A Study on an Explainable Causal-Enhanced LLM Agent for Predicting the Forming Quality of Automotive Component Materials

Jingcheng Zhao*

University of Toronto, Toronto, M5S 1A1, Canada, jingchengzhao0@gmail.com

Jiayu Fan

University of Toronto, Toronto, M5S 1A1, Canada, jiayu.fan@alumni.utoronto.ca

Liangze Li

Carnegie Mellon University, Pittsburgh, PA 15213, United States, liangzel@alumni.cmu.edu

Abstract

Raw material composition, heat treatment, cooling, equipment parameters and die condition have an impact on the forming quality of materials for automotive components. Current prediction models can detect the defect or deviation in the performance, but are not so much supported for finding out the root cause of anomaly and for making the changes in the process. In this paper, a causal enhanced LLM Agent explainable analysis framework is proposed. The study builds a multi-source manufacturing dataset that includes the composition of raw materials, heat treatment curves, forming pressure, cooling time, die condition, inspection outcome, and rework records;Random Forest, Multi-Layer Perceptron, XGBoost, and TabNet are used to predict strength deviation, surface defect grades, and batch pass rates;SHAP, NOTEARS, and counterfactual reasoning are combined to identify the key process variables and their influence paths;In addition, a RAG-LLM Agent is built to fuse the process specifications, material handbooks, historical anomaly cases, and model interpretation results, which helps to generate quality root cause analysis and process adjustment recommendation. As an example, a material forming manufacturing line for automotive parts had

68,000 batch records, 42 process variables and 11 quality metric categories included in the experiment. Across ten repeated group–stratified evaluations, XGBoost achieved a macro–F1 score of 0.892 with a 95% confidence interval of 0.888—0.896, compared with 0.838 for Random Forest, 0.856 for MLP, and 0.874 for TabNet. The increase over Random Forest was 0.054 in absolute terms and 6.4% in relative terms, and the paired bootstrap test confirmed that the difference was statistically significant at $p<0.001$.The causal constraint also improved the key variable identification consistency by 18.7%, Among 200 independently reviewed reports, the Agent's primary root–cause conclusion matched the engineer consensus in 169 cases, yielding an exact agreement rate of 84.5% and a Cohen's kappa of 0.781. The time taken to identify problems was also reduced from 47 minutes to 19 minutes. The study shows that output of causal reasoning and the LLM Agent can improve interpretability of the prediction of material quality and the impact of process interventions.

CCS CONCEPTS:

Computing methodologies~Artificial intelligence~Knowledge representation and reasoning~Causal reasoning and diagnostics

Keywords

automotive components; material forming quality; LLM Agent; causal inference; XGBoost; TabNet; SHAP; Random Forest; multi–layer perceptron; explainable artificial intelligence

1 Introduction

The factors of the material forming process of automotive components are multi–dimensional, such as: the material composition, heat treatment process, forming pressure, cooling speed, forming die conditions, equipment operation parameters, etc. Quality deviations are generally seen as variations in strength, surface defects and lower batch pass rates. Traditional quality prediction

method emphasizes result evaluation, but cannot show the path how variables affect the result, and it is difficult to trace back to the source of the abnormality, cannot clarify the boundary for the process adjustment, and cannot meet the requirements of process optimization in high consistence manufacturing scenarios. Therefore, we propose a causal-enhanced LLM Agent explainable method which combines multi-model prediction, SHAP variable explanation, NOTEARS causal structure learning, counterfactual reasoning and retrieval-augmented reasoning to do prediction, explanation and intervention analysis on the quality of material forming in automotive components. This approach is intended to make better predictions of quality, better stability of key variable identification and easier process recommendations.

2 Data Representation and Problem Modeling for Automotive Component Material Forming Quality Prediction

To predict the forming quality of the materials used for automotive components, it is necessary to translate the heterogeneous production line data into a vector of states that can be used in a computer. Data representation includes 18 fields of raw material composition information, 6 heat treatment temperature curves, information on equipment sampling with a second-by-second resolution, forming pressure, holding time, cooling time, mold temperature, mold wear, surface inspection grayscale features and rework labels, and so on[1]. Set to i The input vectors for each batch are:

$$\mathrm{x}_i = [\mathrm{r}_i, \mathrm{h}_i, \mathrm{p}_i, \mathrm{m}_i, \mathrm{q}_{i\text{-}1}], \hat{\mathrm{y}}_i = f_\theta(\mathrm{x}_i), \min_\theta J = L(\hat{\mathrm{y}}_i, \mathrm{y}_i) + \lambda \left\| \mathrm{A} \nabla_{\mathrm{x}} \hat{\mathrm{y}}_i \right\|_1 \tag{1}$$

where r_i represents the material composition vector, h_i represents the heat treatment timing curve, p_i represents pressure, velocity, and time parameters, m_i represents mold status, $\mathrm{q}_{i\text{-}1}$ represents quality feedback from the previous batch, $\hat{\mathrm{y}}_i$ is the predicted quality output, f_θ is the

prediction model with parameters θ , L represents the supervised loss, A represents the causal adjacency constraint matrix, and λ are the constraint weights.This model provides a common input for the subsequent XGBoost, TabNet, and causal explanation modules.

3 Design of a Causal-Enhanced LLM Agent Method for Predicting Automotive Component Material Forming Quality

3.1 Construction of a Multi-Model Prediction Architecture for Material Forming Quality Prediction

For the prediction of material forming quality, the multi-model architecture uses the material type, the process window, and the quality label as the main input flow, which includes 22MnB5 boron steel, 40Cr, 42CrMo, DP780 high strength steel, 6061 1-t6 Al alloy, 7075-t6 Al alloy, ADC12 die-cast aluminium, AZ91D magnesium alloy, QT450-10 ductile iron, GCr15 bearing steel, and PA66-GF30 reinforced plastics (see Figure 1 for more details). The Random Forest, MLP, XGBoost, and TabNet are used in parallel prediction branches, and the tasks of non-linear fitting, sparse and structured variable selection are assigned to different models [2].

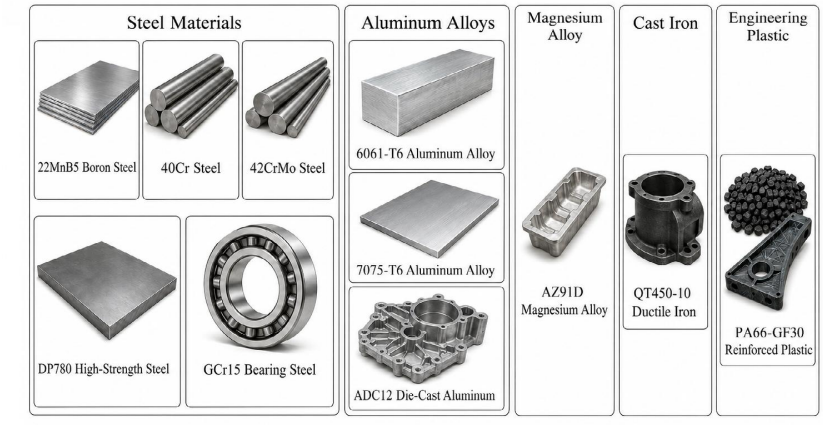

Figure 1: Material type grouping diagram for automotive component forming quality prediction

3.2 SHAP-Based Explanation Method for Key Variables in Material Forming Quality

The SHAP explanation layer creates a variable contribution map for material composition, heat treatment, forming pressure, cooling parameters, and mold conditions after the multi-model

prediction structure produces a quality output. This includes composition fields such as C, Si, Mn, Cr, and Mo; the austenitizing temperature range of 850 — 930 ° C; the 12 — 28 MPa forming pressure window, the 8 — 35 s holding time, the 15 — 60 ° C/s cooling rate, the 0.02 — 0.18 mm die wear, and the rework code into a unified explanatory space [3] . The contribution values of key variables are calculated using the following formula:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|\left(M - |S| - 1\right)}{M} \left[f\left(S \cup \{j\}\right) - f\left(S\right) \right] \tag{2}$$

where ϕ_j represents the marginal contribution of the th process variable (j) to the quality prediction output; F denotes the set of all input variables; M denotes the total number of variables; S denotes the feature subset excluding variable j ; $f\left(S\right)$ denotes the model's prediction function under the feature subset S ; $f\left(S \cup \{j\}\right) - f\left(S\right)$ denotes the change in predicted response after adding variable j . The interpretation results are further classified according to the strength deviation, surface defect grade, and batch pass rate, which provides the basis for the selection of the NOTEARS causal structure and the generation of the LLM Agent process cause [4].

3.3 Design of a Causal Enhancement Mechanism Integrating NOTEARS and Counterfactual Reasoning

Because the manufacturing dataset contains nonlinear interactions and mixed-type variables, the causal module adopts a nonlinear mixed-type extension of NOTEARS rather than the linear Gaussian formulation. Continuous variables, including temperature, pressure, cooling rate, holding time, mold wear, and surface roughness, are modeled by nonlinear structural functions, whereas categorical variables, including material grade and rework code, are modeled by a categorical conditional likelihood. For each variable X_j , the structural relationship is expressed as:

$$X_j = f_j(X_{\mathrm{pa}(j)}; \theta_j) + \varepsilon_j \tag{3}$$

where $f_j(\cdot)$ is implemented by a multilayer perceptron and $X_{\mathrm{pa}(j)}$ denotes the parent variables of X_j. The causal graph is estimated by:

$$\min_{\cdot,W}\left[\sum_j \frac{1}{2n}\left\|X_j - f_j(X;\theta_j)\right\|_2^2 - \eta\sum_{j\in D}\log p\left(X_j\middle|X_{\mathrm{pa}(j)}\right) + \lambda_1\|W\|_1 + \lambda_2\|\cdot\|_2^2\right] \tag{4}$$

subject to:

$$h(W) = \mathrm{tr}\left[\exp(W\circ W)\right] - d = 0 \tag{5}$$

where C and D denote the continuous and categorical variable sets, respectively. Process-order constraints are imposed to prevent quality outcomes and rework results from pointing to upstream material or forming parameters.

The causal constraint is further incorporated into variable identification. For the b-th bootstrap sample, the final score of variable j is calculated as:

$$r_j^{(b)} = \alpha\tilde{s}_j^{(b)} + (1-\alpha)\tilde{c}_j^{(b)} \tag{6}$$

where $\tilde{s}_j^{(b)}$ is the normalized mean absolute SHAP value, $\tilde{c}_j^{(b)}$ is the normalized causal-path support from variable j to the quality output, and α is selected on the validation set. The counterfactual module then searches only over actionable variables, including forming pressure, holding time, and cooling rate, while material grade and historical quality labels remain fixed. This design prevents the Agent from generating infeasible interventions and provides explicit structural evidence for quality root-cause analysis[5-6] .

3.4 An Explainable Reasoning Framework for RAG-LLM Agents Aimed at Quality Root Cause Analysis

The knowledge base integrates three types of manufacturing information through a shared metadata schema. Structured process records are converted into key-value documents containing batch ID, material grade, process stage, parameter name, measured value, unit, timestamp, and

quality label. Unstructured manuals are segmented by material category, section heading, and process stage, with each fragment retaining its document title and parameter range. Semi–structured historical cases are normalized into five fields: anomaly symptom, verified cause, supporting evidence, corrective action, and observed outcome. Material grade, process stage, defect type, and parameter name are used as common metadata keys.Retrieval is performed in two stages. Metadata filtering first removes evidence that does not match the current material and process stage. Dense semantic retrieval and keyword retrieval are then combined, after which the candidate fragments are reranked according to relevance to the predicted defect and identified causal variables. The structured prompt contains six components: system role, current batch context, prediction result, SHAP and causal evidence, retrieved evidence with source identifiers, and counterfactual adjustment boundaries. Engineer–validated cases from the training partition are used as few–shot demonstrations, while test cases are excluded from prompt construction.The Agent is required to return six structured fields: predicted quality state, key process variables, supported causal path, retrieved evidence, recommended parameter interval, and uncertainty or risk statement. It is prohibited from recommending values outside the retrieved process specification or from introducing causes unsupported by either the causal graph or retrieved evidence. This design unifies heterogeneous knowledge sources and makes the generated report traceable to both production data and documentary evidence.The module connections in this reasoning model are shown in Figure 2.

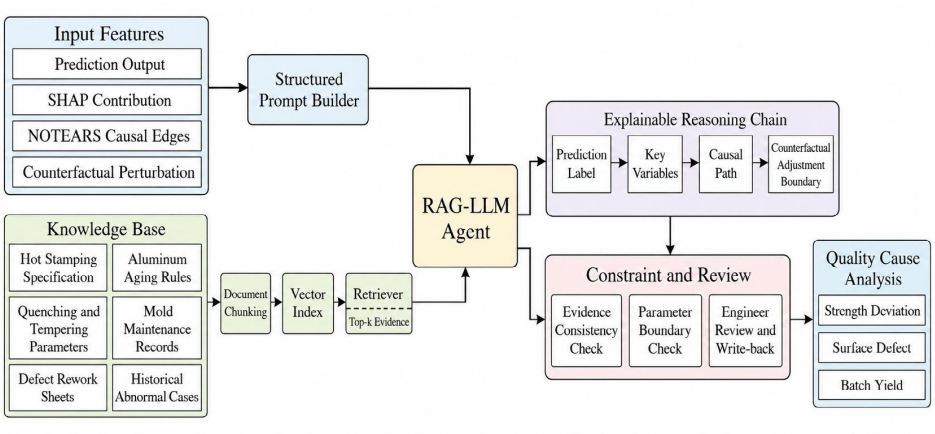

Figure 2: RAG–LLM Agent Explainable Reasoning Framework for Quality Root Cause Analysis

4 Experimental Results and Analysis

4.1 Experimental Design

The experiment was conducted using 68,000 batch records from an automobile component material forming line. The data includes the components of the raw material, the heat treatment curve, the forming pressure, the cooling time, the mold condition, the equipment parameters, the test results, and the rework record.A total of 42 variables are input, including C, Si, Mn, Cr, Mo, heat treatment temperature window 850 — 930 ° C; a holding time of 8 — 35 s; a cooling rate of 15 — 60 ° C/s; and process characteristics such as forming pressure, mold wear, mold temperature, equipment operation state, and batch retreatments.The output contains 11 quality indicators, with strength deviation, surface defect severity, and batch pass rate selected as the principal tasks. Surface defects were coded into four ordered levels: Grade 0 for no visible defect, Grade 1 for a minor defect, Grade 2 for a moderate defect requiring local rework, and Grade 3 for a severe defect requiring batch rejection. Because the distances between these levels are operationally different, model performance was evaluated using macro–F1, quadratic weighted kappa, ordinal mean absolute error, and within–one–grade accuracy. Batch–grouped partitioning was applied to prevent adjacent records from the same production sequence from entering different datasets[8].

4.2 Comparative Analysis of Multi-Model Material Forming Quality Prediction Performance

The four models were evaluated using ten repeated group-stratified data splits to prevent adjacent batches from entering different subsets. Random Forest obtained a macro-F1 score of 0.838±0.006 with a 95% confidence interval of 0.834—0.842. MLP reached 0.856±0.008 with a confidence interval of 0.850—0.862, whereas TabNet achieved 0.874±0.006 with a confidence interval of 0.870—0.878. XGBoost produced the highest macro-F1 score of 0.892±0.005, with a confidence interval of 0.888—0.896. The absolute improvement over Random Forest was 0.054, corresponding to a relative increase of 6.4%.Paired bootstrap comparisons based on 1,000 resamples showed that XGBoost significantly outperformed Random Forest, MLP, and TabNet. The macro-F1 differences were 0.054, 0.036, and 0.018, with $p<0.001$, $p<0.001$, and $p=0.002$, respectively. TabNet performed better than MLP because its attentive feature-selection mechanism reduced the influence of redundant process variables. XGBoost remained superior because its tree-boosting structure captured threshold effects and nonlinear interactions among material composition, heat-treatment temperature, forming pressure, cooling rate, and die condition [9–10].

4.3 Analysis of Key Process Variable Contributions via SHAP Explanation

For further explanation of the variable interpretation, Figure 3 shows the distribution of the contribution degrees of the key process variables with SHAP interpretation.

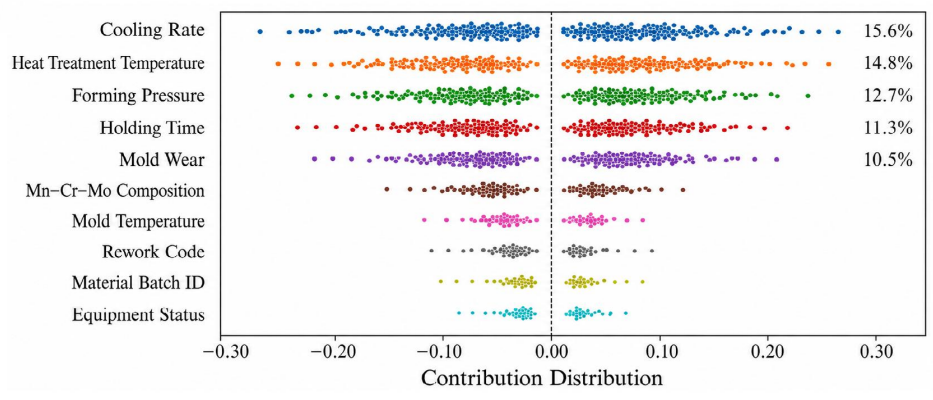

Figure 3: Contribution Analysis of Key Process Variables via SHAP Explanation

As shown in Figure 3, the contributions mainly focus on the continuous process parameters, such as the cooling rate, the heat treatment temperature, the forming pressure, the retention time, and the wear of the mould. The cooling rate is 15.6%, the heat treatment temperature is 14.8%, the forming pressure is 12.7%, the retention time is 11.3%, and the mould wear is 10.5%. This shows that the variation in the quality of the material is more significantly affected by the thermal and mechanical coupling processes and the contact conditions of the mould. The Mn–Cr–Mo component, the mold temperature, the rework code, the batch number of the material, and the operating state of the equipment are all complementary to the information about the source of the material, the stability of the device, and the historical anomaly. The consistency of the key variable identification was improved by 18.7% after the application of the NOTEARS causal constraint, which showed that the SHAP contribution order could provide a stable input for the selection of the cause and the RAG–LLM Agent–based explanation of the quality cause.

4.4 Analysis of Variable Identification Consistency and Intervention Effectiveness Under the Causal Augmentation Mechanism

In order to further assess the effect of NOTEARS and counterfactual inference on the identification of key variables and the stability of intervention, Figure 4 shows the box and whisker plots of each measure.

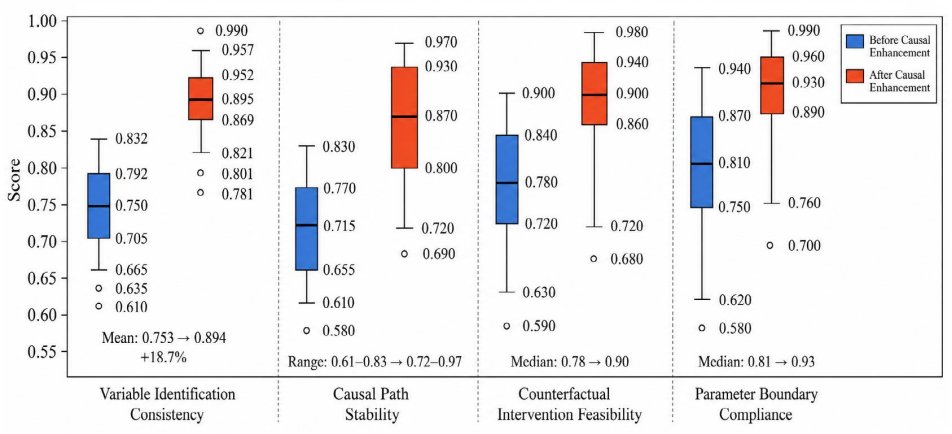

Figure 4: Box plots of variable identification consistency and intervention effectiveness under the causal enhancement mechanism

Figure 4 shows that in a test window containing 68,000 batch records, 42 process variables, and 11 quality metrics, the identification consistency of the SHAP variable was mainly in the range of 0.665 to 0.832, with an average of about 0.753 when causality was not used. Including the NOTEARS causality constraints, the average consistency improved to 0.894, which is 18.7% improvement.Path stability shifted from 0.61 — 0.83 to about 0.72 — 0.97, suggesting that the causal chain between cooling rate, heat treatment temperature, forming pressure, and mold wear has become more concentrated, indicating that the proposed adjustment for pressure, retention time, and cooling rate is better aligned with the process window, providing a stable constraint for the RAG–LLM Agent to produce Quality Cause Analysis and Parameter Correction Recommendations.

4.5 Effectiveness of the LLM Agent's Quality Cause Explanations and Process Adjustment Recommendations

A total of 200 quality analysis reports were randomly selected from the independent test set. Two process engineers with more than eight years of experience independently reviewed the predicted root cause, supporting process evidence, and recommended adjustment range. Disagreements between the two engineers were resolved through joint review, and the resulting

consensus labels were used as the reference standard. The Agent's primary root-cause conclusion exactly matched the engineer consensus in 169 of the 200 reports, corresponding to an exact agreement rate of 84.5%. Cohen's kappa between the Agent output and the reference labels was 0.781, with a 95% confidence interval of 0.724—0.838, indicating substantial agreement beyond chance.The highest agreement occurred in cases involving insufficient cooling, heat-treatment temperature deviation, unstable forming pressure, and excessive die wear. The average issue-localization time decreased from 47 min under manual analysis to 19 min with Agent assistance. The results indicate that retrieved process evidence, causal-path constraints, and counterfactual adjustment limits improved the consistency and operational relevance of the generated reports

4.6 Analysis of Ablation Experiments for Different Modules

In order to validate the contribution of each module to the prediction of material formation quality and explanatory reasoning, Table 1 shows the ablative results after removing key modules.

Table 1: Comparison of Ablation Experiment Results for Different Modules

Experimental Configuration	Defect Grade F1 Score	Variable Identification Consistency	Report Review Consistency Rate	Issue Localization Time
Complete Framework: XGBoost + SHAP + NOTEARS + Counterfactual Reasoning + RAG-LLM Agent	0.892	18.7% improvement	84.50%	19 min
Removing NOTEARS causal constraints	0.892	No improvement of 18.7%	76.80%	27 min
Removing counterfactual reasoning	0.892	Variable ordering retained, intervention boundaries weakened	78.30%	31 min
Removing RAG Retrieval Augmentation	0.892	Causal recognition retained	69.60%	34 min
Random Forest Baseline	Approx. 0.838	Without causal constraints	—	47 min

As shown in Table 1, when the complete framework is used, the F1 score of defect grade prediction is still 0.892, and the consistency rate of key variable identification is also improved by 18.7% with the adoption of the constraint of NOTEARS; without it, the causal edge weights between the cooling rate and the heat treatment temperature, the forming pressure, the mold wear cannot be converged, and the consistency rate of the report review is reduced to 76.8%. With the removal of counterfactual reasoning, the model was still able to identify key variables, however the adjustment boundaries for pressure, holding time and cooling rate were not enforceable and the localization time became 31 min. The results revealed that without RAG retrieval augmentation, the Agent was less effective at retrieving process specifications, historical anomaly cases, and the review consistency dropped to 69.6%, which demonstrates that the ability to explain quality causes is supported by the three factors: the prediction model, causal constraints, and retrieval augmentation.

5 Conclusions

The prediction of forming quality of automotive component materials is a complex problem due to multiple sources of process information, the nonlinearities between the variables and the fact that quality has multiple causes. The explainable analysis framework, which is implemented using XGBoost, SHAP, NOTEARS, counterfactual reasoning, and the RAG–LLM Agent, provides a synergy of defect severity prediction, key variable identification, causal path constraints and process recommendations generation. Experimental results illustrate that this framework has engineering application value in terms of prediction accuracy, stability in identification of the variables, consistency in the review of the report and efficiency in problem localization. Its innovation is in the representation of causal structure learning in the material forming quality explanation process. Because of the limited scale data on a single production line and limited material types, the

cross-factory transferability of the model needs to be further verified. The future research can be extended to multi-production line, multi-material, and online closed-loop control scenarios, which can increase the adaptability in the complex manufacturing cases.

References

[1] Raghunathan V, Sathyamoorthy G, Ayyappan V, et al. Advances in brake friction materials: A comprehensive review of ingredients, processing methods, and performance characteristics[J]. Journal of Vinyl and Additive Technology, 2024, 30(6): 1396-1431.

[2] Lee J, Rew J. Multi-Agent Large Language Model-Based Decision Tree Analysis for Explainable Electric Vehicle Drive Motor Fault Diagnosis[J]. Computers, Materials & Continua, 2026, 87(3): 1.

[3] Zhang L, Liu Z, Ni B, et al. Large language models (llms) for materials design[J]. Advanced Functional Materials, 2026, 36(30): e25897.

[4] Yuan W, Chen G, Wang Z, et al. Empowering generalist material intelligence with large language models[J]. Advanced Materials, 2025, 37(32): 2502771.

[5] Wang W Y, Zhang S, Li G, et al. Artificial intelligence enabled smart design and manufacturing of advanced materials: The endless Frontier in AI+ era[J]. Materials Genome Engineering Advances, 2024, 2(3): e56.

[6] Lee J, Lee J, Rew J. Multimodal Large Language Model-Based Explainable Boosting Machine Analysis for Interpretation of State-of-Health Prediction of Lithium-Ion Batteries[J]. Electronics, 2026, 15(8): 1675.

[7] Pérez-Rosero D A, Pineda-Quintero S, Álvarez-Barreto J C, et al. An Interpretable Artificial Intelligence Approach for Reliability and Regulation-Aware Decision Support in Power Systems[J]. Computation, 2025, 14(1): 2.

[8] Buehler M J. MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities[J]. Applied Mechanics Reviews, 2024, 76(2): 021001.

[9] Zhao L, Liu T, Kim H S, et al. Alloy design paradigms in additive manufacturing: a new era of material innovation[J]. International Journal of Extreme Manufacturing, 2026, 8(3): 033003.

[10]Marandi S, Hu Y S, Modarres M. Complex system diagnostics using a knowledge graph-informed and large language model-enhanced framework[J]. Applied Sciences, 2025, 15(17): 9428.