Evaluating Large Language Model-Generated Explanations for Defective-Part Prediction in Injection Moulding Using Feature-Attribution Evidence

Seyed Iman Hosseini
iman.hosseini@ijs.si
Department for Artificial Intelligence, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Lorenzo Seghesio
lorenzo.seghesio@kuleuven.be
Department of Mechanical Engineering, KU Leuven
Flanders Make@KU Leuven
Leuven, Belgium

Konstantinos Gryllias
konstantinos.gryllias@kuleuven.be
Department of Mechanical Engineering, KU Leuven
Flanders Make@KU Leuven
Leuven, Belgium

Dunja Mladenić
dunja.mladenic@ijs.si
Department for Artificial Intelligence, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

This exploratory study examines LLM-generated explanations for injection-moulding defective-part predictions using local SHAP evidence. CatBoost detected two of ten held-out failures, both short-shot defects. GPT-5.4 mini, Gemini 3.1 Flash-Lite, and Mistral Devstral 2512 each explained both cases, yielding six explanations. One blinded LLM judge and one manufacturing expert assessed evidence faithfulness, attribution caution, and manufacturing-process relevance. The judge assigned both GPT explanations maximum attribution-caution scores, whereas the expert assigned minimum scores for unsupported process-value claims. This observed disagreement illustrates a potential limitation of automated explanation evaluation. With two cases of one defect type and one evaluator of each kind, the findings cannot establish a general ranking of LLMs.

Keywords

Explainable Artificial Intelligence, Large Language Models, injection moulding, model attribution, LLM as Judge

1 Introduction

Injection moulding is widely used for the large-scale manufacturing of plastic components ranging from standard to high-precision parts, but variations in process conditions can lead to quality defects and inconsistent production outcomes [6, 7]. In this study, the term *failure* refers to a production cycle that produces a defective or nonconforming part, rather than to a mechanical failure or breakdown of the injection-moulding machine. Traditional approaches such as Design of Experiments and Analysis of Variance help identify relevant process factors, but they may be limited in representing complex and nonlinear interactions among process variables [7]. Machine-learning studies now use XAI tools like SHAP and LIME to highlight key features and explain injection-moulding quality predictions [6, 7]. XAI-based predictive-quality studies have shown that process measurements, time-dependent information, and operator-adjustable parameters can provide useful evidence for explaining model outputs [6]. Local model-attribution methods are particularly useful because they indicate which variables most strongly influenced a prediction for an individual production cycle [1, 7]. However, model-attribution evidence explains the behavior of the predictive model rather than establishing the physical cause of a defect [1]. Furthermore, the direction of a feature attribution does not by itself indicate that the corresponding process measurement was high, low, abnormal, or outside an acceptable operating range [1]. Large Language Models (LLMs) provide a potential interface for translating technical model-attribution evidence into concise explanations for operators and quality-control personnel [1, 4, 8]. Previous studies have applied LLMs to fault diagnosis, process monitoring, and human–machine collaborative troubleshooting in industrial environments [4, 8]. Research on LLM-based explainability also suggests that LLMs can make technical evidence more accessible to non-specialist users [1]. Nevertheless, fluent natural-language explanations may introduce information that is absent from the supplied evidence, omit influential variables, reverse the direction of a contribution, or present model attribution as proof of physical causation [1, 4]. Therefore, readability alone is insufficient; generated explanations must also be evaluated for their faithfulness to the supplied evidence and for the caution with which they communicate it [1, 2]. Although LLMs are increasingly used to communicate model outputs and support fault diagnosis, limited attention has been given to whether their explanations accurately preserve local attribution evidence in manufacturing settings [1, 4, 8]. In particular, it remains unclear whether an LLM can correctly distinguish variables that push a prediction toward *Fail* from those that push it toward *Pass*, while avoiding unsupported statements about process values or physical root causes [1].

This exploratory study examines attribution errors and judge–expert disagreement in explanations of two short-shot cases detected among ten held-out failures. GPT-5.4 mini, Gemini 3.1 Flash-Lite, and Mistral Devstral 2512 produce six explanations, assessed by one blinded LLM judge and one manufacturing expert

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.95

for evidence faithfulness, attribution caution, and manufacturing-process relevance [2]. The findings describe these cases and do not establish a general model ranking.

2 Methodology and Experiments

This study evaluates the quality of LLM-generated explanations for injection-moulding defective-part prediction. The explanations are assessed by both an LLM judge and an independent manufacturing expert using three criteria: evidence faithfulness, attribution caution, and manufacturing-process relevance, as summarized in Table 3. The methodology consists of five stages: dataset loading and processing, predictive-model training and optimization, extraction of local feature-attribution evidence, generation of explanations by multiple LLMs, and evaluation using both the LLM judge and the human expert. The complete workflow is illustrated in Figure 1. The purpose of the study is not to identify the true physical cause of each defect, but to assess how accurately the generated explanations reflect the supplied attribution evidence, avoid unsupported claims, and remain relevant to manufacturing users.

2.1 Dataset

The study uses an injection-moulding dataset from the Korea AI Manufacturing Platform (KAMP) [3, 5], containing 5,232 production cycles and 35 variables describing machine, product, timing, pressure, and temperature conditions, together with the binary target *PassOrFail*. The dataset is highly imbalanced, with 5,172 passing cycles and 60 failures, corresponding to a failure rate of approximately 1.15%. The failures comprise 35 gas defects, 15 incomplete-moulding or short-shot defects, and 10 startup defects. Startup defects are removed because they occur during process stabilization and are treated as an expected operating condition. The resulting modelling dataset contains 5,222 cycles, including 50 failures. A reason-balanced train/test split is created using a fixed random seed. The test set contains five gas-defect cases, five short-shot cases, and approximately 20% of the passing cycles. This results in 4,178 training records with 40 failures and 1,044 test records with 10 failures. Production cycles are treated as independent samples, so time is not used in the split. Model tuning and threshold selection are performed using stratified four-fold cross-validation on the training set. Leakage-prone and identifier variables, including Reason, TimeStamp, row identifiers, production dates, and serial numbers, are excluded, leaving 29 predictor variables.

2.2 Predictive Model as Supporting Component

The predictive model is used to assign failure probabilities to production cycles and to provide the local model-attribution evidence used for explanation generation. Three gradient-boosting classifiers—CatBoost, LightGBM, and XGBoost—are compared. Because failures represent only a small proportion of the dataset, PR-AUC is used as the main model-selection metric, while failure recall and failure F1-score are reported as additional diagnostic measures. The hyperparameters of each classifier are optimized using Optuna and stratified four-fold cross-validation on the training set. Each model produces a probability that a production cycle will fail. A decision threshold is therefore needed to convert this probability into a final *Pass* or *Fail* prediction. Because failures are rare, the default threshold of 0.5 may miss many

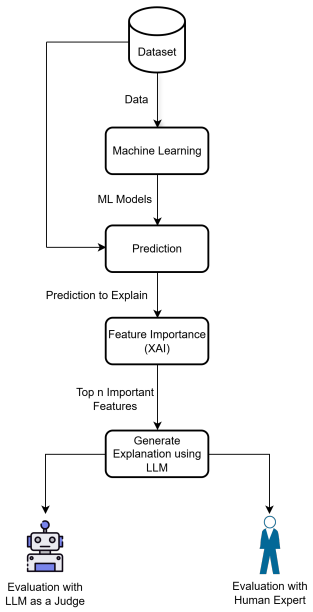

Figure 1: Overview of the study pipeline, from data preparation to LLM-based explanation and evaluation.

actual failure cases. For each model, the threshold is selected using validation predictions from the training data by maximizing the F2-score, which gives greater importance to detecting real failures while still considering false alarms. PR-AUC is used as the main model-selection metric because failures are rare in the dataset, and it measures how well the model identifies failure cases while accounting for false alarms. CatBoost achieved the highest mean cross-validation PR-AUC of 0.1565, followed by XGBoost with 0.1437 and LightGBM with 0.1083 as shown in Table 1. CatBoost was therefore selected as the downstream model according to the predefined cross-validation PR-AUC criterion. Of the ten observed failures in the test set, CatBoost correctly classified two as failures and missed the remaining eight. Only the two true-positive cases were selected for explanation generation because the study evaluates explanations of defective-part predictions made by the model. Both selected cases correspond to incomplete-moulding defects.

2.3 LLM-Based Explanation Generation

For each of the two correctly predicted failure cases in the held-out test set, the selected CatBoost model is used to calculate local SHAP feature attributions.The evidence package provided to the explanation-generating LLMs contains the most influential features, their signed attribution values, the direction in which each feature influenced the prediction, and a short manufacturing-readable description of each process variable. Features are ranked according to the absolute magnitude of their attribution, while the signed attribution is retained to indicate whether a feature pushed the prediction toward *Fail* or toward *Pass*. An example of the evidence package supplied to the explanation-generating LLM is presented in Table 2. The table shows the technical feature name, its manufacturing-readable description, its signed local attribution, and the direction in which it influenced the model prediction. A feature-description dictionary is used to translate technical variable names into terms that are easier to understand in an injection-moulding context. The attribution evidence is

Table 1: Performance comparison of the evaluated predictive models. Model selection was based on stratified cross-validation PR-AUC.

Model	CV PR-AUC	Test PR-AUC	Failure Recall	Failure F1-score	Selected
CatBoost	**0.1565**	**0.1711**	0.2000	0.1143	Yes
XGBoost	0.1437	0.0973	**0.4000**	**0.1509**	No
LightGBM	0.1083	0.0536	**0.4000**	0.0762	No

Note: Defective parts account for approximately 1.15% of the original dataset and 0.96% of the held-out test set.

treated as an explanation of the model's prediction rather than as evidence of the physical cause of the defect. Three LLMs—GPT-5.4 mini, Gemini 3.1 Flash-Lite, and Mistral Devstral 2512—are prompted to generate a short operator-facing explanation for each production cycle. Each model receives the same attribution evidence and feature descriptions to ensure a fair comparison. The prompt instructs the models to identify the most influential process variables, preserve the direction of their contributions, explain their relevance in manufacturing terms, and avoid unsupported statements about process-value levels, abnormal operating conditions, or definitive physical causes. Each of the two true-positive cases is explained by all three LLMs, resulting in six explanations for subsequent LLM-as-Judge and human-expert evaluation.

2.4 LLM-as-Judge Evaluation

The generated explanations are evaluated using a blinded LLM-as-Judge protocol. For each of the two true-positive failure cases, the judge receives the local model-attribution evidence and the three corresponding explanations. Model identities are hidden, and explanation order is randomized. Each explanation is scored from 1 to 5 for evidence faithfulness, attribution caution, and manufacturing-process relevance. The explanations are evaluated using a blinded LLM-as-Judge protocol with ~openai/gpt-mini-latest as the judge model. The overall score is calculated as the mean of these three scores. As shown in Table 4, the LLM judge assigned the GPT-5.4 mini explanations overall scores of 4.67 and 5.00, compared with 4.33 and 4.33 for Gemini 3.1 Flash-Lite and 4.33 and 3.00 for Mistral Devstral 2512. These scores describe the six evaluated explanations and should not be interpreted as estimates of the models' relative performance beyond these two cases.

2.5 Human-in-the-Loop Assessment

A central aim of this study is to compare automated LLM-as-Judge evaluation with human-in-the-loop assessment and examine their similarities and differences in judging LLM-generated manufacturing explanations. The human evaluator evaluates the six explanations generated for the two true-positive failure cases. For each case, the expert receives the same local feature-attribution evidence, manufacturing-readable feature descriptions, and corresponding LLM-generated explanations used in the automated evaluation. The expert scores each explanation from 1 to 5 according to the same three criteria used by the LLM judge: evidence faithfulness, attribution caution, and manufacturing-process relevance. While the LLM-as-Judge provides scalable and structured scores, the human-in-the-loop protocol introduces domain-informed judgment about process plausibility and practical relevance. Comparing the two evaluations allows us to investigate whether automated judging can approximate human assessment.

2.6 Evaluation Objective

The final objective is not to prove the true physical root cause of each failure, but to assess whether LLM-generated explanations are faithful to the local feature-attribution evidence and useful for human interpretation. This makes the study an evaluation-oriented contribution: it examines how generated manufacturing explanations can be assessed for grounding, caution, relevance, and actionability before being trusted in operator-facing settings.

2.7 Human Evaluation of LLM-Generated Explanations

The injection-moulding expert evaluated the explanations generated by the three LLMs, with results reported in Table 5. The expert assigned all six explanations the maximum evidence-faithfulness score, but differed substantially in attribution caution. GPT-5.4 mini and Mistral Devstral 2512 introduced unsupported claims about feature magnitudes, such as describing values as high or low, whereas Gemini 3.1 Flash-Lite avoided this error. Manufacturing relevance was reduced across all models because they stated that the available information could not provide a physical explanation of the failure. Although SHAP attributions alone are insufficient to determine the failure mechanism, the input features are measured process parameters and therefore contain physical information about the manufacturing process. The models also used the term "unshaped" instead of the standard injection-moulding term "short shot". GPT-5.4 mini was further penalized for referring to process parameters measurements as signals, while Gemini received a lower relevance score for part 116 due to imprecise process terminology. For these two cases, the expert assigned the Gemini explanations higher overall scores than the other explanations, principally because they avoided unsupported process-value claims. This case-specific assessment does not establish that Gemini generally produces better manufacturing explanations.

3 Results, Discussion, and Limitations

Across the two short-shot cases, the LLM judge assigned the highest overall scores to GPT-5.4 mini, whereas the manufacturing expert preferred Gemini 3.1 Flash-Lite. The clearest disagreement concerned attribution caution: both GPT explanations received 5 from the judge but 1 from the expert. The expert identified unsupported claims about higher injection speed, longer injection time, and higher injection pressure. Signed SHAP values describe the direction and magnitude of contributions to model output; they do not establish whether process measurements are high, low, abnormal, or physically responsible for a defect. The expert also identified unsupported magnitude claims in Mistral explanations, which the judge penalized inconsistently.

These observations illustrate how automated evaluation can overlook unsupported claims despite assigning high caution

Table 2: Example of the local model-attribution evidence supplied to the explanation-generating LLM for true-positive cases

Feature	Manufacturing-readable description	Attribution	Direction
`Plasticizing_Time`	Time required to melt and meter the material for the next shot through screw rotation	-0.7993	Toward *Pass*
`Max_Screw_RPM`	Maximum rotational speed of the screw during material preparation.	0.5960	Toward *Fail*

Interpretation: A positive attribution indicates that the feature pushed the model output toward *Fail*, whereas a negative attribution pushed it toward *Pass*. The attribution sign does not indicate whether the measured process value was high, low, or abnormal, and it does not establish physical causality.

Table 3: Criteria used for evaluating the LLM-generated explanations.

Criterion	Evaluation focus
Evidence faithfulness (EF)	Whether the explanation accurately reflects the provided model-attribution evidence.
Attribution caution (AC)	Whether the explanation avoids unsupported claims about process-value direction, abnormality, or physical causation.
Manufacturing-process relevance (MPR)	Whether the explanation is meaningful and relevant in the context of injection-moulding operations.

Table 4: LLM-as-Judge scores for the two true-positive cases.

Model	EF	AC	MPR	Overall
Case 1: Example 115				
GPT-5.4 mini	4	**5**	**5**	**4.67**
Gemini 3.1 Flash-Lite	4	**5**	4	4.33
Mistral Devstral 2512	4	4	**5**	4.33
Case 2: Example 116				
GPT-5.4 mini	**5**	**5**	**5**	**5.00**
Gemini 3.1 Flash-Lite	4	**5**	4	4.33
Mistral Devstral 2512	3	2	4	3.00

Note: EF = Evidence Faithfulness; AC = Attribution Caution; MPR = Manufacturing-Process Relevance. Bold values indicate the highest score within each true-positive example; ties are shown in bold.

Table 5: Manufacturing-expert evaluation of the LLM-generated explanations for the two true-positive cases.

Model	EF	AC	MPR	Overall
Case 1: Part 115				
GPT-5.4 mini	**5**	1	3	3.00
Gemini 3.1 Flash-Lite	**5**	**5**	**4**	**4.67**
Mistral Devstral 2512	**5**	1	**4**	3.33
Case 2: Part 116				
GPT-5.4 mini	**5**	1	3	3.00
Gemini 3.1 Flash-Lite	**5**	**5**	3	**4.33**
Mistral Devstral 2512	**5**	1	**4**	3.33

Note: EF = Evidence Faithfulness; AC = Attribution Caution; MPR = Manufacturing-Process Relevance. Scores range from 1 to 5. Bold values indicate the highest score for each criterion within a case.

scores. They do not establish a general ranking of explanation-generating models. The six explanations share only two underlying production cases of one defect type, assessed by one human expert and one LLM judge. Consequently, score differences are descriptive and cannot establish reliable model superiority or the prevalence of evaluator disagreement. Both cases were correctly detected failures, leaving explanation quality for incorrect predictions unexplored. Broader conclusions require more diverse cases, including false positives and false negatives, together with multiple experts and LLM judges.

The source code and experimental materials used in this study are publicly available at https://github.com/iman-xosseini/injection_molding_study_sikdd2026.

Acknowledgements

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) via the Research Program Knowledge Technologies (grant P2-0103) and the Gravity project AI for Science (grant GC-0001); by the European Commission Marie Sklodowska Curie program through the Horizon Europe DN APRIORI project (GA 101073551). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

During the preparation of this manuscript, the authors used Microsoft Copilot, powered by GPT-5, to support grammar refinement, reference formatting, and overall manuscript organization. All AI-generated suggestions were carefully reviewed and revised by the authors, who take full responsibility for the final content of this publication.

References

[1] A. Bilal, D. Ebert, and B. Lin. 2025. LLMs for explainable AI: a comprehensive survey. arXiv preprint arXiv:2504.00125. (2025). arXiv: 2504.00125.

[2] J. Gu et al. 2025. A survey on LLM-as-a-judge. arXiv preprint arXiv:2411.15594. (2025). arXiv: 2411.15594.

[3] KAMP AI Data Center. 2026. AI dataset detail. Retrieved May 2026 from https://www.kamp-ai.kr/aidataDetail?AI_SEARCH=&page=1&DATASET_SEQ=4&EQUIP_SEL=&GUBUN_SEL=&FILE_TYPE_SEL=&WDATE_SEL=.

[4] A. Khan, R. Nahar, H. Chen, G. E. Constante Flores, and C. Li. 2025. Faultexplainer: leveraging large language models for interpretable fault detection and diagnosis. *Computers & Chemical Engineering*, 199, 109152. doi:10.1016/j.compchemeng.2025.109152.

[5] J. W. S. Lee. 2026. Injection molding analysis. Github repository. Retrieved May 2026 from https://github.com/johnwslee/injection_molding_analysis/tree/main.

[6] A. Presciuttini, A. Cantini, S. Cramer, M. Huber, D. Wolfschläger, R. H. Schmitt, and A. Portioli-Staudacher. 2025. An integrated framework for predictive quality in injection molding: combining explainable ai and time series analysis in a german industry case study. *IFAC-PapersOnLine*, 59, 10, 1677–1682. doi:10.1016/j.ifacol.2025.09.282.

[7] F. Tayalati, I. Boukrouh, A. Azmani, and M. Azmani. 2025. Interpreting injection molding quality defect using explainable artificial intelligence and analysis of variance. *Engineering Applications of Artificial Intelligence*, 147, 110362. doi:10.1016/j.engappai.2025.110362.

[8] S. Wen, F. Li, W. Zhuang, X. Pan, W. Yu, J. Bao, and X. Li. 2025. Leveraging large language models for human–machine collaborative troubleshooting of complex industrial equipment faults. *Advanced Engineering Informatics*.