# Identifying the Cause of Performance Issues of Pretrained Language Model for Educational Technology

Pak Yuen Patrick Chan
*Department of Computer Science,*
*City University of Hong Kong,*
Kowloon, Hong Kong.
ppychan2-c@my.cityu.edu.hk

Jacky Keung
*Department of Computer Science,*
*City University of Hong Kong,*
Kowloon, Hong Kong.
jacky.keung@cityu.edu.hk

***Abstract*— Online self-learning platforms, such as question-answering (Q&A) websites, are popular technology for learning, and utilising Pretrained Language Models (PLMs) to maintain their content qualities is a common practice. However, it is challenging to identify the cause that affects the performance of PLMs. In this study, we propose using machine common sense to identify the cause affecting the performance of PLMs. We conducted an empirical experiment with three PLMs using a publicly available dataset. We select 45000 data points as the training data and 1000 as the testing data. We first train the PLMs with training data, then run machine common sense tests to examine their reasoning abilities. We define the causal relationship between content quality and reasoning ability, and use the cause to derive Metamorphic Relations (MR) of Metamorphic Testing (MT) for creating follow-up testing datasets. We analyse the changes between source and follow-up outputs to see whether the identified cause affects the performance. Results show that the reasonableness of the content is the cause that affects the performance of PLM, which has reasoning abilities. In addition, the proposed approach in this study is effective for identifying and validating the cause that affects the performance of PLMs, even on devices with limited computer resources. Future research can apply our approach and seek different machine common sense tests and counterfactual analysing techniques to identify different causes of performance issues of different PLMs.**

***Keywords*— *Content quality prediction, pre-trained language model, self-learning platform, metamorphic testing, machine common sense, causality.***

## I. Introduction

Online self-learning platforms, such as Q&A educational websites, are becoming a popular technology for learning and utilising PLMs to maintain their content quality is typical [1-6]. Thus, the performance of PLMs is crucial to maintaining the quality of these education platforms. This study aims to address the performance issue of PLMs by exploring the use of machine common sense to identify the cause that affects PLM's performance in content quality classification and use MT to validate the cause. We conduct experiments on three different PLMs to demonstrate the practicality of our approach.

We propose utilising machine common sense tests to examine the subject PLMs' reasoning capacity. Then, we analyse the causal relationship between machine common sense and content quality classification. Reasonableness of the content is identified as the cause that can affect PLMs' performance in content quality classification. We then derive MRs from this cause to create different testing datasets.

Subject PLMs then make predictions on various testing datasets to evaluate the effectiveness of MRs, in order to validate the identified cause. The results demonstrate that our approach is effective on distilled version BERT-based PLM and that improving the reasonableness of content can improve this PLM's performance in content quality classification. To the best of our knowledge, we are the first to introduce this novel approach for identifying and validating the cause that affects the performance of PLM-based educational technology.

The remainder of this paper is structured as follows: Section II reviews on the self-learning platform with PLM, causality, machine common sense, and MT. Section III presents an overview of the evaluation methodology, including the proposed MRs, experiment design, and implementation process. The results are discussed in Section IV. Finally, we conclude and discuss future research directions in Section V.

## II. Literature Review

This section briefly introduces the related items of the proposed approach, including the self-learning platform with PLM, causality, machine common sense, the preliminary knowledge of MT, and related works.

### A. Self-learning platform with Pretrained Language Model

Self-learning platforms, such as Q&A websites, play a significant role in educational technology by offering a wide range of content that caters to different levels of expertise. Popular websites like Quora and Stack Overflow have become particularly valuable for addressing software-related inquiries [3, 5, 7]. To ensure a positive user experience and encourage active participation from experts in providing valuable answers, it is essential to maintain high-quality content on these websites [5]. PLMs have been widely used to maintain the content quality on these websites, such as medical and software Q&A platforms [1-6]. While PLMs generally perform well in this context, their effectiveness does not fully reach that of human experts. Therefore, human reviews and evaluations are still essential to ensure the safe and appropriate use of PLMs [1, 3, 8].

### B. Machine Common Sense

The performance of PLM indicates that they lack sophisticated semantic comprehension and human-level common sense; instead, these PLMs rely on the statistical likelihood of words and patterns to perform sentence comprehension tasks [9, 10]. Therefore, evaluating the level of machine common sense in these PLMs is essential to determine how far they are from achieving robust human-level reasoning [9]. Additionally, further research on the common sense knowledge of PLMs is needed as those PLMs are applied to various important natural language tasks [9]. The findings from these investigations will lead to more reliable and trustworthy outcomes in diverse fields that utilize

2025 International Symposium on Educational Technology (ISET) | 979-8-3315-9550-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISET65607.2025.00044

2766-2144/25/$31.00 ©2025 IEEE
DOI 10.1109/ISET65607.2025.00044

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:26:12 UTC from IEEE Xplore. Restrictions apply.

machine learning. Ultimately, one of the most significant contributions of the science of machine common sense may be to enhance our understanding of human being [11].

### C. Causality

Causality means the cause-and-effect relationship among predicates, and it is crucial for generating realistic counterfactuals to explain decisions in machine learning [12]. Different graphical notions for generating counterfactuals can represent causality [13]. [14]'s study even showed using CAusal-Reasoning-driven Performance Testing (CARPT) to leverage causal reasoning for effectively exploring the space of operational conditions and to identify those causes that lead to performance issues of Microservices Architectures (MSA). Thus, we propose using machine common sense to generate counterfactuals for identifying the cause.

### D. Metamorphic Testing

MT has been widely used to address the oracle problems encountered in various domains [15-19]. MT tackles the test oracle issue by analysing the expected relationships between input-output pairs in consecutive executions of a system, referred to as Metamorphic Relations (MRs) [20]. If the output from different software executions violate an MR, it indicates the presence of a fault [18, 21]. [18]'s study expanded on the concept of MRs by utilizing symmetry to define Metamorphic Relation Patterns (MRPs), which can generate multiple MRs to provide deeper insights into the system under investigation. Thus, we propose using MT to validate the identified cause.

Conclusively, this paper attempts to examine PLM's machine common sense and define the causal relationship between machine common sense and content quality. Based on this causal relationship, we identified the potential cause and used it to derive MRs for validation. This research addresses the question: "Can we use machine common sense to identify the cause that affects the performance of PLMs?".

## III. Methodology

### A. Scenario

This study aims to identify the cause that influence the performance of PLMs in classifying content quality on the Stack Overflow Q&A website. Our objective is to apply our approach to identify and validate the cause that affects their performance.

### B. Proposed Counterfactual and Metamorphic Relations

Inspired by [12], we first define the causal relationship between machine common sense and content quality classification. If PLM has machine common sense, it should have reasoning capacity. Thus, making the content (C) more reasonable (R), such as with fewer errors or enough context, PLMs should give the content a higher quality rating (Q). Therefore, fewer spelling mistakes or adding more related content can improve R, which leads to reasonable PLM giving C higher Q. The causal relationship is defined as Equation 1.

$$Q = f(R, \epsilon Q), \quad \text{where } \epsilon Q \text{ is random error term and PLM is reasonable} \tag{1}$$

Based on the causal relationship, we define the counterfactual as R. We propose two MRs aimed at generating testing datasets to validate the improved R, which can lead to a more effective performance of PLM, resulting in higher quality rating (Q) for C.

The first proposed MR referred to as **MR1**, assumes that if C has fewer spelling errors, R will be increased, leading to an increase in Q (Table I).

The second proposed MR, denoted as **MR2**, assumes that if C includes additional related information, R will be increased, and Q will be increased (Table II). In this experiment, we duplicated the title text and added it to the content of the body text in the dataset to simulate the inclusion of more related context.

TABLE I. Illustration of MR1.

| (Original) Source content | (Counterfactual) MR1 content | (Reasonable PLM) Classification Result |
|---|---|---|
| **"**Can you **expian** how to **prgroam?"** | "Can you **explain** how to **program**?" | **Counterfactual** is classified as **higher quality rating** |

TABLE II. Illustration of MR2.

| (Original) Source content | (Counterfactual) MR2 content | (Reasonable PLM) Classification Result |
|---|---|---|
| Title: "Java question" Body: "How to write a for loop?" | Title: "Java question" Body: "**Java question** How to write a for loop?" | **Counterfactual** is classified as **higher quality rating** |

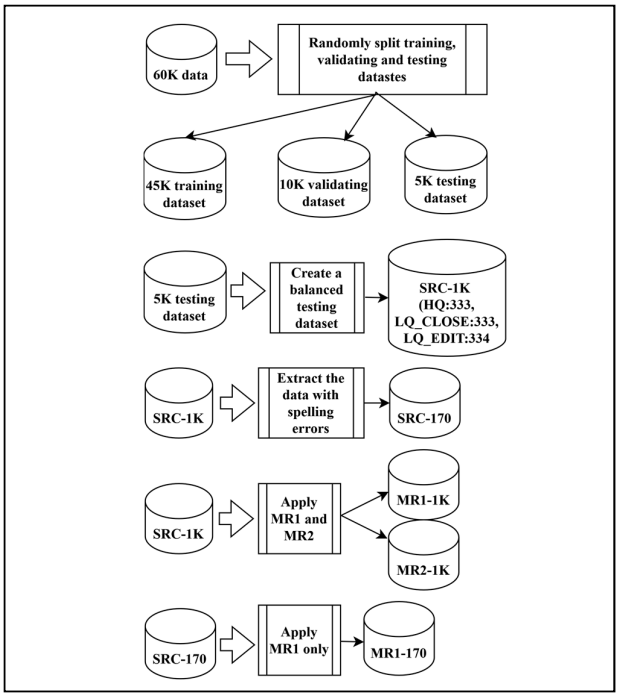



Fig. 1. Illustration of creating datasets for this study.

### C. Dataset

We selected a publicly available dataset from Kaggle[1] created to study the content quality classification of Stack Overflow questions [5]. The dataset contains sixty thousand Stack Overflow questions in English and serves the intent of

[1] The dataset is downloaded on 23/09/2023 from the link https://www.kaggle.com/datasets/imoore/60k-stack-overflow-questionswith-quality-rate



Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:26:12 UTC from IEEE Xplore. Restrictions apply.

this study. Questions are classified into three labels. They are "High-quality posts without a single edit" (HQ), "Low-quality posts with a negative score, and multiple community edits. However, they still remain open after those changes" (LQ_EDIT), and "Low-quality posts that were closed by the community without a single edit" (LQ_CLOSE) [5, p.147].

Forty-five thousand labelled Stack Overflow questions (45K) are selected randomly as the training dataset, ten thousand labelled data (10K) are selected randomly as the validating dataset, and the rest of five thousand labelled data (5K) are used to create the one thousand labelled data as this study's testing dataset (SRC-1K), which contains balanced data of one-third of each of the three labels[2], which includes 333 data labelled with HQ, 333 data labelled with LQ_CLOSE, and 334 data labelled with LQ_EDIT. We then apply MR1 and MR2 to 1Kdata to create follow-up testing datasets MR1-1K and MR2-1K, respectively. In addition, as MR1 only modified 170 lines of data, which contain incorrect spellings, we extracted all the related data and created two more testing datasets, which are "SRC-170" and "MR1-170", for further comparison use. All datasets are cleansed and set to lowercase before training and testing (Fig. 1). Additionally, the datasets of common ability tests developed by [9] are employed to determine the reasonableness of PLMs.

### D. Measurements

This experiment utilizes a common ability tests developed by [9] to assess the level of machine common sense and robustness of PLMs. We selected three token-level common sense ability tests (CSAT), and they are "Sense Making" (sm), "Winograd Schema Challenge" (wsc), and "Conjunction Acceptability" (ca). Additionally, we selected one sentence-level CSAT, the "Argument Reasoning Comprehension Task" (arct1). Four robust tests (RT) are selected: "add", "del", "sub", and "swap" [9]. All the test instances are created by replacing pronouns, verbs, adjectives, conjunction words or keywords to form dual pairs of statements with labels for models to predict. The scores are calculated by the number of correct predictions made by PLMs [9], and we assess the scores against 0.5, which is the chance of random guess, to determine the reasonableness of PLMs.

We use two metrics to measure the performance: the prediction accuracy rate (ACC) and the Matthew correlation coefficient (MCC) [22]. Let $T_p$ be true positives that are the actual positives and correctly predicted, $T_n$ be true negatives that are the actual negatives and correctly predicted, $F_p$ be false positives that are the actual negatives and wrongly predicted as positives, and $F_n$ be false negatives are the actual positives and wrongly predicted negatives. *ACC* is calculated as shown in Equation (2), and *MCC* is calculated as shown in Equation (3). Additionally, we assess the reclassification percentage (RP) and the number of predictions for each label between the source and follow-up outputs to validate the effectiveness of MRs and the identified cause.

$$ACC = (T_p + T_n)/(T_p + T_n + F_p + F_n) \quad (2)$$

$$MCC = \frac{(T_p \times T_n - F_p \times F_n)}{\sqrt{((T_p + F_p) \times (T_p + F_n) \times (T_n + F_p) \times (T_n + F_n))}} \quad (3)$$

Assume that the more reasonable the PLM is, the more accurate its classification results will be. Therefore, if the PLM achieves high score in the common ability tests, it indicates that PLM can recognize improvements from MRs and subsequently reclassify the data to a higher quality rating.

### E. Subject large language model

Due to the constraints on computational resources, three different distilled or basic versioned PLMs, which employ a minimal number of parameters, are chosen from the machine learning community "HuggingFace"[3]. **DistBERT** is a distilled version of the BERT, which uses the BERT base model as a teacher model and receives the teacher model's knowledge, and it can achieve comparable performance [23]. **DistBART** is also a distilled PLM, which uses the BART base model as a teacher model [24]. **MegaGPT2** is a small uni-directional generative PLM created by the Applied Deep Learning Research team at NVIDIA and works similarly to GPT2 [25].

### F. Setup

This experiment environment included using "Intel Core" i7 CPU with 64 GB RAM, "Windows 10" operating systems, "Jupyter Notebook" development platform and "Python" programming language and related libraries, such as "Pytorch", "transformers", "tensorflow", "scikit_learn", for machine learning algorithms. Standard parameters are applied to all subject PLMs, and the rest of the options are set to default values. In addition, Microsoft Excel and Notepad++ are used to create the follow-up testing datasets, including correcting spellings.

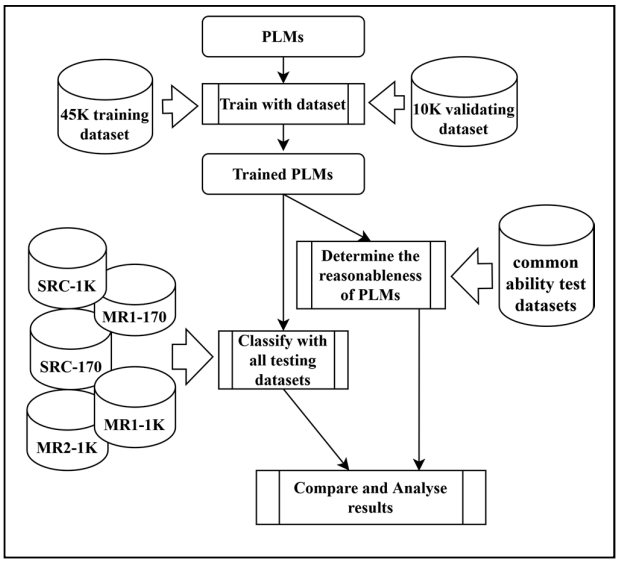



Fig. 2. Illustration of Step 1 to Step 4.

### G. Implementation

Our experiment includes five steps: **Step 1.** We train the PLMs with the training dataset as the currently operating PLMs. **Step 2.** We examine PLMs with CSAT and RT and

[2] The rest of the five thousand labelled data is not balanced. We have to randomly select around three hundred and thirty-three data with each label to create a balanced testing dataset with one thousand data to reduce the effect of imbalanced testing data.

[3] Hugging Face website is https://huggingface.co/



Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:26:12 UTC from IEEE Xplore. Restrictions apply.

collect the results to see how reasonable the PLMs are. **Step 3.** We test all PLMs with all the source and follow-up testing datasets. **Step 4**. We compare and analyse all the results from steps 2 and 3 to evaluate whether our approach can identify and validate the cause of performance issues of PLMs (Fig. 2).

## IV. Discussion

### A. Machine Common Sense

Table III shows that "sm" and "arct1" results of DistBART and MegaGPT2 are not as good as DistBERT, and DistBERT achieved results of "sm" and "arct1" higher than "RANDOM". Based on the CSAT results, DistBERT has the best reasoning abilities in the study. However, the RT results of all PLMs are below "RANDOM" (Table III). These results are similar to [9] as the subject PLMs might be confused with the modifications in RT; thus, RT results might not reflect the reasoning skills of PLMs.

TABLE III. Results of The Common Ability Tests.

| PLM | | DistBERT | DistBART | MegaGPT2 |
|---|---|---|---|---|
| CSAT | *sm* | **0.5477** | **0.5178** | 0.5029 |
| | *wsc* | 0.4912 | 0.4982 | 0.5018 |
| | *ca* | 0.4044 | 0.4918 | **0.5137** |
| | *arct1* | **0.5135** | 0.4459 | 0.4820 |
| RT | *add* | 0.1413 | 0.1522 | 0.0000 |
| | *del* | 0.1098 | 0.1463 | 0.0000 |
| | *sub* | 0.1333 | 0.1467 | 0.1333 |
| | *swap* | 0.2297 | 0.3919 | 0.0811 |
| **RANDOM** | | **0.5000** | **0.5000** | **0.5000** |

TABLE IV. Results of DistBERT.

| Dataset | 1K data | | | 170 data | |
|---|---|---|---|---|---|
| | *SRC-1K* | *MR1-1K* | *MR2-1K* | *SRC-170* | *MR1-170* |
| **ACC** | 0.6490 | 0.6460 | 0.5520 | 0.6647 | 0.6471 |
| **MCC** | 0.4751 | 0.4710 | 0.3304 | 0.4688 | 0.4625 |
| **RP** | 0.0000 | 0.0170 | 0.3190 | **0.0000** | **0.1000** |
| **HQ** | 380 | 376 | 276 | **41** | **37** |
| **LQ_CLOSE** | 333 | 349 | 311 | **52** | **68** |
| **LQ_EDIT** | 287 | 275 | 413 | **77** | **65** |

### B. MR

Table IV demonstrates that DistBERT increased the number of predictions for LQ_CLOSE while decreasing the predictions for LQ_EDIT when using MR1-1K and MR1-170. These results suggest that DistBERT recognized an improvement in content quality after spelling correction, which aligns with its reasoning capabilities shown in Table III. In contrast, the other two PLMs did not recognize any improvement in content quality after spelling correction, which reflects their weaker reasoning abilities (Tables III and IV). These findings indicate that the reasonableness of content is the cause that affects the performance of PLM with reasoning skills in classifying the content quality of stack overflow questions. Moreover, it justifies our approach of identifying the underlying cause of the PLM performance issues through machine common-sense testing and further validates this cause using MT.

### C. Further Analysis

Table V reveals that DistBART achieves the highest ACC and MCC in this study and over 86% prediction accuracy rate with SRC-1K, indicating that a small PLM can deliver comparable performance. However, DistBART is insensitive to MR1 and highly sensitive to MR2 (Table V). These results suggest that DistBART does not depend on semantic understanding; instead, it relies on statistical patterns. Since MR2 adds more tokens to the content, DistBART considers these additional contexts detrimental to content quality (Table V). This observation aligns with the findings in Table III and demonstrates that DistBART's reasoning abilities are limited, meaning that the reasonableness of content does not contribute to its performance issues.

Furthermore, Table VI indicates that MegaGPT2 is also insensitive to both MR1 and MR2 and exhibits the lowest ACC and MCC in this study. Table III shows that MegaGPT2 similarly has low reasoning skills, reinforcing that the reasonableness of content is not responsible for its performance problems. Lastly, MR2 does not help PLMs to consider the content quality is improved (Table VI). Therefore, further investigation is needed to understand why providing more related content does not positively impact PLMs in evaluating content quality.

TABLE V. Results of DistBART.

| Dataset | 1K data | | | 170 data | |
|---|---|---|---|---|---|
| | *SRC-1K* | *MR1-1K* | *MR2-1K* | *SRC-170* | *MR1-170* |
| **ACC** | 0.8640 | 0.8640 | 0.3310 | 0.9589 | 0.9589 |
| **MCC** | 0.7965 | 0.7965 | -0.0311 | 0.9290 | 0.9291 |
| **RP** | 0.0000 | 0.0020 | 0.6720 | 0.0000 | 0.0118 |
| **HQ** | 318 | 316 | 1 | 24 | 22 |
| **LQ_CLOSE** | 353 | 355 | 6 | 52 | 54 |
| **LQ_EDIT** | 329 | 329 | 993 | 94 | 94 |

TABLE VI. Results of MegaGPT2.

| Dataset | 1K data | | | 170 data | |
|---|---|---|---|---|---|
| | *SRC-1K* | *MR1-1K* | *MR2-1K* | *SRC-170* | *MR1-170* |
| **ACC** | 0.4540 | 0.4530 | 0.4540 | 0.3647 | 0.3588 |
| **MCC** | 0.1946 | 0.1928 | 0.1946 | 0.1527 | 0.1415 |
| **RP** | 0 | 0.004 | 0 | 0 | 0.0235 |
| **HQ** | 576 | 574 | 576 | 91 | 89 |
| **LQ_CLOSE** | 226 | 229 | 226 | 40 | 43 |
| **LQ_EDIT** | 198 | 197 | 198 | 39 | 38 |

## V. Conclusion

Online self-learning platforms have become a popular educational technology, and leveraging PLMs to ensure content quality is a common practice. Therefore, the performance of these PLMs is crucial for maintaining the quality of content on these educational platforms. This study demonstrates that machine common sense can be used to identify the cause that affects PLM's performance, while MT can validate the cause.



Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:26:12 UTC from IEEE Xplore. Restrictions apply.

We conducted machine common sense tests on the subject PLMs to evaluate their reasoning capacities. Then, we based on the causal relationship between machine common sense and content quality to identify the reasonableness of the content as the cause that affects their performance. We derived MRs from the identified cause to create testing datasets for validation. The results indicate that our approach is effective, particularly with the reasonable PLM: DistBERT. By improving the reasonableness of content through the correcting spelling errors, DistBERT provided better quality ratings for these contents. Thus, employing machine common sense to identify causal factors and use the cause to derive MRs for validation is an effective approach.

### *A. Future Directions*

To further advance the field, future studies could build on our approach by exploring more specific machine common sense tests and counterfactual analysing techniques. This would help pinpoint the specific skills necessary for identifying the causes that impact PLM performance more accurately. Additionally, our experiment revealed that smaller PLMs, like DistBART, can achieve high prediction accuracies, while DistBERT demonstrates its robust reasoning capabilities. Future research could explore the functionality of similar PLMs in environments with limited computational resources, such as mobile devices.

## ACKNOWLEDGMENT

This work is supported in part by the General Research Fund of the Research Grants Council of Hong Kong and the research funds of the City University of Hong Kong (6000796 6000796, 9229109, 9229098, 9220103, 9229029).

## REFERENCES

[1] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature,* vol. 620, no. 7972, pp. 172-180, 2023.

[2] R. Mousavi, T. Raghu, and K. Frey, "Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums," *Journal of Management Information Systems,* vol. 37, no. 4, pp. 1073-1098, 2020.

[3] B. Xu *et al.*, "Are we ready to embrace generative AI for software Q&A?," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2023: IEEE, pp. 1713-1717.

[4] A. Wen, M. Y. Elwazir, S. Moon, and J. Fan, "Adapting and evaluating a deep learning language model for clinical why-question answering," *JAMIA open,* vol. 3, no. 1, pp. 16-20, 2020.

[5] I. Annamoradnejad, J. Habibi, and M. Fazli, "Multi-view approach to suggest moderation actions in community question answering sites," *Information Sciences,* vol. 600, pp. 144-154, 2022.

[6] B. Sen, N. Gopal, and X. Xue, "Support-BERT: predicting quality of question-answer pairs in MSDN using deep bidirectional transformer," *arXiv preprint arXiv:2005.08294,* 2020.

[7] F. Zhang, J. Liu, Y. Wan, X. Yu, X. Liu, and J. Keung, "Diverse title generation for Stack Overflow posts with multiple-sampling-enhanced transformer," *Journal of Systems and Software,* vol. 200, p. 111672, 2023.

[8] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology,* vol. 15, no. 3, pp. 1-45, 2024.

[9] X. Zhou, Y. Zhang, L. Cui, and D. Huang, "Evaluating commonsense in pre-trained language models," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, no. 05, pp. 9733-9740.

[10] J. Browning and Y. LeCun, "Language, common sense, and the Winograd schema challenge," *Artificial Intelligence,* p. 104031, 2023.

[11] M. Kejriwal, H. Santos, A. M. Mulvehill, K. Shen, D. L. McGuinness, and H. Lieberman, "Can AI have common sense? Finding out will be key to achieving machine intelligence," *Nature,* vol. 634, no. 8033, pp. 291-294, 2024.

[12] S. Dasgupta, "Generating Causally Compliant Counterfactual Explanations using ASP," *arXiv preprint arXiv:2502.09226,* 2025.

[13] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *ACM Computing Surveys,* vol. 56, no. 12, pp. 1-42, 2024.

[14] L. Giamattei, A. Guerriero, I. Malavolta, C. Mascia, R. Pietrantuono, and S. Russo, "Identifying Performance Issues in Microservice Architectures through Causal Reasoning," in *Proceedings of the 5th ACM/IEEE International Conference on Automation of Software Test (AST 2024)*, 2024, pp. 149-153.

[15] X. Xie, Z. Zhang, T. Y. Chen, Y. Liu, P.-L. Poon, and B. Xu, "METTLE: A metamorphic testing approach to assessing and validating unsupervised machine learning systems," *IEEE Transactions on Reliability,* vol. 69, no. 4, pp. 1293-1322, 2020.

[16] Z. Ying, D. Towey, A. Bellotti, Z. Q. Zhou, and T. Y. Chen, "Preparing SQA Professionals: Metamorphic Relation Patterns, Exploration, and Testing for Big Data," *Proceedings of the International Conference on Open and Innovation Education (ICOIE 2021),* pp. 22-30, 2021.

[17] M. Zhang, J. W. Keung, T. Y. Chen, and Y. Xiao, "Validating class integration test order generation systems with Metamorphic Testing," *Information and Software Technology,* vol. 132, p. 106507, 2021.

[18] Z. Q. Zhou, L. Sun, T. Y. Chen, and D. Towey, "Metamorphic relations for enhancing system understanding and use," *IEEE Transactions on Software Engineering,* vol. 46, no. 10, pp. 1120-1154, 2020.

[19] B. Stacy, J. Hauzel, M. Lindvall, A. Porter, and M. Pop, "Metamorphic Testing in Bioinformatics Software: A Case Study on Metagenomic Assembly," in *2022 IEEE/ACM 7th International Workshop on Metamorphic Testing (MET)*, 2022: IEEE, pp. 31-33.

[20] A. Duque-Torres, D. Pfahl, C. Klammer, and S. Fischer, "Bug or not Bug? Analysing the Reasons Behind Metamorphic Relation Violations," in *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2023: IEEE, pp. 905-912.

[21] S. Segura, A. Durán, J. Troya, and A. Ruiz-Cortés, "Metamorphic relation patterns for query-based systems," in *2019 IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*, 2019: IEEE, pp. 24-31.

[22] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics,* vol. 21, no. 1, pp. 6-6, 2020, doi: 10.1186/s12864-019-6413-7.

[23] HuggingFace. "distilbert/distilbert-base-cased." Hugging Face. https://huggingface.co/distilbert/distilbert-base-cased (accessed 20/01/2025, 2025).

[24] HuggingFace. "sshleifer/distilbart-cnn-12-6." Hugging Face. https://huggingface.co/sshleifer/distilbart-cnn-12-6 (accessed 20/01/2025, 2025).

[25] HuggingFace. "robowaifudev/megatron-gpt2-345m." Hugging Face. https://huggingface.co/robowaifudev/megatron-gpt2-345m (accessed 20/01/2025, 2025).



Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:26:12 UTC from IEEE Xplore. Restrictions apply.