A Symmetric Metamorphic Relations Approach Supporting LLM for Education 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

2024 International Symposium on Educational Technology (ISET) | 979-8-3503-6141-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISET61814.2024.00017

***Abstract*—Question-Answering (Q&A) educational websites are widely used as self-learning platforms, and pre-trained large language models (LLMs) play a crucial role in maintaining content quality. Despite their usefulness, LLMs still fall short of human performance. To tackle this issue, we propose leveraging symmetric Metamorphic Relations (MRs) to enhance LLMs' performance by improving their machine common sense. The goal is to ensure that learners receive more relevant content. This work presents an empirical experiment using one specific symmetric MR, three LLMs, and a publicly available dataset of labelled Stack Overflow data. We employ the symmetric MR to generate training data that augments the machine common sense of LLMs. Additionally, we prepare a separate set of training data consisting of labelled Stack Overflow data for comparison purposes. By comparing the results of a common ability test and the predictions made by LLMs trained with different training datasets, we can assess the potential practicality of our proposed approach. Our experimental results demonstrate that a Bert-based LLM trained with MR-generated data outperforms a Bert-based LLM trained solely with regular labelled data. This outcome highlights the effectiveness of symmetric MRs in enhancing LLMs' performance by improving their machine common sense. Subsequent studies can extend our approach to other domains related to education technology and explore additional MRs to further enhance the study experience of students.**

Keywords— Content quality prediction, large language model, question-answering (Q&A) website, metamorphic relations, machine common sense, natural language processing data augmentation.

I. Introduction

Question-Answering (Q&A) educational websites have gained widespread popularity as self-learning platforms, providing users with guidelines and expert answers for effective learning [1]. Ensuring high content quality on these platforms is crucial, and multiple studies showed that large language models (LLMs) have been employed to maintain content quality across various Q&A websites [1-6]. However, the performance of LLMs has revealed limitations in deep semantic comprehension and human-level common sense, often relying on statistical likelihood and patterns for sentence understanding tasks [7, 8].

This paper aims to address these limitations by exploring the use of symmetric metamorphic relations (MRs) to enhance LLM performance in content quality classification, specifically by improving their machine common sense. We propose a specific symmetric MR and conduct experiments on three different LLMs to demonstrate the practicality of our approach. The symmetric MR generates a training dataset that incorporates common sense by replacing abbreviations with their full forms, aiming to preserve the semantic meaning of sentences and prediction results. We train the LLMs using different datasets to create distinct groups for comparison. These LLMs then make predictions on various testing datasets to evaluate their levels of machine common sense and performance. The results demonstrate that training LLMs with symmetric MR-generated data can significantly improve the machine common sense level and performance of Bert-based LLMs in the content quality classification of Stack Overflow questions compared to regular training data solely.

Additionally, the symmetric MR can be employed as a formally derived natural language processing data augmentation (NLPDA) technique to create reliable training or testing datasets. Future studies can expand on our approach by extending it to other domains related to education technology and exploring additional MRs to enhance machine common sense and further improve the performance of Bert-based LLMs.

The remainder of this paper is structured as follows: Section 2 provides background information on Q&A websites, LLMs, machine common sense, MR, and NLPDA. Section 3 presents an overview of the evaluation methodology, including the proposed MR, experiment design, and implementation process. The results are discussed in Section 4. Finally, we conclude and discuss future research directions in Section 5.

II. Background

A. Question-Answering (Q&A) websites

Q&A websites play a significant role as self-learning platforms, offering a wide range of user-generated content that spans from basic guidelines to advanced and expert-level answers. These platforms, such as Quora and Stack Overflow, have become immensely popular, particularly in the realm of software-related queries [1, 5, 9]. Maintaining high-quality content on Q&A websites is essential to ensure a positive user experience and encourage active participation from experts in providing valuable answers [1].

B. Large language model (LLM)

LLMs have been widely utilized to uphold content quality on various Q&A websites, including medical sites and Stack Overflow platforms [1-6]. While the performance of LLMs in this context is generally satisfactory, it falls short of matching human performance. As a result, human reviews and evaluations remain necessary to ensure the safe utilization of LLMs [4, 5, 10]. In addition, LLMs' performances showed that they do not possess deep semantic comprehension or human-level common sense, and only use the statistical likelihood of words and patterns to handle the sentence comprehension tasks [7, 8]. Thus, there is still a gap for LLM to reach robust human-level common sense reasoning [7].

979-8-3503-6141-4/24/$31.00 ©2024 IEEE
DOI 10.1109/ISET61814.2024.00017

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:23:35 UTC from IEEE Xplore. Restrictions apply.

C. Machine common sense and natural language processing data augmentation (NLPDA)

Improving machine common sense can be achieved by either learning from experience or the Web [11]. Learning from experience is similar to learning from data, and theoretically, more training data can improve the performance of machine learning models. Data augmentation is a widely used approach for increasing training data, even in a data scarcity situation [12-14]. However, studies found that NLPDA approaches are mostly derived from heuristics instead of formal or theoretically based approaches [12-14]. Thus, creating a training dataset for machine common sense by NLPDA is an oracle problem because it is difficult to distinguish the correct approach from the potentially incorrect approach [15, 16], and the training data might not be trustworthy if they are not formally or theoretical-based derived.

D. Metamorphic relations (MRs)

MRs of metamorphic testing (MT) have long been used for test oracle problems of different domains [17-27]. MT addresses the test oracle problem by examining the expected relationships between input-output pairs in consecutive executions of the systems [28]. These "expected" relationships are expressed as MRs, and if the output results in various software executions violate an MR, then a fault is revealed [20, 27]. [20] further utilized the concept of MRs and symmetry to define metamorphic relation patterns (MRPs) that can derive multiple MRs, whereas [27] presented a list of MRPs for identifying multiple MRs in testing query-based systems.

This paper attempts to follow [20]'s study and base the concept of symmetry on formally derived MRs as a formally derived NLPDA approach for creating machine common sense training datasets. It demonstrates the practicality of this approach in answering the research question, "Can symmetry MR enhance the performance of LLMs through enhancing their machine common sense?" with a comprehensive scaled experiment.

III. Methodology

A. Scenario

Inspired by [1], this study focuses on enhancing the performance of a currently operating LLM specifically designed for content quality classification on the Stack Overflow Q&A website. Our objective is to augment the capabilities of the LLM by training it with symmetric MR-generated data, with the aim of improving its performance through the enhancement of its machine common sense. By incorporating the principles of symmetric MRs, we seek to address the limitations of the LLM and elevate its ability to effectively classify and maintain the quality of content on the Stack Overflow platform.

B. Proposed metamorphic relation (MR)

In this study, we devised a specific symmetric MR inspired by the work of [20] and centered around the concept of symmetry, which posits that predictions should remain unaffected when the semantic meanings of sentences remain the same. By leveraging this symmetric MR, we aimed to generate a training dataset that would enhance the machine common sense of the LLM by incorporating patterns associated with common sense reasoning.

Specifically, our proposed symmetric MR, denoted as **MR1**, assumes that the semantic meaning of a sentence remains unchanged even when all abbreviations within the content are replaced with their corresponding full forms. For instance, the sentences "I don't eat" and "I do not eat" are expected to convey the same meaning to the LLMs. This metamorphic relation primarily focuses on replacing contractions containing "n't" with their expanded forms, such as substituting "n't" with "not".

C. Measurements

This experiment uses the common ability test created by [7] to evaluate the level of machine common sense and robustness of LLMs. We selected three token-level common sense ability tests (CSAT), and they are "Sense Making" (sm), "Winograd Schema Challenge" (wsc), and "Conjunction Acceptability" (ca). We also selected one sentence-level CSAT, which is "Argument Reasoning Comprehension Task" (arct1). Four robust tests (RT) are chosen, and they are "add", "del", "sub", and "swap" [7]. All the test instances are created by replacing either pronouns, verbs, adjectives, conjunction words or keywords to create dual pairs of statements with labels for models to predict. The scores are calculated by the number of correct predictions the model makes [7].

We also employ the prediction accuracy rate (ACC) and Matthew correlation coefficient (MCC) [29]. These measurements can show the prediction accuracy. 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 Equation (1), and *MCC* is calculated as Equation (2).

$$ACC = (T_p + T_n)/(T_p + T_n + F_p + F_n) \quad (1)$$

$$MCC = \frac{(T_p \times T_n - F_p \times F_n)}{\sqrt{((T_p + F_p) \times (T_p + F_n) \times (T_n + F_p) \times (T_n + F_n))}} \quad (2)$$

D. Dataset

This study uses a publicly available dataset from Kaggle[1] created to study the content quality classification of Stack Overflow questions [1]. The dataset contains 60000 Stack Overflow questions in English and serves the intent of 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).

45000 labelled Stack Overflow questions are used as the "training dataset". We then separate the rest of the 15000 labelled data into two datasets. The first 5000 labelled data

[1] 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 08:23:35 UTC from IEEE Xplore. Restrictions apply.

becomes the "extra training dataset" (5K-EX), and the second 10000 labelled data becomes this study's "testing dataset" (10Ktest). At last, we apply symmetric MR, as the NLPDA, into the "training dataset" to create a "common sense training dataset" containing 5000 labelled data (5K-CS)2 (Fig. 1). All datasets are cleansed and set to lowercase before training and testing.

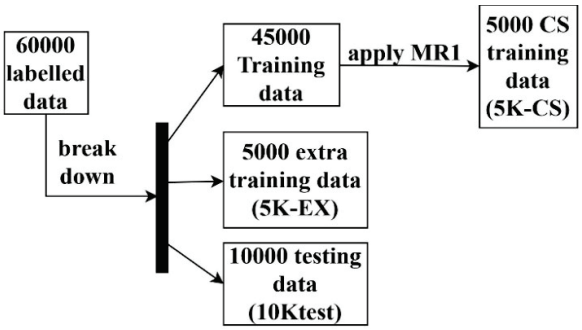


Fig. 1. Illustration of dataset creations.

E. Subject large language model

Three different LLMs are chosen from the machine learning community "HuggingFace" [3]. **BERT** is a bi-directional LLM that makes semantic understanding "from the raw text by jointly conditioning on both left and right context in all layers" [30, 31]. **T5** is an encoder-decoder transfer learning model for natural language processing that converts every language problem into a text-to-text format [32, 33]. **GPT1** is an uni-directional language model, and it is the first generative pre-training transformer-based language model created and released by the technology company "OpenAI", and it uses language modelling on a large corpus with the ability to process long-range dependencies [34, 35].

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 LLMs, and the rest of the options are set to default values.

G. Implementation

Our experiment includes five steps: Step 1 - We train the subject LLMs with a 45000 training dataset to create Group 1 (G1) as the currently operating LLMs. Step 2 - We train LLMs(G1) with 5K-EX to create Group 2 (G2). Step 3 - We train LLMs(G1) with 5K-CS to create Group 3 (G3) (Fig. 2). Step 4 - LLMs(G1), LLMs(G2) and LLMs(G3) make predictions on 10Ktest and common ability testing datasets. Step 5 - We compare and analyse all the outputs from Step 4 to evaluate whether there is any improvement (Fig. 3).

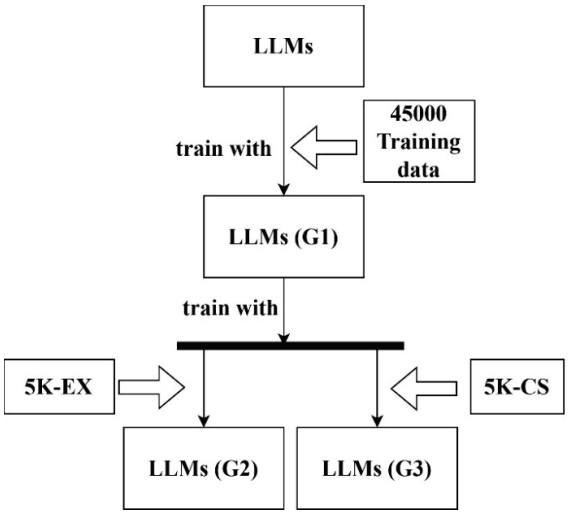


Fig. 2. Illustration of Step 1 to Step 3.

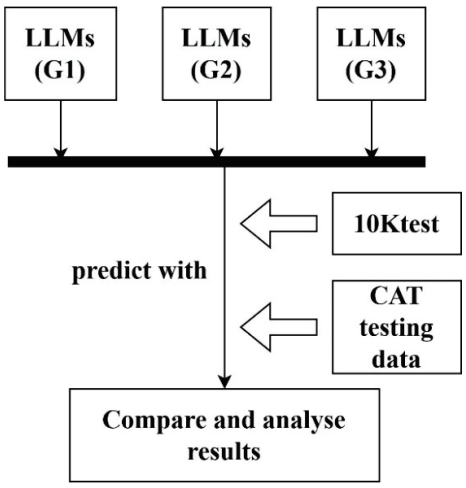


Fig. 3. Illustration of Step 4 to Step 5.

IV. Discussion

A. Machine common sense and performance

The results of our study indicate that the proposed approach is most effective when applied to Bert-based LLMs, as demonstrated in Table I[4]. Specifically, Bert(G3) exhibited superior accuracy (ACC) and Matthews correlation coefficient (MCC) compared to Bert(G2). These findings suggest that Bert-based LLMs can benefit significantly from the integration of symmetric MR-generated data, surpassing the performance achieved with regular data alone. In contrast, GPT1-based and T5-based LLMs did not exhibit any improvement in ACC and MCC when subjected to the proposed approach. Notably, GPT1(G1), GPT1(G2), and GPT1(G3) demonstrated identical results in terms of content quality classification accuracy, MCC, CSAT, and RT. Consequently, it is evident that this approach does not yield meaningful enhancements for GPT1-based and T5-based

[2] This study created 5000 common sense training data to match the size of the extra training dataset, in order to reduce the influence of size differences in training datasets. In addition, this study did not replace "ain't", "men't", "differn't", "itn't", and "Cullen't" because it is hard to identify the true meaning of those words, and they only related to seven changes out of 45000 data.

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

[4] Due to page limitation, we omit showing the results of GPT1 in Table I because they are all identical.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:23:35 UTC from IEEE Xplore. Restrictions apply.

LLMs. Therefore, based on the study's findings, it is clear that the proposed approach is most suitable for enhancing Bert-based LLMs, while it does not provide significant improvements for GPT1-based and T5-based LLMs.

Table I also shows that "sm", "wsc", and "arct1" results are positively related to ACC and MCC, especially "sm" but not "ca". This can be explained by the fact that LLMs are focusing on the content quality classification of stack overflow questions, so LLMs are required to understand the content of the stack overflow question to make a judgement. "sm", "wsc" and "arct1" are tests that change the pronouns or nouns of sentences, whereas "ca" changes the conjunction words, which might not be significant enough for content quality classification of stack overflow questions. Thus, enhancing machine common sense that related to "sm", "wsc", and "arct1" tests can improve LLMs' performance in classifying the content quality of stack overflow questions.

However, the RT results are not positively related to ACC and MCC. Table I shows that despite T5(G3) having better RT results than the other two T5-based LLMs, it has worse ACC and MCC. Similar to Bert(G3) having worse RT results than the other two Bert-based LLMs, it has better prediction results. These RT results are similar to [7] as the subject LLMs might be confused with the modifications in RT; thus, enhancing the robustness related to RT cannot improve the performance in classifying the content quality of stack overflow questions.

TABLE I. RESULTS OF THE EXPERIMENT.

		Bert (G1)	Bert (G2)	Bert (G3)	T5 (G1)	T5 (G2)	T5 (G3)
	ACC	0.8696	0.8525	**0.8580**	0.8672	0.8649	0.8508
	MCC	0.8040	0.7873	**0.7920**	0.8008	0.7997	0.7851
CSAT	**sm**	**0.5242**	0.5168	0.5136	**0.4885**	0.4832	0.4832
	wsc	0.4947	0.5053	**0.5088**	**0.4982**	0.4947	0.4947
	ca	0.5628	**0.5738**	0.5464	0.5683	**0.5738**	0.5574
	arct1	0.4932	0.4910	**0.4955**	**0.4707**	0.4662	0.4685
RT	**add**	**0.1630**	0.1413	0.1304	0.2826	**0.2935**	**0.2935**
	del	**0.2561**	0.2439	0.2195	0.3171	0.3049	**0.3902**
	sub	0.1733	**0.2267**	0.2000	0.2000	0.1733	**0.1867**
	swap	0.3514	**0.4054**	0.3514	**0.4459**	**0.4459**	0.4054

B. Further Analysis

The analysis of the results suggests that Bert-based LLMs can effectively utilize symmetric MR-generated training datasets for enhancing their machine common sense and performance. This approach becomes particularly valuable when acquiring additional training data is challenging. Furthermore, employing symmetric MR as a formally derived NLPDA approach, which is based on the concept of symmetry, allows for the creation of trustworthy training and testing datasets.

While BERT(G1) demonstrates the best ACC and MCC results, it aligns with previous research [7] indicating that simply increasing the amount of data may not necessarily lead to improved performance. It is important to note that this study focused on employing a single symmetric MR to enhance machine common sense, and the MR-generated training data proved more effective in improving the performance of Bert-based LLMs compared to regular training data. These findings provide guidance for future research in the creation of metamorphic relations (MRs) to enhance the machine common sense and performance of Bert-based LLMs. Ultimately, a standardized set of machine common sense MRs can be developed to incorporate various machine common sense patterns into different LLMs to bridge the gap between machines and humans regarding common sense reasoning.

V. CONCLUSION

Q&A websites serve as popular self-learning platforms, and the performance of LLMs is crucial in maintaining content quality on these online education platforms. This study has demonstrated that utilizing symmetric MRs to generate training data can significantly enhance the performance of Bert-based LLMs in the content quality classification of Stack Overflow questions and improve their machine common sense, surpassing the performance achieved with regular training data alone. Therefore, symmetric MRs have proven to be effective in enhancing LLMs' performance by improving their machine common sense.

Moreover, the application of symmetric MRs as a formally derived NLPDA approach can provide trustworthy training or testing datasets based on the concept of symmetry. This finding opens up possibilities for utilizing symmetric MRs in other education technology domains to improve machine common sense and enhance the performance of Bert-based LLMs, supporting contents needed for students.

To further advance the field, future studies can expand upon our approach by exploring different MRs and extending it to various domains. By continuing to enhance machine common sense and LLM performance, we can foster improved automated content quality and provide enhanced learning experiences for users. This work highlights the potential of symmetric MRs in enhancing LLMs and demonstrates their effectiveness in improving machine common sense in technology education. By leveraging these techniques, we can pave the way for advancements in the performance and reliability of LLMs, benefiting students in various educational and self-learning contexts, and develop a standardized set of machine common sense MRs to bridge the gap between machines and humans regarding common sense reasoning.

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).

REFERENCES

[1] 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.

[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] 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.

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

[5] 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.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:23:35 UTC from IEEE Xplore. Restrictions apply.

[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] 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.
[8] J. Browning and Y. LeCun, "Language, common sense, and the Winograd schema challenge," *Artificial Intelligence,* p. 104031, 2023.
[9] 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.
[10] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology,* 2023.
[11] D. Gunning, "Machine common sense concept paper," *arXiv preprint arXiv:1810.07528,* 2018.
[12] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *Ai Open,* vol. 3, pp. 71-90, 2022.
[13] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Applied Soft Computing,* vol. 132, p. 109803, 2023.
[14] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in nlp," *Transactions of the Association for Computational Linguistics,* vol. 11, pp. 191-211, 2023.
[15] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, "The oracle problem in software testing: A survey," *IEEE transactions on software engineering,* vol. 41, no. 5, pp. 507-525, 2014.
[16] A. R. Ibrahimzada, Y. Varli, D. Tekinoglu, and R. Jabbarvand, "Perfect is the enemy of test oracle," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 70-81.
[17] 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.
[18] 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.
[19] 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.
[20] 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, 2018.
[21] 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.
[22] J. D. Ellis, R. Iqbal, and K. Yoshimatsu, "Verification of the neural network training process for spectrum-based chemical substructure prediction using metamorphic testing," *Journal of Computational Science,* vol. 55, p. 101456, 2021.
[23] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, and P. Tonella, "Testing machine learning based systems: a systematic mapping," *Empirical Software Engineering,* vol. 25, no. 6, pp. 5193-5254, 2020.
[24] P. Saha and U. Kanewala, "Fault Detection Effectiveness of Metamorphic Relations Developed for Testing Supervised Classifiers," in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 4-9 April 2019 2019, pp. 157-164, doi: 10.1109/AITest.2019.00019.
[25] X. Xie, J. W. K. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," *J SYST SOFTWARE,* vol. 84, no. 4, pp. 544-558, 2011, doi: 10.1016/j.jss.2010.11.920.
[26] S. Pugh, M. S. Raunak, D. R. Kuhn, and R. Kacker, "Systematic testing of post-quantum cryptographic implementations using metamorphic testing," in *2019 IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*, 2019: IEEE, pp. 2-8.
[27] 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.
[28] 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.
[29] 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.
[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.
[31] HuggingFace. "BERT base model (uncased)." Hugging Face. https://huggingface.co/google-bert/bert-base-uncased (accessed 2024).
[32] HuggingFace. "Google's T5 Version 1.1." Hugging Face. https://huggingface.co/google/t5-v1_1-base (accessed 2024).
[33] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research,* vol. 21, no. 1, pp. 5485-5551, 2020.
[34] HuggingFace. "OpenAI GPT 1." Hugging Face. https://huggingface.co/openai-community/openai-gpt (accessed 2024).
[35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on October 06,2026 at 08:23:35 UTC from IEEE Xplore. Restrictions apply.