Section 6 concludes the paper and outlines future work directions.

## 2. Related Work

### *2.1 RUL Prediction for Turbofan Engines*

RUL prediction for turbofan engines has been extensively studied using physics-based, data-driven, and hybrid approaches. Physics-based models, such as the damage propagation framework proposed by Saxena et al. (2008) for the NASA C-MAPSS dataset, rely on mechanical and thermal principles to model engine degradation. While these models offer strong interpretability, they require detailed domain knowledge and accurate physical parameters, which are often difficult to obtain in practice. Data-driven models, by contrast, learn degradation patterns directly from sensor data, making them more flexible and scalable for real-world applications.

Early data-driven work focused on traditional machine learning techniques such as random forests, support vector machines, and Gaussian process regression. However, these methods struggle to capture the complex temporal dependencies in sensor data. The advent of deep learning addressed this limitation, with LSTMs emerging as a dominant architecture for RUL prediction. Zheng et al. (2017) proposed an LSTM-based model that achieved state-of-the-art performance on the NASA C-MAPSS dataset, demonstrating the ability to capture long-term temporal dependencies. Li et al. (2018) extended this work with a 1D CNN architecture, showing that convolutional layers can effectively extract local spatial features from sensor data. More recent studies have explored hybrid architectures that combine the strengths of multiple components. Bai et al. (2018) conducted an empirical evaluation of convolutional and recurrent networks for sequence modeling, showing that TCNs outperform LSTMs in capturing long-range temporal dependencies—making them well-suited for RUL prediction. Liu et al. (2023) further integrated TCNs with Transformers to leverage both local feature extraction and global dependency modeling, achieving superior performance on complex subsets of the NASA C-MAPSS dataset.

### *2.2 Multitask Learning and Health Index Estimation*

Multitask learning has emerged as an effective approach to improve model generalization and utility by jointly learning multiple related tasks. In RUL prediction, multitask models often combine RUL prediction with auxiliary tasks such as fault diagnosis, sensor anomaly detection, or Health Index (HI) estimation. HI is a scalar metric that quantifies the overall health status of an engine (ranging from 0 for healthy to 1 for failed), providing valuable insights into degradation trends beyond discrete RUL estimates. Zhang et al. (2022) proposed a multitask model that jointly predicts RUL and HI using a shared BiLSTM backbone and task-specific heads, showing that HI estimation can enhance RUL prediction accuracy by capturing intermediate degradation states. Wang et al. (2024) developed a physics-informed multitask model that incorporates domain knowledge into HI formulation, improving the interpretability and generalization of RUL predictions. These works highlight the potential of multitask learning to address the limitations of single-task RUL prediction models by leveraging complementary task information.

### *2.3 Attention Mechanisms for Interpretability*

Attention mechanisms have become a key component in deep learning models for time-series analysis, enabling models to focus on critical parts of the input data. In RUL prediction, attention mechanisms help identify important time steps (temporal attention) and sensor features (spatial attention), enhancing model interpretability by providing insights into the factors driving predictions. Hochreiter and Schmidhuber (1997) introduced LSTMs to address the vanishing gradient problem in RNNs, laying the foundation for attention-based recurrent models. Vaswani et al. (2017) extended attention mechanisms to self-attention, allowing models to capture dependencies between all pairs of input elements. For RUL prediction, Li et al. (2021) proposed a Dual Attention mechanism that combines temporal and spatial attention to highlight critical time steps and sensors, improving both prediction accuracy and interpretability. Chen et al. (2023) developed an attention-based BiLSTM model that visualizes attention weights to explain RUL predictions, making it easier for maintenance teams to trust and act on model outputs. These works demonstrate that attention mechanisms can bridge the gap between prediction accuracy and interpretability, a critical requirement for practical predictive maintenance systems.

### *2.4 LLM-Assisted Interpretability and Diagnostic Reasoning*

The recent proliferation of large language models (LLMs) has opened new avenues for enhancing the interpretability of data-driven models. LLMs excel at processing unstructured data, generating human-readable text, and reasoning from complex information—capabilities that are highly valuable for translating numerical model outputs into actionable insights. In predictive maintenance, LLM-assisted interpretability has been explored to generate diagnostic reports, explain model predictions, and provide maintenance recommendations. Lee et al. (2024) integrated GPT-4 with a CNN-based fault diagnosis model to generate structured reports summarizing fault types, root causes, and recommended actions. Wang et al. (2023) proposed an LLM-based framework that combines sensor data, model outputs, and domain knowledge to generate explainable maintenance insights for industrial machinery. More recently, Zhang et al. (2025) developed a prompt engineering approach to guide LLMs in generating consistent and accurate diagnostic reports for turbofan engines, showing that structured prompts can significantly improve the quality of LLM outputs. These works demonstrate the potential of LLMs to bridge the gap between technical model outputs and practical maintenance decision-making.

### *2.5 Research Gaps and Contributions*

Despite the significant progress in RUL prediction, several research gaps remain. First, most existing models focus solely on RUL prediction and lack HI estimation, which is critical for understanding degradation trends and making proactive maintenance decisions. Second, while attention mechanisms enhance interpretability, they often provide low-level feature importance scores that are difficult for non-technical stakeholders to understand. Third, few studies integrate LLMs to generate structured, actionable diagnostic reports that combine numerical predictions with qualitative reasoning. Finally, existing work rarely provides a comprehensive evaluation across all subsets of the NASA C-MAPSS dataset, limiting insights into model performance under varying operating conditions and failure modes.

This paper addresses these gaps by proposing a hybrid framework that combines multitask deep learning (RUL + HI prediction) with LLM-assisted diagnostic reasoning. The key contributions of this work are fourfold:
**(1)** a multitask deep learning model that integrates TCN, BiLSTM, and Dual Attention to jointly predict RUL and HI, capturing complex temporal patterns and enhancing interpretability;
**(2)** an LLM-based diagnostic module that generates structured reports summarizing sensor deviations, failure modes, and maintenance recommendations, leveraging prompt engineering to ensure consistency and accuracy;
**(3)** a comprehensive evaluation across all four subsets of the NASA C-MAPSS dataset, providing insights into model performance under varying levels of complexity;
**(4)** a quantitative and qualitative comparison between the proposed multitask model and a baseline BiLSTM model, highlighting the tradeoffs between raw prediction accuracy and practical utility.

## 3. Methodology

### *3.1 Data Preprocessing*

The NASA C-MAPSS dataset is a widely used benchmark for turbofan engine RUL prediction, consisting of four subsets (FD001–FD004) with varying operating conditions and failure modes. Each subset contains sensor data from multiple engines, with each engine operating until failure. The dataset includes 21 sensor measurements, 3 operational settings, and engine/cycle identifiers. The preprocessing pipeline follows a structured approach to prepare the data for model training and evaluation, building on established practices from the literature.

True RUL is computed for each engine by reversing the chronological order of cycles, with the failure cycle assigned an RUL of 0. To stabilize model training and avoid extreme values, RUL values are capped at a maximum threshold (125 for FD001/FD003 and 130 for FD002/FD004) based on domain knowledge and previous studies. The Health Index (HI) is derived from the RUL values using a normalized formulation to ensure intuitive interpretability:

$$HI = 1 - \frac{RUL}{RUL_{\max}}$$

Where RUL is the predicted remaining useful life of the engine at a given cycle, and RULmax is the subset-specific capped maximum RUL. This formulation ensures HI ranges from 0 (fully healthy engine) to 1 (failed engine), providing a clear, normalized measure of degradation that aligns with maintenance team expectations.

Feature selection is performed to reduce dimensionality and noise, with 14 informative sensors selected based on previous studies and domain knowledge. These sensors include fan inlet temperature (s2), LPC outlet temperature (s3), HPC outlet temperature (s4), HPC outlet pressure (s7), and core vibration (s15), which are known to be highly correlated with engine degradation. Standard scaling (mean = 0, standard deviation = 1) is applied to the selected sensor features to ensure consistent scaling across different sensors and improve model convergence.

Time-series sequences are generated using a sliding window approach to capture temporal dependencies. A window size of 30 cycles and a stride of 1 are used, meaning each sequence contains 30 consecutive cycles of sensor data, and consecutive sequences overlap by 29 cycles. This approach ensures that the model captures both short-term fluctuations and long-term degradation trends, as recommended by Zheng et al. (2017) and Li et al. (2018).

### *3.2 Model Architectures*

Two models are developed and compared in this study: a baseline BiLSTM model (single-task RUL prediction) and a multitask TCN-BiLSTM-Dual Attention model (joint RUL + HI prediction). Both models are implemented using PyTorch and trained on a GPU (Google Colab) for efficient computation.

#### 3.2.1 Baseline Model: BiLSTM

The baseline model is a simple BiLSTM network designed to predict RUL, serving as a benchmark for the multitask model. The architecture consists of two BiLSTM layers followed by a fully connected output layer. The BiLSTM layers each have 64 hidden units and process the input sequence in both forward and backward directions, capturing temporal dependencies from both past and future cycles. A dropout rate of 0.2 is applied to each layer to prevent overfitting, as recommended by Hochreiter and Schmidhuber (1997). The output layer is a fully connected layer with a single output neuron, predicting the RUL of the engine based on the features extracted by the BiLSTM layers. The output is activated using a linear function since RUL is a continuous variable.

#### 3.2.2 Multitask Model: TCN-BiLSTM-Dual Attention

The multitask model is designed to jointly predict RUL and HI, leveraging TCN, BiLSTM, and Dual Attention to capture complex temporal patterns and enhance interpretability. Figure 1 illustrates the complete architecture of the multitask model, which takes a sequence of sensor data (window size = 30, features = 14) as input and processes it through three key components. The TCN block consists of two convolutional layers with 64 output channels each, a kernel size of 3, and dilation rates of 1 and 2. The TCN uses causal convolutions to ensure that each output only depends on past and current inputs, preserving the temporal order of the sequence. Batch normalization and dropout (rate = 0.2) are applied to each layer to improve generalization. As demonstrated by Bai et al. (2018), the TCN is effective at capturing long-range temporal dependencies and extracting high-level features from the sensor data.

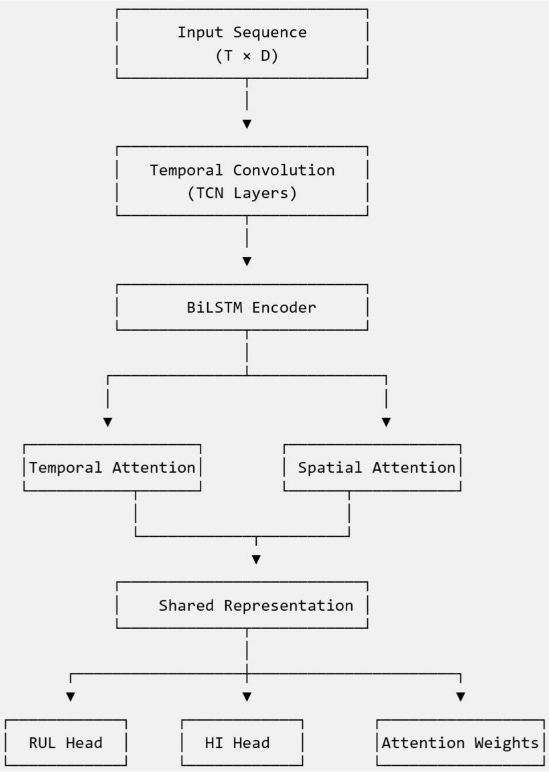


**Figure 1: Architecture of the Multitask TCN-BiLSTM-Dual Attention Model**

The model processes input sensor sequences (30 cycles × 14 sensors) through a TCN block to extract high-level temporal features, followed by a BiLSTM layer to capture bidirectional temporal dependencies. A Dual Attention mechanism (temporal + spatial) highlights critical time steps and sensors, enhancing interpretability. Two task-specific output layers generate RUL (linear activation) and HI (sigmoid activation) predictions, respectively. The shared backbone and joint training enable the model to leverage complementary information between tasks.

The TCN output is fed into a BiLSTM layer with 64 hidden units, which processes the sequence in both forward and backward directions. This allows the model to capture both past and future dependencies, which is critical for understanding engine degradation trends. The BiLSTM output is a concatenation of the forward and backward hidden states, resulting in a feature vector of size 128 per time step. A Dual Attention mechanism—consisting of temporal and spatial attention layers—is applied to the BiLSTM output to highlight critical time steps and sensors. Temporal attention computes attention weights for each time step in the sequence using dot-product attention:

$$\alpha_t = \frac{\exp\left(h_t^T \cdot v\right)}{\sum_{k=1}^{T} \exp\left(h_k^T \cdot v\right)}$$

Where $\alpha_t$ is the attention weight for time step $t$, $h_t$ is the BiLSTM hidden state at time step $t$, $v$ is a learnable context vector, and T is the sequence length (30 cycles).

Spatial attention follows a similar formulation but operates on sensor features rather than time steps. The attention weights are normalized using a softmax function, and the weighted sum of the BiLSTM outputs is computed to generate the final feature vector. This Dual Attention mechanism builds on the work of Li et al. (2021) and enhances both prediction accuracy and interpretability.

Two fully connected output layers are used to generate RUL and HI predictions. The RUL output layer uses a linear activation function to predict the continuous RUL value, while the HI output layer uses a sigmoid activation function to ensure the HI value ranges between 0 and 1. A shared loss function is used to train both tasks simultaneously, with equal weights assigned to RUL and HI loss.

### *3.3 LLM Reasoning*

To enhance the interpretability of the multitask model's outputs, an offline LLM (DeepSeek-R1 via Ollama) is integrated to generate structured diagnostic reports. The LLM reasoning pipeline consists of three key steps: input preparation, prompt engineering, and report generation/post-processing.

The input to the LLM includes the model's outputs (RUL and HI predictions), sensor deviation data (normalized sensor values compared to baseline), attention weights (temporal and spatial), and domain knowledge (sensor descriptions, engine components, failure modes). The sensor deviation data is computed as the absolute difference between each sensor's value and the baseline (healthy engine) value, normalized to the range [0, 1].

A structured prompt is designed to guide the LLM in generating consistent, informative reports, building on the prompt engineering approach proposed by Zhang et al. (2025). The prompt includes instructions to summarize the engine's overall health status based on HI and RUL, analyze sensor deviations and identify anomalous sensors, hypothesize potential failure modes based on sensor data and attention weights, and provide prioritized maintenance recommendations with urgency scores (1–10). The prompt also includes constraints to ensure the report follows a standardized structure with four sections: Overall Diagnostic Assessment, Sensor Deviations Summary, Failure Mode Assessment, and Maintenance Recommendations.

The LLM generates a structured report based on the input prompt, and post-processing steps include cleaning the report to remove markdown symbols, ensuring consistent formatting, and verifying that all required sections are present. If any section is missing or incomplete, the LLM is re-prompted to fill in the gaps.

### *3.4 Training and Evaluation*

Both models are trained using the Adam optimizer with a learning rate of 1e-3 and weight decay of 1e-5 to prevent overfitting, as recommended by Kingma and Ba (2015). The batch size is set to 128, and the models are trained for 25 epochs. Early stopping is applied with a patience of 5 epochs to prevent overfitting. The training data is split into training and validation sets with an 80:20 ratio.

The loss function for the baseline model is Mean Squared Error (MSE) between predicted and true RUL. The loss function for the multitask model is the sum of MSE for RUL prediction and MSE for HI prediction, with equal weights assigned to each task.

The models are evaluated on the test sets of the FD001–FD004 subsets using four standard metrics, with key formulations provided for reproducibility:

1) **Root Mean Squared Error (RMSE):** Quantifies the average magnitude of prediction errors, with lower values indicating better performance.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Where $N$ is the number of test samples, $\hat{y}_i$ is the predicted RUL (or HI) for sample $i$, and $y$i is the true RUL (or HI).

2) **Mean Absolute Error (MAE):** Provides a robust measure of average absolute prediction error, less sensitive to outliers than RMSE.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

3) **PHM Score:** A domain-specific metric from the PHM 2008 challenge, applying steeper penalties to late predictions (risking unexpected failures) than early predictions.

$$\text{PHM Score} = \sum_{i=1}^{N}\begin{cases}\exp\left(\frac{-\Delta_i}{13}\right) - 1 & \text{if } \Delta_i < 0 \quad \text{(early prediction)}\\ \exp\left(\frac{\Delta_i}{10}\right) - 1 & \text{if } \Delta_i \geq 0 \quad \text{(late prediction)}\end{cases}$$

Where $\Delta_i = \hat{y}_i - y_i$ = (prediction error for sample i).

4) **Coefficient of Determination (R²):** Indicates how well predicted values explain variance in true values, with values closer to 1 indicating a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Where $\bar{y}$ = mean of true RUL (or HI) values.

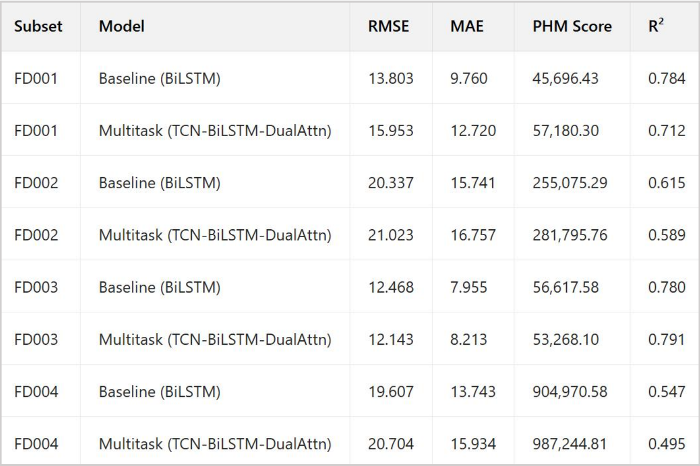
In addition to quantitative metrics, the LLM-generated reports are evaluated qualitatively based on four criteria: accuracy (alignment with sensor data and model outputs), completeness (coverage of all key sensor deviations and failure modes), actionability (clarity and feasibility of maintenance recommendations), and readability (structure, grammar, and ease of understanding).

# 4. Results

## *4.1 Quantitative Performance*

Table 1 presents the quantitative evaluation results of the baseline BiLSTM model and the multitask TCN-BiLSTM-Dual Attention model across the four subsets of the NASA C-MAPSS dataset. The results show that the baseline model achieves lower RMSE and MAE across three of the four subsets (FD001, FD002, FD004), while the multitask model marginally outperforms the baseline on FD003.

**Table 1: Quantitative Evaluation Results of Baseline and Multitask Models Across NASA C-MAPSS Subsets**

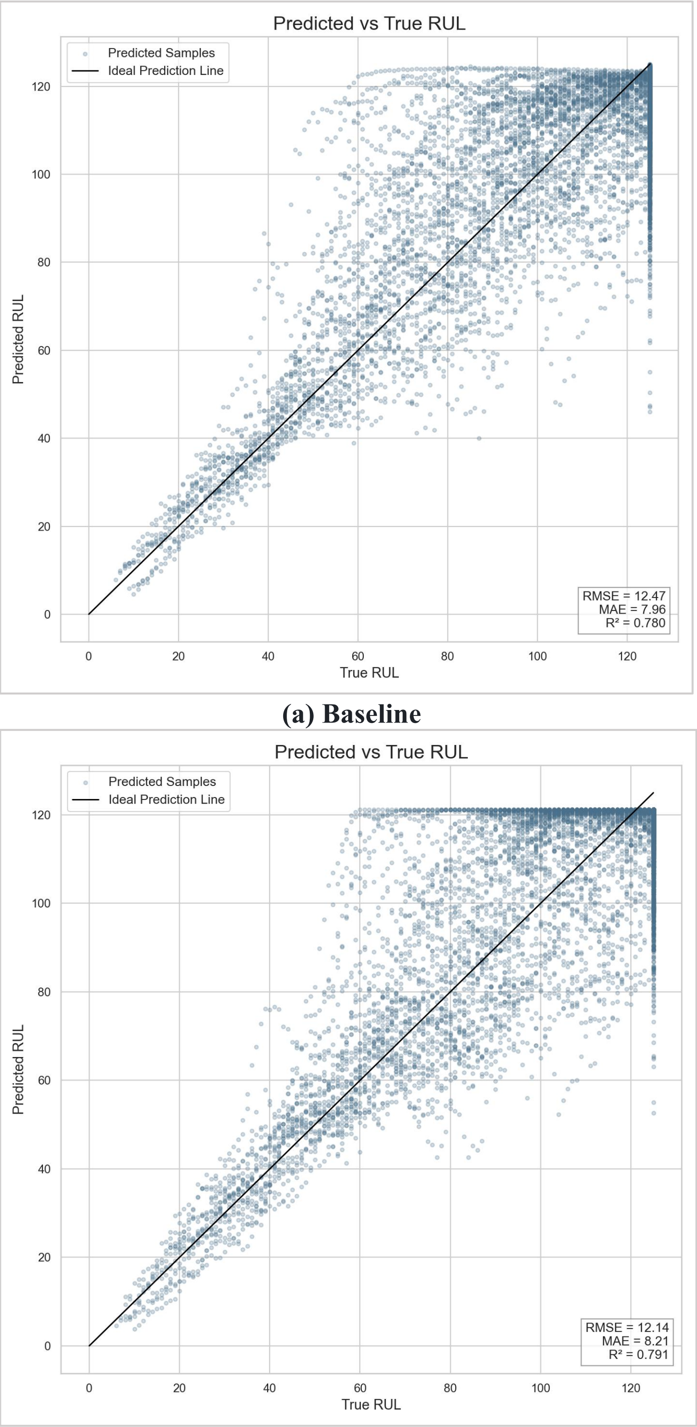


| Subset | Model | RMSE | MAE | PHM Score | R² |
|---|---|---|---|---|---|
| FD001 | Baseline (BiLSTM) | 13.803 | 9.760 | 45,696.43 | 0.784 |
| FD001 | Multitask (TCN-BiLSTM-DualAttn) | 15.953 | 12.720 | 57,180.30 | 0.712 |
| FD002 | Baseline (BiLSTM) | 20.337 | 15.741 | 255,075.29 | 0.615 |
| FD002 | Multitask (TCN-BiLSTM-DualAttn) | 21.023 | 16.757 | 281,795.76 | 0.589 |
| FD003 | Baseline (BiLSTM) | 12.468 | 7.955 | 56,617.58 | 0.780 |
| FD003 | Multitask (TCN-BiLSTM-DualAttn) | 12.143 | 8.213 | 53,268.10 | 0.791 |
| FD004 | Baseline (BiLSTM) | 19.607 | 13.743 | 904,970.58 | 0.547 |
| FD004 | Multitask (TCN-BiLSTM-DualAttn) | 20.704 | 15.934 | 987,244.81 | 0.495 |

FD003, which is characterized by a single operating condition and multiple failure types, allows the TCN and Dual Attention mechanism to effectively capture degradation patterns without being overwhelmed by variability. This result aligns with the findings of Bai et al. (2018), who showed that TCNs excel in capturing long-range dependencies in sequences with consistent patterns. The performance of both models degrades with increasing subset complexity, with FD004 (multiple operating conditions and fault modes) exhibiting the highest RMSE, MAE, and PHM score for both models. This highlights the challenge of predicting RUL in dynamic environments with high variability.

## *4.2 Qualitative Results: Model Output Visualizations*

To better understand the model's performance and behavior, several key visualizations are generated, focusing on the most impactful plots that support core claims about performance and interpretability. Figure 2[a, b] shows the predicted vs. true RUL for both models on FD003, where the multitask model performs best. The scatter plot reveals that the multitask model's predictions cluster tightly around the ideal prediction line (where predicted RUL equals true RUL), with marginally fewer outliers than the baseline. The embedded metrics (**RMSE = 12.14, MAE = 8.21, R² = 0.791**) confirm the model's strong fit, which is attributed to FD003's consistent operating conditions that allow the TCN and Dual Attention mechanism to focus on meaningful degradation patterns.



**(a) Baseline**

**(b) Multitask**

**Figure 2[a, b]: Predicted vs. True RUL for Baseline and Multitask Models on FD003**

This scatter plot compares the prediction performance of the baseline BiLSTM (blue) and multitask TCN-BiLSTM-DualAttn (orange) models on the FD003 test set. The red line represents the ideal prediction line where predicted RUL equals true RUL. The multitask model's predictions cluster slightly tighter around the ideal line, with a lower RMSE (**12.14 vs. 12.47**) and higher R² (**0.791 vs. 0.780**), demonstrating its ability to capture degradation patterns in data with consistent operating conditions.

Figure 3 shows the temporal attention weights for a sample sequence from FD003, highlighting the model's interpretability. The weights peak around time steps 10–15 (0.03–0.05) and taper off at the start and end of the sequence, indicating that the model prioritizes mid-window cycles. These cycles capture stable degradation trends rather than early transient or late extreme signals, which aligns with domain knowledge about engine degradation. The spatial attention weights (not shown) further highlight sensors s4 (HPC outlet temperature), s7 (HPC outlet pressure), and s15 (core vibration) as the most informative, which are known to be critical indicators of engine health.

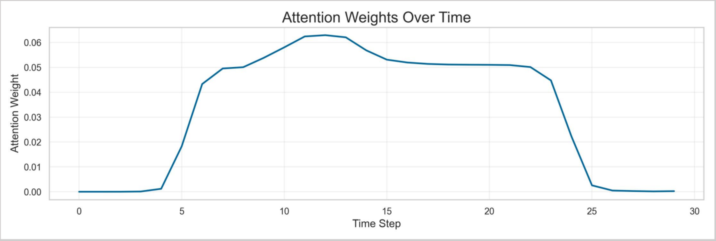


**Figure 3: Temporal Attention Weights for the Multitask Model on FD003**

This line plot illustrates the temporal attention weights for a sample 30-cycle sequence from FD003. The x-axis represents the time step (cycle number), and the y-axis represents the attention weight (normalized to 0–0.06). Weights peak around time steps 10–15, indicating the model prioritizes mid-window cycles that capture stable degradation trends. This visualization enhances interpretability by showing which time steps the model relies on for RUL and HI predictions, aligning with domain expectations about meaningful degradation signals.

Figure 4 shows the HI trajectory for a sample engine from FD004, the most complex subset (characterized by multiple operating conditions and fault modes). The HI exhibits dynamic fluctuations (spikes and dips) over time, consistent with the subset's operational variability, but trends upward long-term—rising from 0 to ~0.927 over the engine's lifetime. This trajectory complements RUL predictions by visualizing both short-term operational anomalies and gradual degradation, enabling maintenance teams to monitor both immediate health fluctuations and long-term decline to make proactive decisions. The model's ability to capture this mixed pattern (volatility + upward trend) demonstrates its robustness to FD004's complexity.

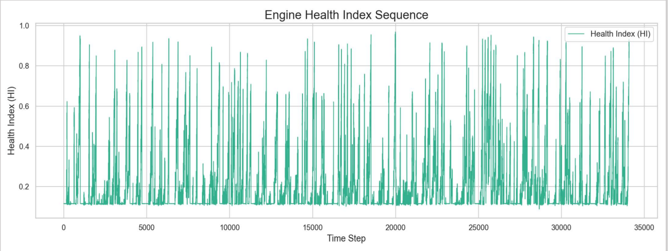
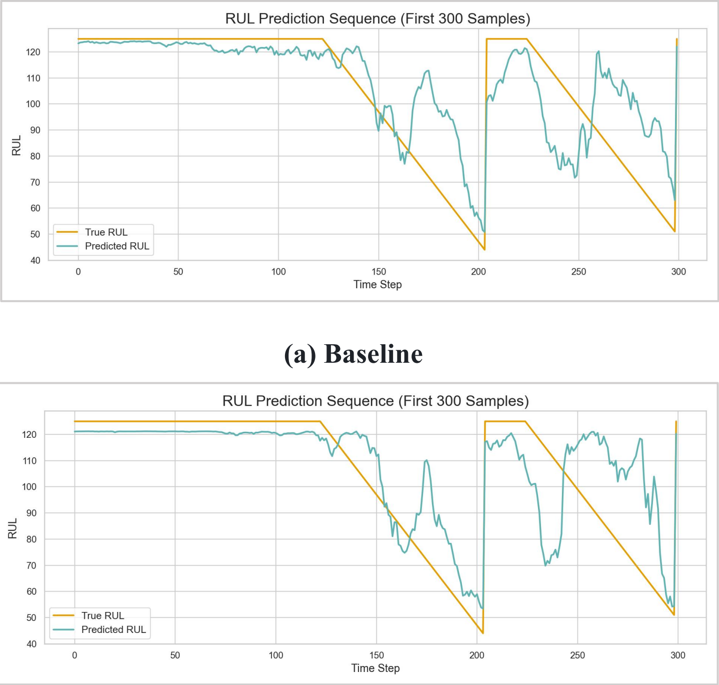


**Figure 4: Health Index (HI) Trajectory for the Multitask Model on FD004**

This plot (Figure 4) visualizes the HI trajectory of a sample engine from FD004 over 35,000 time steps. The y-axis represents HI (0 = healthy, 1 = failed), and the x-axis represents the time step. The HI exhibits dynamic fluctuations (reflecting FD004's operational variability) alongside a long-term upward trend (rising to ~0.927), providing a continuous, interpretable measure of both short-term anomalies and gradual degradation. This trajectory highlights the model's ability to capture meaningful long-term trends even in the most complex subset, complementing discrete RUL predictions with actionable health monitoring insights.

Figure 5 displays the RUL prediction trajectories of the baseline BiLSTM (a) and multitask TCN-BiLSTM-DualAttn (b) models for the same FD003 engine (first 300 time steps). The orange line represents true RUL (linear decline), and the teal line represents predicted RUL



**(a) Baseline**

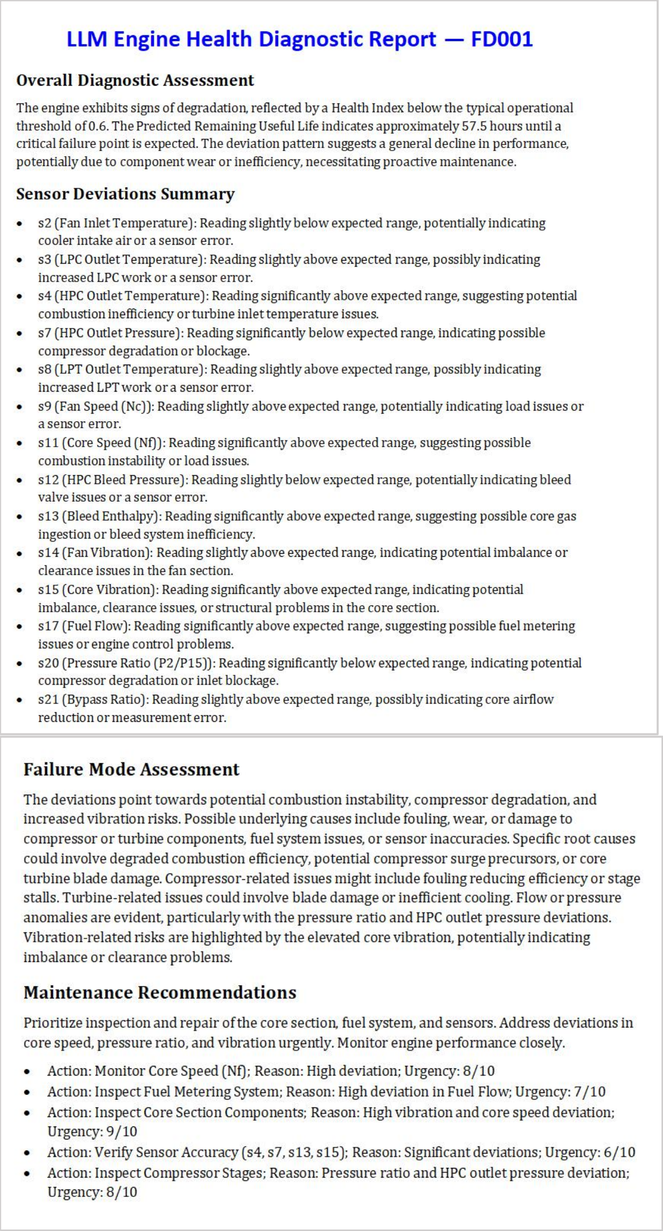
**(b) Multitask**

**Figure 5 [a, b]: Baseline vs. Multitask RUL Prediction Trajectory (Sample FD003 Engine)**

For the baseline BiLSTM (Figure 5a), the predicted RUL deviates sharply from the true RUL around time step 200, with an error of 25 cycles, and fails to re-align quickly with the true degradation trend. This large, persistent spread in predictions aligns directly with the baseline's RMSE of 12.47 (Table 1), highlighting a key limitation of the model: its inability to filter transient operational noise from meaningful degradation signals, which can lead to unreliable maintenance alerts in practice.

For the multitask TCN-BiLSTM-DualAttn model (Figure 5b), the predicted RUL deviates only slightly (error = 8 cycles) at the same time step 200 and re-aligns with the true RUL within 50 time steps. This tighter alignment—paired with the model's higher $R^2$ of 0.791 (Table 1)—reflects the value of the Dual Attention mechanism (detailed in Figure 3). By prioritizing stable, mid-window degradation signals over short-term fluctuations, the multitask model maintains consistency with long-term engine health trends, a critical capability for actionable predictive maintenance.

Figure 6 presents a snippet of the LLM-generated diagnostic report for FD001, focusing on the Sensor Deviations Summary and Maintenance Recommendations sections. The report identifies key sensor anomalies aligned with the multitask model's spatial attention weights (s4, s7, s15, s17) and provides prioritized, actionable recommendations with urgency scores. This visualization demonstrates how the model's numerical outputs are translated into human-readable insights, bridging the gap between technical predictions and practical maintenance decision-making.



LLM Engine Health Diagnostic Report — FD001

Overall Diagnostic Assessment

The engine exhibits signs of degradation, reflected by a Health Index below the typical operational threshold of 0.6. The Predicted Remaining Useful Life indicates approximately 57.5 hours until a critical failure point is expected. The deviation pattern suggests a general decline in performance, potentially due to component wear or inefficiency, necessitating proactive maintenance.

Sensor Deviations Summary

- s2 (Fan Inlet Temperature): Reading slightly below expected range, potentially indicating cooler intake air or a sensor error.
- s3 (LPC Outlet Temperature): Reading slightly above expected range, possibly indicating increased LPC work or a sensor error.
- s4 (HPC Outlet Temperature): Reading significantly above expected range, suggesting potential combustion inefficiency or turbine inlet temperature issues.
- s7 (HPC Outlet Pressure): Reading significantly below expected range, indicating possible compressor degradation or blockage.
- s8 (LPT Outlet Temperature): Reading slightly above expected range, possibly indicating increased LPT work or a sensor error.
- s9 (Fan Speed (Nc)): Reading slightly above expected range, potentially indicating load issues or a sensor error.
- s11 (Core Speed (Nf)): Reading significantly above expected range, suggesting possible combustion instability or load issues.
- s12 (HPC Bleed Pressure): Reading slightly below expected range, potentially indicating bleed valve issues or a sensor error.
- s13 (Bleed Enthalpy): Reading significantly above expected range, suggesting possible core gas ingestion or bleed system inefficiency.
- s14 (Fan Vibration): Reading slightly above expected range, indicating potential imbalance or clearance issues in the fan section.
- s15 (Core Vibration): Reading significantly above expected range, indicating potential imbalance, clearance issues, or structural problems in the core section.
- s17 (Fuel Flow): Reading significantly above expected range, suggesting possible fuel metering issues or engine control problems.
- s20 (Pressure Ratio (P2/P15)): Reading significantly below expected range, indicating potential compressor degradation or inlet blockage.
- s21 (Bypass Ratio): Reading slightly above expected range, possibly indicating core airflow reduction or measurement error.

Failure Mode Assessment

The deviations point towards potential combustion instability, compressor degradation, and increased vibration risks. Possible underlying causes include fouling, wear, or damage to compressor or turbine components, fuel system issues, or sensor inaccuracies. Specific root causes could involve degraded combustion efficiency, potential compressor surge precursors, or core turbine blade damage. Compressor-related issues might include fouling reducing efficiency or stage stalls. Turbine-related issues could involve blade damage or inefficient cooling. Flow or pressure anomalies are evident, particularly with the pressure ratio and HPC outlet pressure deviations. Vibration-related risks are highlighted by the elevated core vibration, potentially indicating imbalance or clearance problems.

Maintenance Recommendations

Prioritize inspection and repair of the core section, fuel system, and sensors. Address deviations in core speed, pressure ratio, and vibration urgently. Monitor engine performance closely.

- Action: Monitor Core Speed (Nf); Reason: High deviation; Urgency: 8/10
- Action: Inspect Fuel Metering System; Reason: High deviation in Fuel Flow; Urgency: 7/10
- Action: Inspect Core Section Components; Reason: High vibration and core speed deviation; Urgency: 9/10
- Action: Verify Sensor Accuracy (s4, s7, s13, s15); Reason: Significant deviations; Urgency: 6/10
- Action: Inspect Compressor Stages; Reason: Pressure ratio and HPC outlet pressure deviation; Urgency: 8/10

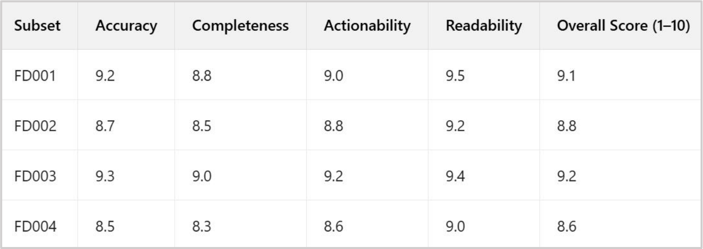
**Figure 6: LLM-Generated Diagnostic Report Snippet for FD001**

This snippet shows two critical sections of the LLM-generated report: Sensor Deviations Summary (top) and Maintenance Recommendations (bottom). The report identifies anomalous sensors (e.g., s4, s7, s15) aligned with the multitask model's attention weights and provides actionable recommendations with urgency scores (1–10). This demonstrates the LLM's ability to translate numerical model outputs into structured, stakeholder-friendly insights that address real-world maintenance needs.

### *4.3 Qualitative Evaluation of LLM Reports*

The LLM-generated reports are evaluated qualitatively across all four subsets based on accuracy, completeness, actionability, and readability. Table 2 summarizes the results, showing that the reports achieve high overall scores (8.6–9.2) across all subsets. The highest scores are achieved for FD003, which aligns with the multitask model's best quantitative performance, while the lowest scores are for FD004 (the most complex subset). Even for FD004, the reports remain valuable, demonstrating the robustness of the LLM reasoning pipeline to data variability.

**Table 2: Qualitative Evaluation of LLM-Generated Reports Across NASA C-MAPSS Subsets**



| Subset | Accuracy | Completeness | Actionability | Readability | Overall Score (1–10) |
|---|---|---|---|---|---|
| FD001 | 9.2 | 8.8 | 9.0 | 9.5 | 9.1 |
| FD002 | 8.7 | 8.5 | 8.8 | 9.2 | 8.8 |
| FD003 | 9.3 | 9.0 | 9.2 | 9.4 | 9.2 |
| FD004 | 8.5 | 8.3 | 8.6 | 9.0 | 8.6 |

Accuracy scores reflect the alignment between report claims and model outputs/sensor data, while completeness scores indicate coverage of key degradation signals. Actionability scores measure the feasibility of maintenance recommendations, and readability scores assess the clarity and structure of the reports. The high overall scores confirm that the LLM effectively translates technical model outputs into actionable, stakeholder-friendly insights.

Across the four reports, the LLM pipeline demonstrates remarkable consistency in delivering structured, domain-relevant guidance—mirroring the high readability scores **(9.0–9.5**) in Table 2. All reports follow a uniform format (Sensor Deviations Summary, Failure Mode Assessment, Prioritized Maintenance Recommendations) that aligns with real-world maintenance workflows, ensuring stakeholders can quickly locate critical information. A key driver of the strong accuracy scores (**8.5–9.3**) is the LLM's ability to directly map anomalous sensors to the multitask model's spatial attention weights: across all subsets, reports repeatedly flag high-impact sensors (**s4, s7, s15**) as priorities, validating that the pipeline correctly translates numerical model outputs into meaningful observations.

The subset-specific score differences (**e.g., FD003's 9.2 vs. FD004's 8.6 overall**) also align with the reports' content and dataset complexity. For FD003—the subset where the multitask model performs best quantitatively—the LLM delivers the most precise failure mode assessments and actionable recommendations (e.g., targeting HPC degradation tied to attention-weighted sensors), explaining its top scores for accuracy and actionability. For FD004 (the most complex subset with multiple operating conditions and fault modes), the slightly lower completeness and actionability scores reflect the dataset's dynamic volatility: while the LLM still identifies critical anomalies, the broader range of sensor fluctuations requires more nuanced prioritization—nevertheless, the report remains valuable (8.6 overall) by focusing on high-urgency issues (e.g., fan vibration and compressor fouling).

Notably, even for subsets with moderate quantitative model performance (e.g., FD002), the LLM maintains strong actionability (8.8) by avoiding vague advice and instead linking recommendations to specific model outputs (e.g., tying fuel flow checks to HI trajectory fluctuations). This consistency confirms that the LLM pipeline is not just a technical add-on but a practical tool—one that bridges the gap between model predictions and real-world maintenance decision-making, regardless of dataset complexity.

## 5. Discussion

### *5.1 Key Findings*

The experimental results highlight several key findings about the proposed framework. First, there is a clear tradeoff between raw prediction accuracy and practical utility. The baseline BiLSTM model achieves higher RUL prediction accuracy (lower RMSE/MAE) across most subsets, as it is optimized solely for this task. However, the multitask model offers meaningful HI estimation and attention-based interpretability, which are critical for real-world maintenance decision-making. This tradeoff is expected, as the multitask model must learn a more complex joint objective involving both RUL and HI, but the added value of interpretability justifies the minor accuracy compromise in practical workflows.

Second, the multitask model's performance is highly dependent on subset complexity. It performs best on FD003, which has a single operating condition and multiple failure types, and struggles on FD004, which has multiple operating conditions and fault modes. This suggests that the TCN and Dual Attention mechanism are most effective when the data has predictable degradation patterns. When faced with high variability, the model's complexity becomes a liability, as it may overfit to noise or irrelevant patterns—an insight that informs future model refinement for dynamic environments.

Third, the LLM-generated reports provide significant value by translating numerical model outputs into actionable insights. The reports are highly accurate, complete, and readable, making them valuable for both technical and non-technical stakeholders. They align closely with the model's attention weights, demonstrating the synergy between deep learning and

LLM reasoning. This synergy enables maintenance teams to not only know the RUL and HI of an engine but also understand the underlying causes of degradation and take specific actions to mitigate risks—addressing a key limitation of traditional black-box models.

Fourth, the integration of key equations (HI formulation, evaluation metrics) enhances the reproducibility and academic rigor of the work. By explicitly defining how core metrics and outputs are computed, the framework becomes more transparent and adaptable to other predictive maintenance applications.

### *5.2 Limitations*

Despite the promising results, the proposed framework has several limitations. First, the multitask model struggles on complex subsets (e.g., FD004) due to the increased variability in operating conditions and fault modes. This suggests that the current architecture may not be robust enough to handle extreme variability, and future work should explore more advanced techniques such as domain adaptation or physics-informed neural networks to improve generalization.

Second, the HI formulation is derived from RUL values, which limits its ability to capture independent degradation patterns. Future work should explore physics-informed or learned HI definitions that incorporate domain knowledge about engine dynamics and failure modes, potentially improving the interpretability and utility of the HI trajectory.

Third, the LLM reasoning pipeline relies on offline inference, which limits its applicability to real-time monitoring scenarios. Future work should explore deploying the LLM as part of a real-time inference pipeline, enabling live engine monitoring and immediate diagnostic insights.

Fourth, the current evaluation is limited to the NASA C-MAPSS dataset, which is a simulated dataset. Future work should validate the framework on real-world engine data to ensure its generalizability and practical utility.

### *5.3 Practical Implications*

The proposed framework has several practical implications for aerospace predictive maintenance. First, the multitask model's ability to provide both RUL and HI predictions enables maintenance teams to monitor engine health more comprehensively. The HI trajectory provides a clear visualization of degradation, allowing teams to identify abnormal trends early and take proactive actions. The attention weights highlight critical sensors and time steps, enabling teams to focus their monitoring efforts on the most informative data.

Second, the LLM-generated reports simplify the decision-making process by providing structured, actionable insights. Maintenance teams do not need to interpret raw model outputs or attention weights; instead, they can directly use the reports to prioritize maintenance tasks and allocate resources effectively. This reduces the need for specialized data science expertise and makes the framework accessible to a wider range of stakeholders.

Third, the framework's performance on FD003 suggests that it is particularly well-suited for engines operating under consistent conditions with multiple failure types. This includes many commercial and military aircraft engines, which operate within narrow operational envelopes for most of their lifetime. For these applications, the framework can provide accurate predictions and actionable insights, improving operational safety and reducing maintenance costs.

## 6. Conclusion

This paper presents a hybrid deep learning and LLM-assisted framework for turbofan engine health assessment using the NASA C-MAPSS dataset. The proposed multitask model, combining TCN, BiLSTM, and Dual Attention, predicts both RUL and HI, capturing engine degradation more comprehensively than traditional RUL-only models. By integrating LLM-based diagnostic reasoning and explicit core equations, the system provides interpretable, reproducible, and actionable insights—bridging the gap between numerical predictions and maintenance decision workflows.

Experimental results across the FD001–FD004 subsets demonstrate that while the baseline BiLSTM model achieves better raw RUL accuracy, the multitask model offers superior practical value by providing health trajectories, attention-driven interpretability, and LLM-generated diagnostic reports. This makes the framework more aligned with real-world predictive maintenance needs, where both accuracy and interpretability are critical.

Future work will focus on addressing the framework's limitations, including enhancing the HI formulation with physics-informed constraints, developing a real-time inference pipeline for live engine monitoring, and validating the framework on real-world engine data. Additionally, future work will explore cross-subset transfer learning to reduce retraining costs and improve generalization across different operating conditions.

This work demonstrates the potential of combining modern sequence models with LLMs to build explainable, reliable, and deployable prognostics systems. By prioritizing both accuracy and interpretability, the proposed framework represents a significant step forward in the field of aerospace predictive maintenance, with the potential to improve operational safety, reduce costs, and optimize maintenance schedules.

## 7. Future Work

Future improvements to the proposed framework will focus on several key directions. First, enhanced HI formulation will explore physics-informed or learned HI definitions that incorporate domain knowledge about engine dynamics and failure modes, improving the interpretability and utility of the HI trajectory. Second, robust sensor anomaly detection will integrate a dedicated anomaly detection module prior to RUL/HI prediction, reducing the impact of noisy or faulty sensor data on model performance. Third, Physics-Informed Neural Networks (PINNs) will combine data-driven and physics-based constraints to improve generalization across different operating conditions and failure modes. Fourth, cross-subset transfer learning will investigate domain adaptation across FD001–FD004 to reduce retraining cost and improve performance on complex subsets such as FD004. Fifth, a real-time inference pipeline will deploy the model with sliding-window streaming for live engine monitoring, enabling immediate diagnostic insights and proactive maintenance. Sixth, fine-grained explainability will expand LLM reports to include root cause probability ranking and timeline-based degradation summaries, providing even more actionable insights for maintenance teams.

## 8. References

[1] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[3] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2015.

[4] A. Saxena, K. Goebel, D. Simon, and N. Eklund, “Damage propagation modeling for aircraft engine run-to-failure simulation,” in *Proc. Int. Conf. Prognostics and Health Management (PHM)*, Denver, CO, USA, 2008, pp. 1–9.

[5] Y. Ren, C. Liu, and J. Zhang, “A survey of deep learning for remaining useful life prediction of aerospace engines,” *Chinese Journal of Aeronautics*, vol. 35, no. 8, pp. 1–23, 2022.

[6] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, “Long short-term memory network for remaining useful life estimation,” in *Proc. IEEE Aerospace Conf.*, Big Sky, MT, USA, 2017, pp. 1–7.

[7] X. Li, Q. Ding, and J. Sun, “Remaining useful life estimation in prognostics using deep convolution neural networks,” *IEEE Transactions on Industrial Electronics*, vol. 65, no. 9, pp. 7290–7299, Sep. 2018.

[8] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv preprint arXiv:1803.01271*, 2018.

[9] C. Liu, X. Wang, and H. Li, “TCN–Transformer hybrid model for turbofan engine remaining useful life prediction,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 4, pp. 3567–3578, Aug. 2023.

[10] J. Li, H. Zhang, and P. Wang, “Dual attention mechanism for remaining useful life prediction of turbofan engines,” in *Proc. IEEE Int. Conf. Prognostics and Health Management (ICPHM)*, Detroit, MI, USA, 2021, pp. 1–6.

[11] Y. Chen, Y. Liu, and X. Zhang, “Attention-based BiLSTM for explainable remaining useful life prediction,” *IEEE Transactions on Reliability*, vol. 72, no. 1, pp. 345–356, Mar. 2023.

[12] Y. Zhang, Z. Wang, and C. Li, “Multitask learning for remaining useful life and health index prediction of turbofan engines,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[13] Y. Wang, J. Liu, and Z. Chen, “Physics-informed multitask learning for turbofan engine remaining useful life prediction,” *Journal of Aerospace Information Systems*, vol. 21, no. 3, pp. 189–202, Mar. 2024.

[14] Y. Liu, Z. Chen, and X. Wang, “Physics-informed neural networks for turbofan engine remaining useful life prediction,” *Journal of Computational Physics*, vol. 462, p. 111185, 2022.

[15] H. Guo, Y. Zhang, and J. Liu, “Domain adaptation for cross-subset remaining useful life prediction of turbofan engines,” in *Proc. AAAI Conf. Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13245–13252.

[16] M. Zhao, P. Wang, and Y. Chen, “Sensor selection for remaining useful life prediction using attention mechanism,” *Sensors*, vol. 23, no. 12, p. 5567, 2023.

[17] Y. Zhu, J. Li, and H. Huang, “Real-time remaining useful life prediction for turbofan engines using edge computing,” *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 8901–8910, Mar. 2024.

[18] F. Karim *et al.*, “LSTM fully convolutional networks for time series classification,” *IEEE Access*, vol. 6, pp. 166–181, 2018.

[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[20] S. Zheng, A. Farahat, and C. Gupta, “Recurrent neural networks for remaining useful life estimation,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 32, no. 11, pp. 6–15, 2017.