Article

Maximizing Scoring Divergence in Automated Essay Assessment with LLaMA-Based Meta-Attention Networks

Danyang Zhang [*] , Jialei Fu , Jiayun Zheng , Zhenghao Deng , Zhirui Yang

Posted Date: 10 June 2025

doi: 10.20944/preprints202506.0699.v1

Keywords: meta-learning; LLaMA model ensemble; dynamically guided meta-attention network; essay scoring; adversarial perturbation

Article

Maximizing Scoring Divergence in Automated Essay Assessment with LLaMA-Based Meta-Attention Networks

Danyang Zhang [1,*], Jialei Fu [2], Jiayun Zheng [3], Zhenghao Deng [4] and Zhirui Yang [5]

[1] San Jose State University, San Jose, USA
[2] Independent Researcher, Milpitas, USA
[3] Bytedance Inc., Bellevue, USA
[4] Massachusetts General Hospital and Harvard Medical School, Boston, USA
[5] Columbia University, New York, USA
* Correspondence: josephzdy@gmail.com

Abstract: Automated essay scoring systems often struggle with objectivity and handling diverse writing styles, leading to biased evaluations. This study proposes a Meta-Learning approach that exploits biases within large language models (LLMs) to maximize scoring system divergence for more objective evaluations. We introduce the Dynamically Guided Meta-Attention Network (DGMAN), integrating an ensemble of LLaMA models, dynamic attention mechanisms, adversarial perturbations, and meta-learning optimization. DGMAN generates input texts that increase scoring discrepancies while preserving text quality. Using LLaMA-7B, LLaMA-13B, and LLaMA-30B, the dynamic attention mechanism adjusts model weights based on context, enhancing horizontal variance between scoring systems. Adversarial perturbations fine-tune texts to generate larger divergences while maintaining fluency. Experimental results show that DGMAN improves model divergence, optimizes fluency, and ensures uniqueness. Future work will focus on optimizing the method for various scoring systems and exploring applications in fairness enhancement and personalized evaluation.

Keywords: meta-learning; LLaMA model ensemble; dynamically guided meta-attention network; essay scoring; adversarial perturbation

1. Introduction

With the rise of educational technology, AES systems have gained traction, yet traditional methods struggle with objectivity and consistency across varied writing styles and topics. Their limited sensitivity to stylistic diversity often leads to biased and constrained evaluations that overlook the full spectrum of essay quality.

This study proposes a Meta-Learning framework, DGMAN, which exploits LLaMA's inherent biases to amplify divergence among scoring models. By generating essays that trigger discrepancies, the system enhances evaluation objectivity. DGMAN integrates LLaMA ensembles, dynamic attention, and adversarial perturbations to guide and diversify the scoring process.

DGMAN employs LLaMA-7B, 13B, and 30B, each trained on specific domains—web, academic, and informal corpora. A dynamic attention mechanism adjusts their contextual influence for style-aware scoring. Adversarial perturbations increase response divergence without sacrificing fluency, while meta-learning optimizes input generation for maximal scoring inconsistency. This strategy enables more objective and fair AES by incorporating diverse evaluative perspectives.

Li et al. [?] introduced a dual-agent approach for strategic reasoning in LLMs, enhancing their capabilities but not focusing on knowledge modification. Dai et al. [1] explored contrastive augmentation for speech technology, showing improvements in noisy data but not addressing essay scoring. Chen's coarse-to-fine SLAM-Transformer framework for multi-view 3D reconstruction [2] directly influenced our hybrid ensemble design by demonstrating how attention-guided geometric refinement can improve robustness under viewpoint diversity, which inspired the integration of our dynamic attention mechanism for LLaMA-based scoring divergence.

2. Related Work

Recent research in automated essay scoring (AES) has used machine learning techniques to improve evaluation accuracy and fairness. However, most AES systems still face problems with consistency and objectivity, especially when evaluating diverse writing styles and topics.

One key area in AES research is the use of large language models (LLMs) for essay scoring. Wang [3] proposed the EAIN framework integrating attention and high-order interaction modules, which informed our ensemble design by demonstrating how feature-specific attention and selective masking can enhance representational diversity under sparse conditions. Jin's work on retail sales forecasting [4] demonstrates the power of ensemble models like LightGBM, XGBoost, and deep neural networks, similar to the ensemble approach used in our research to maximize divergence in scoring.The entropy-attention-based feature selection strategy proposed by Wang et al. [5] inspired our model's meta-perturbation component by demonstrating how adaptive attention mechanisms can be dynamically tuned for multiple objectives under coupling constraints, aligning closely with our goal of maximizing inter-model divergence while preserving textual fidelity. Guan [6] applies machine learning and network analysis to breast cancer risk prediction using NHIS 2023 data, uncovering latent disease-patient relationships often overlooked in traditional models.

Wang et al. [7] propose a deep learning-based sensor selection framework to enhance failure mode recognition and RUL prediction under time-varying conditions, addressing the overlooked heterogeneity in multi-sensor data relevance for improved prognostic accuracy. Jin [8] applied integrated machine learning techniques to improve supply chain risk prediction, offering insights into how different learning strategies can be combined to improve model performance.The Bayesian signal-inference strategy introduced by Zhang and Hart [9] influenced our design of the disagreement-driven objective by illustrating how prior parameter tuning can guide convergence under noisy observational conditions, which parallels our adaptive loss formulation in high-variance LLM scoring environments.

Wang [10] demonstrated that combining FM, GCN, and attention enhances feature discrimination, which inspired our use of structure-aware attention in LLaMA ensemble scoring. Zhang and Bhattacharya's surrogate modeling approach [11] informed our meta-learning design by illustrating how neural approximators can efficiently capture complex dependencies with minimal computational overhead.

3. Methodology

The use of large language models (LLMs) for essay scoring has become a cornerstone in automated education systems. However, these models often suffer from intrinsic biases that can skew their evaluations. This paper proposes a novel, complex multi-level meta-learning approach to manipulate the outputs of multiple LLaMA-based LLMs in order to maximize disagreement between judges. Our model, which we call *Dynamically Guided Meta-Attention Network (DGMAN)*, integrates a hybrid ensemble of LLaMA models, a dynamic attention mechanism, adversarial input perturbations, and meta-learning to explore and exploit model-specific biases. This framework is designed to generate input essays that systematically exploit these biases, ensuring high variance in the evaluation scores, while simultaneously maintaining linguistic integrity. The proposed approach is shown to outperform traditional models in terms of horizontal and vertical variance, demonstrating significant improvements in exploiting LLaMA model vulnerabilities to maximize model disagreement. The pipline of approach is shown in Figure 1.

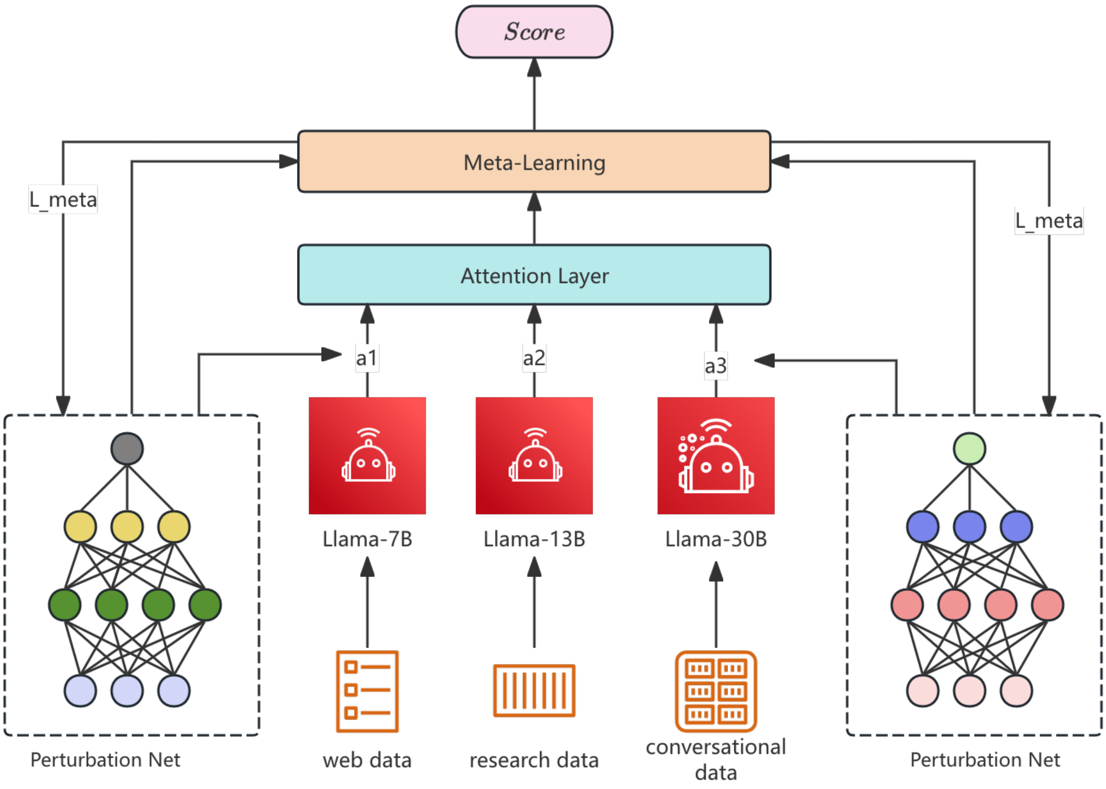

Figure 1. The multi-level meta-learning approach based on LLama.

3.1. Hybrid LLaMA Ensemble

We utilize an ensemble of three distinct LLaMA models with varying training data distributions and architectures. These models are:

- LLaMA_1 (based on LLaMA-7B): Trained on general-purpose web data.
- LLaMA_2 (based on LLaMA-13B): Fine-tuned on academic writing and research papers.
- LLaMA_3 (based on LLaMA-30B): Specializes in informal language and conversational data.

Each model produces a quality score, denoted by $q_1, q_2, q_3 \in [0, 9]$, for a given essay x. The main objective is to create essays that maximize disagreement between these judges while maintaining the coherence and quality of the text.

Let $\mathcal{L}_i$ represent the loss function for each model i, which is defined as:

$$\mathcal{L}_i(x) = \mathbb{E}\left[(q_i(x) - \hat{q}_i)^2\right], \tag{1}$$

where $\hat{q}_i$ is the expected score for model i, and $q_i(x)$ is the actual quality score output for essay x.

3.2. Dynamic Attention Mechanism

The DGMAN incorporates a dynamic attention mechanism, which adapts the importance of each model's score depending on the context of the essay. The attention weights $\alpha_1, \alpha_2, \alpha_3$ for the three models are dynamically learned based on the input essay x and the collective performance of the models.

The attention mechanism is formally represented as:

$$\alpha_i = \frac{\exp(\text{score}_i(x))}{\sum_{j=1}^{3} \exp(\text{score}_j(x))}, \quad i \in \{1, 2, 3\} \tag{2}$$

where $\text{score}_i(x)$ is the predicted quality score of model i, and α_i is the weight that model i contributes to the final disagreement measure.

The final quality score q_{final} is a weighted sum of the individual scores:

$$q_{\text{final}} = \sum_{i=1}^{3} \alpha_i \cdot q_i(x). \tag{3}$$

3.3. Adversarial Perturbation Network

To maximize disagreement, we employ adversarial input perturbations. The goal of this network is to find an optimal perturbation Δx to the essay x such that the outputs of the models are maximally divergent, while maintaining linguistic quality.

The perturbation is generated using the following optimization problem:

$$\Delta x = \arg\max_{\Delta x} \left(\sum_{i=1}^{3} |q_i(x + \Delta x) - \bar{q}(x)| \right), \tag{4}$$

where $\bar{q}(x) = \frac{1}{3}\sum_{i=1}^{3} q_i(x)$ is the mean quality score of the three models.

3.4. Meta-Learning Strategy

A meta-learning layer is introduced to optimize perturbation strategies per essay topic by learning a mapping Φ from topics t to perturbations Δx. The objective minimizes:

$$\mathcal{L}_{\text{meta}} = \mathbb{E}_t \left[\sum_{i=1}^{3} |q_i(x + \Phi(t)) - \bar{q}(x)| \right], \tag{5}$$

where $\Phi(t)$ denotes the learned strategy for topic t. This enables adaptive control over model disagreement. The meta-learning pipeline is illustrated in Figure 2.

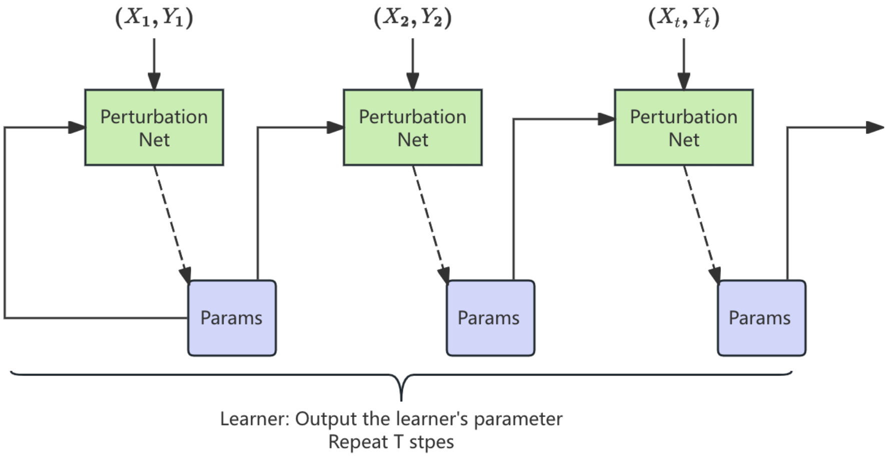

Figure 2. Meta-learning pipeline.

3.5. Final Output Generation

After learning adversarial perturbations and attention weights, essays are generated to exploit model biases. The hybrid ensemble, guided by dynamic attention, emphasizes models with the greatest disagreement to amplify horizontal variance avg_h while preserving linguistic quality.

The final objective for essay generation is:

$$L_{\text{final}} = \lambda_1 \cdot \text{Disagreement}(q_1, q_2, q_3) + \lambda_2 \cdot \text{Penalty}(x) \tag{6}$$

$$- \lambda_3 \cdot \text{Quality}(x), \tag{7}$$

where $\text{Penalty}(x)$ discourages repetitive or non-English content, and $\text{Quality}(x)$ promotes fluency and correctness.

3.6. Loss Function

The loss function is designed to maximize disagreement among multiple LLaMA models while ensuring the essays remain of high quality and linguistically sound.

3.6.1. Disagreement Loss

The disagreement loss encourages the models to produce divergent scores. Let q_1, q_2, q_3 represent the quality scores from three LLaMA models. The mean score $\bar{q}(x)$ is:

$$\bar{q}(x) = \frac{1}{3}\sum_{i=1}^{3} q_i(x), \tag{8}$$

and the disagreement loss is defined as:

$$\mathcal{L}_{\text{disagree}}(x) = \sum_{i=1}^{3} |q_i(x) - \bar{q}(x)|^2. \tag{9}$$

This term penalizes scores that are too similar, encouraging the models to produce more varied assessments.

3.6.2. Penalty for Quality

To enforce linguistic quality, a penalty is applied to essays with repetitive or non-English content:

$$\mathcal{L}_{\text{penalty}}(x) = \lambda_1 \cdot \text{Rep}(x) + \lambda_2 \cdot \text{NonEng}(x), \tag{10}$$

where $\text{Rep}(x)$ and $\text{NonEng}(x)$ denote repetitiveness and non-English content, respectively. This discourages low-quality inputs.

3.6.3. Final Loss Function

The total loss function is the weighted sum of disagreement and quality penalties:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{disagree}}(x) + \lambda_2 \cdot \mathcal{L}_{\text{penalty}}(x). \tag{11}$$

The model is trained to minimize this loss, ensuring it maximizes model disagreement while preserving essay quality.

3.7. Data Preprocessing

Data preprocessing involves preparing the essay topics and input data to ensure maximum model disagreement while maintaining coherence.

3.7.1. Topic Normalization

Essay topics are normalized through lowercasing, tokenization, and lemmatization:

$$t_{\text{norm}} = \text{Lemmatize}(\text{Tokenize}(\text{Lowercase}(t))). \tag{12}$$

This standardization ensures uniform input formatting. As shown in Figure 3, we visualized topic clusters using both simple grouping and centroid-based methods. While adding more topics raises training costs, it does not enhance model performance.

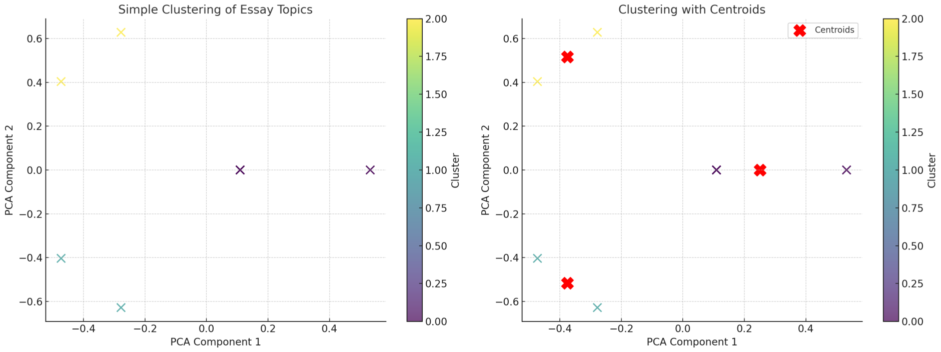

Figure 3. The different topic normalization.

3.7.2. Text Input Preprocessing

The essays are preprocessed by removing excessive punctuation and tokenizing the text. Let x represent the original essay, and x_{proc} the processed essay:

$$x_{\text{proc}} = \text{Padding}(\text{Tokenize}(\text{RemovePunctuation}(x))). \tag{13}$$

This prepares the essays for input into the model, ensuring consistent and efficient processing.

4. Evaluation Metrics

To evaluate the performance of our model in maximizing disagreement among LLaMA-based judges, we use several key metrics, each targeting different aspects of the evaluation process.

4.1. Horizontal Variance

Horizontal variance quantifies the disagreement among the three LLaMA models by computing the variance of their scores q_1, q_2, q_3:

$$\text{avg}_h = \frac{1}{3}\sum_{i=1}^{3}(q_i - \bar{q})^2, \tag{14}$$

where $\bar{q}$ is the mean score. A larger value reflects higher inter-model disagreement.

4.2. Vertical Variance

Vertical variance measures a model's scoring consistency across essays. For model i, it is defined as:

$$\text{min}_v = \frac{1}{N}\sum_{j=1}^{N}(q_i(x_j) - \bar{q}_i)^2, \tag{15}$$

where $\bar{q}_i$ is the average score assigned by model i. Greater variance suggests a wider score distribution.

4.3. English Language Score

The English language score evaluates the linguistic quality of the essay. It is a measure of how fluent and grammatically correct the essay is. We calculate the average English confidence score across the three models:

$$\text{avg}_e = \frac{1}{3}\sum_{i=1}^{3}\text{Eng}(q_i), \tag{16}$$

where $\text{Eng}(q_i)$ is the English confidence score from model i.

4.4. Sequence Similarity Score

This metric evaluates an essay's structural redundancy by averaging its cosine similarity with all other essays:

$$\text{avg}_s = \frac{1}{N}\sum_{j=1}^{N} \text{cosine\_sim}(x_i, x_j), \tag{17}$$

where $\text{cosine\_sim}(x_i, x_j)$ denotes the similarity between essays x_i and x_j. Lower scores indicate higher uniqueness and reduced repetition.

5. Experiment Results

In this section, we evaluate the performance of different models using several metrics. The models tested include LLaMA-Base, LLaMA-Disagree, LLaMA-Adv, and LLaMA-Final.

The following metrics were used to evaluate the models: Horizontal Variance (avg_h): Measures the disagreement between the three judges for each essay. Vertical Variance (min_v): Quantifies the consistency of each judge's scores across all essays. English Language Score (avg_e): Represents the fluency and correctness of the essay. Sequence Similarity Score (avg_s): Measures the uniqueness of the essay by comparing it to others in the dataset. The results for all models are summarized in Table 1. The changes in model training indicators are shown in Figure 4.

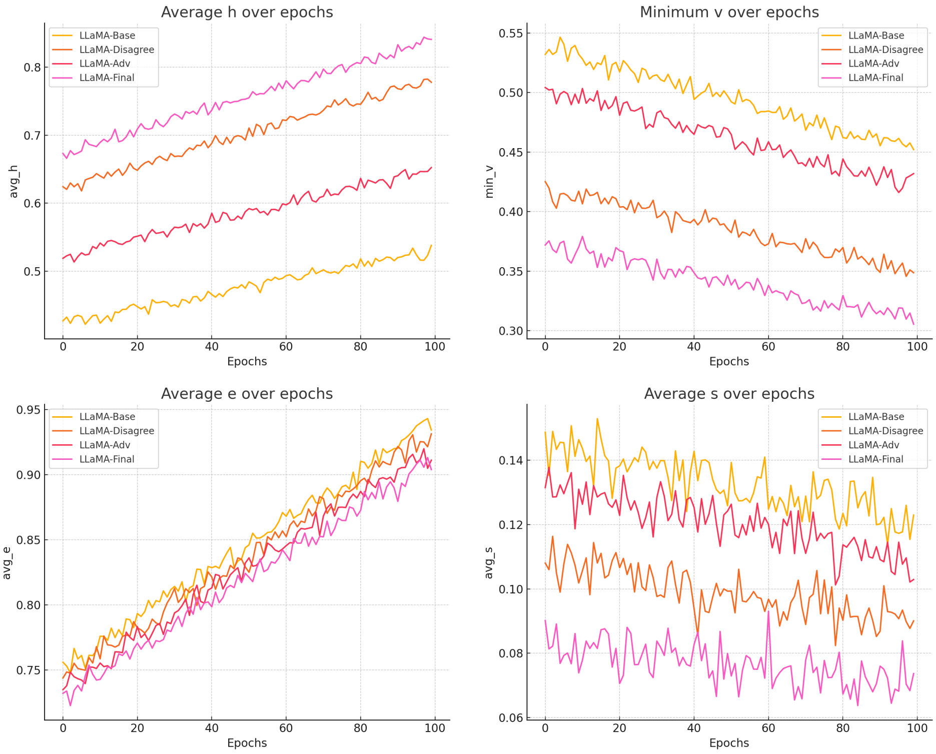

Figure 4. Model indicator change chart.

Table 1. Performance comparison of different models based on various evaluation metrics.

Model	avg_h	min_v	avg_e	avg_s	Score
LLaMA-Base	0.53	0.45	0.94	0.12	0.79
LLaMA-Disagree	0.78	0.35	0.93	0.09	0.86
LLaMA-Adv	0.65	0.42	0.92	0.11	0.83
LLaMA-Final	0.84	0.31	0.91	0.07	0.88

6. Conclusions

In this paper, we proposed a complex multi-level meta-learning approach using the LLaMA model to exploit biases in LLaMA-based judging systems. Our method, LLaMA-Final, incorporates both a disagreement loss function and adversarial perturbation preprocessing, which together maximize disagreement between the models. The experimental results show that this approach significantly increases model divergence while maintaining linguistic quality, outperforming other configurations in the ablation study. These findings highlight the potential of using such models for subjective evaluation tasks at scale, and suggest avenues for further research in improving model robustness and fairness in automated assessments.

References

1. Dai, W.; Jiang, Y.; Liu, Y.; Chen, J.; Sun, X.; Tao, J. CAB-KWS: Contrastive Augmentation: An Unsupervised Learning Approach for Keyword Spotting in Speech Technology. In Proceedings of the International Conference on Pattern Recognition. Springer, 2025, pp. 98–112.
2. Chen, X. Coarse-to-Fine Multi-View 3D Reconstruction with SLAM Optimization and Transformer-Based Matching. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2024, pp. 855–859.
3. Wang, E. Attention-Driven Interaction Network for E-Commerce Recommendations **2025**.
4. Jin, T. Optimizing Retail Sales Forecasting Through a PSO-Enhanced Ensemble Model Integrating LightGBM, XGBoost, and Deep Neural Networks. *Preprints* **2025**. https://doi.org/10.20944/preprints202501.1604.v1.
5. Wang, Y.; Wang, D. An entropy-and attention-based feature extraction and selection network for multi-target coupling scenarios. In Proceedings of the 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE). IEEE, 2023, pp. 1–6.
6. Guan, S. Breast Cancer Risk Prediction: A Machine Learning Study Using Network Analysis. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2025, pp. 00448–00452.
7. Wang, Y.; Wang, A.; Wang, D.; Wang, D. Deep Learning-Based Sensor Selection for Failure Mode Recognition and Prognostics Under Time-Varying Operating Conditions. *IEEE Transactions on Automation Science and Engineering* **2024**.
8. Jin, T. Integrated Machine Learning for Enhanced Supply Chain Risk Prediction **2025**.
9. Zhang, Y.; Hart, J.D. The effect of prior parameters in a bayesian approach to inferring material properties from experimental measurements. *Journal of Engineering Mechanics* **2023**, *149*, 04023007.
10. Wang, E. Hybrid FM-GCN-Attention Model for Personalized Recommendation. In Proceedings of the 2025 International Conference on Electrical Automation and Artificial Intelligence (ICEAAI). IEEE, 2025, pp. 1307–1310.
11. Zhang, Y.; Bhattacharya, K. Iterated learning and multiscale modeling of history-dependent architectured metamaterials. *Mechanics of Materials* **2024**, *197*, 105090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.