Validating pretrained language models for content quality classification with semantic-preserving metamorphic relations

Chan, Pak Yuen Patrick; Keung, Jacky

Published in:
Natural Language Processing Journal

Published: 01/12/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1016/j.nlp.2024.100114

Publication details:
Chan, P. Y. P., & Keung, J. (2024). Validating pretrained language models for content quality classification with semantic-preserving metamorphic relations. *Natural Language Processing Journal*, *9*, Article 100114. https://doi.org/10.1016/j.nlp.2024.100114

Natural Language Processing Journal 9 (2024) 100114

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Validating pretrained language models for content quality classification with semantic-preserving metamorphic relations

Pak Yuen Patrick Chan *, Jacky Keung

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Content quality prediction
Pretrained language model
Validation
Metamorphic testing
Simulation testing

ABSTRACT

Context: Utilizing pretrained language models (PLMs) has become common practice in maintaining the content quality of question-answering (Q&A) websites. However, evaluating the effectiveness of PLMs poses a challenge as they tend to provide local optima rather than global optima.
Objective: In this study, we propose using semantic-preserving Metamorphic Relations (MRs) derived from Metamorphic Testing (MT) to address this challenge and validate PLMs.
Methods: To validate four selected PLMs, we conducted an empirical experiment using a publicly available dataset comprising 60000 data points. We defined three groups of Metamorphic Relations (MRGs), consisting of thirteen semantic-preserving MRs, which were then employed to generate "Follow-up" testing datasets based on the original "Source" testing datasets. The PLMs were trained using a separate training dataset. A comparison was made between the predictions of the four trained PLMs for "Source" and "Follow-up" testing datasets in order to identify instances of violations, which corresponded to inconsistent predictions between the two datasets. If no violation was found, it indicated that the PLM was insensitive to the associate MR; thereby, the MR can be used for validation. In cases where no violation occurred across the entire MRG, non-violation regions were identified and supported simulation metamorphic testing.
Results: The results of this study demonstrated that the proposed MRs could effectively serve as a validation tool for content quality classification on Stack Overflow Q&A using PLMs. One PLM did not violate the "Uppercase conversion" MRG and the "Duplication" MRG. Furthermore, the absence of violations in the MRGs allowed for the identification of non-violation regions, confirming the ability of the proposed MRs to support simulation metamorphic testing.
Conclusion: The experimental findings indicate that the proposed MRs can validate PLMs effectively and support simulation metamorphic testing for PLMs. However, further investigations are required to enhance the semantic comprehension and common sense knowledge of PLMs and explore highly informative statistical patterns of PLMs, in order to improve their overall performance.

1. Introduction

Question-answering (Q&A) websites, which serve as a dynamic self-learning platform, provide a source of learning or solution searching for different users (Annamoradnejad et al., 2022). Stack Overflow and Quora are among the largest and most popular self-learning platforms for different levels of computer and software practitioners (Annamoradnejad et al., 2022; Xu et al., 2023; Zhang et al., 2023). Given their significance, the content quality of the Q&A websites must be maintained to remain popular and provide quality support to their users. (Annamoradnejad et al., 2022).

Machine learning is a cutting-edge technology that has been applied to maintain the quality of Q&A websites by applying pretrained language models (PLMs) to either replying the users' questions or classifying the content quality of the sites automatically (Annamoradnejad et al., 2022; Singhal et al., 2023; Xu et al., 2023). Although PLMs are capable of performing the replying and classification tasks, a robust evaluation approach is required to ensure the correctness of PLMs (Chang et al., 2024).

Evaluating the performance of machine learning models, including PLMs, is a test oracle problem since machine learning models can deliver local optima rather than global optima. Metamorphic Testing (MT) was selected for this challenge as it is a widely used technique to alleviate the test oracle problem (Sun et al., 2022; Xiao et al., 2022; Xie et al., 2020; Ying et al., 2021; Zhang et al., 2021; Zhou

* Corresponding author.
E-mail address: ppychan2-c@my.cityu.edu.hk (P.Y.P. Chan).

https://doi.org/10.1016/j.nlp.2024.100114
Received 3 March 2024; Received in revised form 13 September 2024; Accepted 12 October 2024

et al., 2020a). MT addresses the test oracle problem by examining the expected relationships between input–output pairs in consecutive executions of the systems (Duque-Torres et al., 2023). These expected relationships are expressed as metamorphic relations (MRs), and if the output results in various software executions violate an MR, a fault is revealed (Segura et al., 2019; Zhou et al., 2020a). We can avoid the test oracle problem and evaluate the models through this approach.

In this study, we adopted the MT approach to validate PLMs for content quality classification of Q&A websites. We derived three groups of semantic-preserving MRs inspired by Zhou et al. (2020a), and these MR groups (MRGs) assume that predictions should not be affected as the semantic meanings are preserved. We experimented with four different PLMs on labelled datasets containing 60,000 Stack Overflow questions with quality ratings. We applied MRs to generate thirteen follow-up testing datasets for testing. The source and follow-up predictions of PLMs were compared, and if there was no violation, meaning there was no inconsistent prediction between the two datasets, the PLM was insensitive to the related MR and that MR could be used for validation. If the PLM was insensitive to data transformation, then any violation to MR could be interpreted as a fault in the model. Furthermore, if there was no violation with the entire MRG, non-violation regions were identified for supporting simulation metamorphic testing.

The results show that the proposed MRs can effectively validate PLMs on classifying the content quality of Stack Overflow Q&A and can support simulation metamorphic testing by identifying the non-violation regions. The main contributions of this study include:

(1) We validate four different PLMs with our proposed MRs and analyse the range of MRs to create non-violation regions in support of simulation metamorphic testing. To our best knowledge, this study is the first study to introduce this approach to validate PLMs for classifying the content quality of Stack Overflow Q&A.

(2) The semantic-preserving MRs are proposed to validate PLMs effectively. One PLM did not violate two types of MRs. These results demonstrate that the proposed MRs can be extended to validate other domains using different languages and PLMs.

(3) We created a systematic validation approach using semantic-preserving MRs for researchers and practitioners to follow and apply.

(4) We outlined several research challenges and opportunities that researchers may encounter in extending the proposed approach for PLMs in other areas.

This paper is structured in the following manner. Section 2 introduces Question-answering (Q&A) websites using the machine learning approach. It also discusses the challenges in test oracle problem, MT, and the recent related works. Section 3 describes the methodology of this study to determine the effectiveness of the proposed MRs in validating PLMs. Section 4 discusses and analyses the results. Section 5 covers the threats to the validity of this study. Finally, Section 6 concludes this paper with the limitations of this study and future directions.

2. Literature review

This section briefly introduces question-answering (Q&A) websites using the machine learning approach, test oracle problem and the preliminary knowledge of MT, and related works.

2.1. Question-answering (Q&A) websites

Question-answering (Q&A) is a crucial technology in the field of human–computer interaction (Chang et al., 2024), and Q&A websites provide a self-learning platform for many users. Quora and Stack Overflow are among the largest and most popular Q&A self-learning platforms for software-related issues (Annamoradnejad et al., 2022; Xu et al., 2023; Zhang et al., 2023). These platforms contain different user-provided content ranging from simple guidelines to highly sophisticated and experienced answers, serving as valuable sources of learning and solutions for users at different levels (Annamoradnejad et al., 2022).

Thus, their content quality must be maintained to protect the sites' reputation, attract more visitors, provide a better user experience, encourage experts answering more questions, and increase search result rankings of the sites (Annamoradnejad et al., 2022). Maintaining the content quality is accomplished by keeping questions clean and clear, removing unpleasant content and duplicated questions, and safeguarding high-quality content from unwanted changes (Annamoradnejad et al., 2022).

2.2. Using machine learning approach

Machine learning models, such as PLMs, have been applied to maintain the content quality of Q&A websites and respond to questions across various Q&A platforms (Gao et al., 2023; Mousavi et al., 2020; Singhal et al., 2023; Wen et al., 2020; Xu et al., 2023). Machine learning models have also been applied to generate hints for the question, aiding users in getting the final answers and clarifying the questions (Zhang and Yang, 2024). Furthermore, they have been used to construct a multimodal medical Q&A system with medical text and image data (Zhi, 2024).

PLMs have been utilized on Medical Q&A sites for answering medical questions; although the performance of PLMs is reasonably acceptable, human reviews and evaluation frameworks are required for creating safe and helpful clinically used PLMs (Singhal et al., 2023). Generative Pretrained Transformers (GPT), which is a PLM developed by the company OpenAI, has been used to reply to Stack Overflow questions; however, human responses to Stack Overflow questions are still better than PLMs (Xu et al., 2023).

Despite human responses are better at answering questions, it is challenging to maintain the content quality of Q&A from community support websites, such as Stack Overflow, and using PLMs has demonstrated the ability to automate moderation and content quality prediction (Annamoradnejad et al., 2022; Sen et al., 2020). PLMs were also applied to evaluate and examine the content quality of health-related Q&A sites, and the results were acceptable (Mousavi et al., 2020; Wen et al., 2020). Additionally, PLMs have been employed to identify duplicate questions on Game Development Stack Exchange and Stack Overflow Q&A websites to enhance content quality (Kamienski et al., 2023).

Overall, PLMs exhibit good performance on Q&A tasks and have the potential for further enhancing their proficiency in social, event, and temporal common sense knowledge in the future; however, there is a need for an evaluation method to ensure the accurate assessment of PLMs' capabilities and limitations (Chang et al., 2024).

2.3. Test oracle problem and metamorphic testing

Performance evaluation of machine learning models, such as PLMs, is a test oracle problem since machine learning models deliver local optima rather than global optima (Xie et al., 2020; Ying et al., 2021; Zhang et al., 2021; Zhou et al., 2020a). While human evaluation of the responses of PLMs is feasible (Singhal et al., 2023), it becomes more challenging when dealing with a large volume of the responses of PLMs. To address this issue, a systematic and practical approach for validating PLMs is needed, and Metamorphic testing (MT) can be employed to mitigate the test oracle problem (Xie et al., 2020; Ying et al., 2021; Zhang et al., 2021; Zhou et al., 2020a). MT addresses the test oracle problem by examining the expected relationships between input–output pairs in consecutive executions of the systems (Chen et al., 2020; Duque-Torres et al., 2023). These "expected" relationships are expressed as metamorphic relations (MRs), and if the output results in various software executions violate an MR, then a fault is revealed (Segura et al., 2019; Zhou et al., 2020a).

For example, a mathematical property of the $sine$ function is that $sin(y) = sin(\pi - y)$. The corresponding MR is that if two function inputs, $y1$ and $y2$, satisfy $y1 + y2 = \pi$, then two function outputs should be

Table 1
Table of the proposed semantic-preserving MRs.

MR group (MRG)	Descriptions
MRG1: Duplication	MR1.1 — duplicate the content of the title two times MR1.2 — duplicate the content of the title three times MR1.3 — duplicate the content of the title four times MR1.4 — duplicate the content of the title five times MR1.5 — duplicate the content of the title six times
MRG2: Abbreviation replacement[a]	MR2.1 — Replace [don't, doesn't, didn't] with [do not, does not, did not] MR2.2 — Replace [don't, doesn't, didn't, can't, couldn't] with [do not, does not, did not, cannot, could not] MR2.3 — Replace [don't, doesn't, didn't, can't, couldn't, won't, wouldn't] with [do not, does not, did not, cannot, could not, will not, would not] MR2.4 — Replace [don't, doesn't, didn't, can't, couldn't, won't, wouldn't, shalln't, shouldn't, isn't, wasn't, aren't, weren't] with [do not, does not, did not, cannot, could not, will not, would not, shall not, should not, is not, was not, are not, were not]
MRG3: Uppercase conversion	MR3.1 — Convert 25% texts to uppercase MR3.2 — Convert 50% texts to uppercase MR3.3 — Convert 75% texts to uppercase MR3.4 — Convert 100% texts to uppercase

[a] This experiment does not check the spelling errors, such as "cant", "dont".

equal, i.e., *sin(y1) = sin(y2)*. Based on MR, *y2* is constructed based on *y1* and *sin(y1)*. Therefore, we refer to *y1* as a source input and *y2* as a follow-up input. Such relations can be used to test the program as if the outputs of *sin(y1)* and sin *(y2)* are different. The implementation for this function *sin()* is faulty as it violates the above MR (Zhang et al., 2021).

MT has been applied to alleviate test oracle problems in many areas (Segura et al., 2018), for example, bioinformatic software (Stacy et al., 2022), machine learning classification (Ellis et al., 2021; Riccio et al., 2020; Saha and Kanewala, 2019; Xie et al., 2011), cryptographic software (Pugh et al., 2019), online search engines (Segura et al., 2019), and so on. Zhou et al. (2020a) further the concept of metamorphic relations and symmetry to defined metamorphic relation patterns (MRPs) that can derive multiple MRs. Segura et al. (2019) presented a list of MRPs for identifying multiple MRs in testing query-based systems. Furthermore, Ying et al. (2022) have incorporated the simulation testing concept into MT and introduced the concept of MR-violation regions to enable testers or researchers to determine whether the test case is violated or not by the location of the test case.

Therefore, this paper aims to demonstrate MT's practicality in validating PLMs and supporting simulation testing through a comprehensive scaled experiment.

3. Methodology

The main objective of this experiment is to attempt to answer two research questions.

RQ1 Can MRs validate the content quality classification of Stack Overflow problems using Pretrained Language Models?

RQ2: Can MRs support simulation metamorphic testing for Pretrained Language Models?

In this section, we presented the proposed semantic-preserving MRs group and evaluated the effectiveness through the experiment's design and implementation.

3.1. Metamorphic relations (MRs)

In this study, we adopted the MT approach to validate PLMs and derived three semantic-preserving MR groups (MRGs) inspired by Zhou et al. (2020a). These relation groups revolve around the principle of symmetry, positing that the quality predictions for content – specifically, for "Source" and "Follow-up" questions on Stack Overflow – should remain consistent, as the semantic content does not change (Table 1). Thus, systems should classify the content quality of "Source" Stack Overflow questions and the "Follow-up" Stack Overflow questions created by MRGs with the same quality rating. MRGs should be held. Otherwise, there is a violation with MRGs and a fault in the PLM.

MRG1 — Duplication

"Source" and "Follow-up" prediction results should be consistent regardless of duplicating the content of title texts. Due to the limited computer resources, this experiment only duplicates the content of title texts (not title and body text). As the contents of title texts are duplicated but not modified, there is no change in semantic meaning; thus, they should give the same outputs.

For example, we assume two sentences, "I don't know" and "I don't know I don't know", are semantic equivalent to the models, and models will classify both sentences with the same label.

We create five associated MRs duplicating the content of the title texts at different times, and they are MR1.1, MR1.2, MR1.3, MR1.4, and MR1.5 (Table 1).

MRG2 — Abbreviation replacement

"Source" and "Follow-up" prediction results should be consistent regardless of whether terms are presented in their abbreviated form or full form. This consistency is expected as the transformation involves expanding abbreviations in the title and body texts to their full forms without altering the overall semantic meaning. Thus, they should give the same outputs.

For example, we assume two sentences, "I don't know" and "I do not know", are semantic equivalent to the models, and models will classify both sentences with the same label.

We create four associated MRs that replace different numbers of abbreviations into full forms, and they are MR2.1, MR2.2, MR2.3, and MR2.4 (Table 1).

MRG3 — Uppercase conversion

"Source" and "Follow-up" prediction results should be consistent regardless of whether uppercase or lowercase is used. As only the case format of the title and body texts are changed, there is no change in semantic meaning; thus, they should give the same outputs.

For example, we assume two sentences, "I don't know" and "I DON'T KNOW", are semantic equivalent to the models, and models will classify both sentences with the same label.

We create four associated MRs that convert texts from lowercase into uppercase at different percentages of texts in the title and the body texts, and they are MR3.1, MR3.2, MR3.3, and MR3.4 (Table 1).

3.2. Test scenario

Inspired by Annamoradnejad et al. (2022), this study used the self-learning Q&A website "Stack Overflow" as the test scenario. It is crucial to maintain the content quality of the Q&A site, and using PLMs to classify content quality is a time-saving approach for maintaining and improving the quality of the site (Annamoradnejad et al., 2022). Four PLMs are chosen in this study as the classifiers of the Stack Overflow questions.

3.3. Subject pretrained language model

This experiment involved four PLMs provided by the machine learning community "Hugging Face".[1]

XLNet was firstly introduced by Yang et al. (2019) and is pretrained on a large-scale dataset, including Wikipedia, BookCorpus, Giga5, ClueWeb, and Common Crawl. It is an unsupervised generalized pretraining learning method based on a permutation-based language modelling objective and employs the Transformer-XL model as the core model to preserve the advantages of the autoregressive language model and bidirectional model (Annamoradnejad et al., 2022; HuggingFace, 2024d; Yang et al., 2019; Zhou et al., 2020b). XLNet has been applied in language tasks that use sentences to make decisions, such as text generation, text classification and question answering, and this study uses the cased version (Annamoradnejad et al., 2022; HuggingFace, 2024d).

XLM-Roberta. Roberta is a transformer model similar to BERT, which operates as a bidirectional PLM that processes from "unlabelled text by jointly conditioning on both left and right context in all layers" (Devlin et al., 2018). XLM-Roberta is enhanced with dynamic masking, sentences without next sentence prediction loss, a larger batch size, and a more extensive vocabulary size, which gives it access to more knowledge than other PLMs. As for XLM-Roberta, it is a multilingual cased version of Roberta and is pretrained with filtered CommonCrawl data which is at the size of 2.5 terabytes and contains over 100 languages (Conneau et al., 2019; HuggingFace, 2024c; Zhou et al., 2020a).

T5, which is developed by Google and pretrained on the Colossal Clean Crawled Corpus, is an encoder–decoder transfer learning model for natural language processing that converts every language problem into a unified text-to-text format (HuggingFace, 2024a; Raffel et al., 2020; Shazeer, 2020). The model has demonstrated proficiency across various language tasks, including question-answering and text classification (Raffel et al., 2020). This study utilizes version 1.1 of T5.

GPT1 is an uni-directional language model and the pioneering generative pretraining 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 (HuggingFace, 2024b; Radford et al., 2018; Zhou et al., 2020a). GPT1 is pretrained with the BooksCorpus dataset, containing over 7000 unique unpublished books and has proven to solve discriminative tasks such as question answering, semantic similarity assessment, entailment determination, and text classification (Radford et al., 2018).

3.4. Dataset

This study utilizes a publicly available dataset from Kaggle,[2] which was created to study the content quality classification of Stack Overflow questions (Annamoradnejad et al., 2022). The dataset contains 60,000 Stack Overflow questions in English and serves the intent of this study. Questions are classified into three labels: "High quality questions that require no moderation action" (HQ), "Low-quality questions that require community edits" (LQ_EDIT), and "Low-quality questions that were closed by the community without a single edit" (LQ_CLOSE) ((Annamoradnejad et al., 2022), p. 147).

For this study, the training dataset includes 45,000 Stack Overflow questions with labels for training PLMs. The testing dataset contains 15,000 Stack Overflow questions with labels and is the "Source" testing dataset. We applied MRGs with "Source" testing datasets to create thirteen "Follow-up" testing datasets for comparison. In addition, datasets are cleansed and all set to lowercase before training and testing.

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

[2] The dataset is downloaded on 23/09/2023 from https://www.kaggle.com/datasets/imoore/60k-stack-overflow-questions-with-quality-rate.

3.5. Environment

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 "transformers", "tensorflow", "scikit_learn", for machine learning algorithms. Standard parameters are applied to all PLMs, and the rest of the options are set to default values.

3.6. Measurement

This experiment measures the re-labelling rate (RR), inspired by Xie et al. (2020), to measure the differences between the source and follow-up outputs. Source outputs were the prediction outcomes of PLMs on the "Source" testing dataset, whereas follow-up outputs were the prediction outcomes of PLMs on the "Follow-up" testing datasets. Suppose there is no violation, which is prediction inconsistency, between source outputs and follow-up outputs. In that case, RR is zero, and the PLM is insensitive to the associated MR, which means the MR could be used for validation.

Let "Source" testing dataset be $Ts = [x1, x2 \ldots xn]$ and "Follow-up" testing dataset be $Tf = [xf1, xf2 \ldots xfn]$. PLM was trained by the training dataset. The prediction outcomes of PLM with Ts be $Ys = [ys1, ys2 \ldots ysn]$, and the prediction outcomes of PLM with Tf be $Yf = [yf1, yf2 \ldots yfn]$. Let dif be the number of cases with label differences between Ys and Yf. RR is calculated as Eq. (1).

$$\mathrm{RR} = \text{ dif } / \text{total size of testing dataset} \tag{1}$$

The above calculation indicates that no difference between Ys and Yf if RR is zero, which is considered a zero violation of this MR. If RR is larger than zero, then violations have occurred. Concerning validation, if RR is zero, no violation is caused by MR, and vice versa.

As the subject datasets are labelled, we also employ the prediction accuracy rate (ACC) and Matthew correlation coefficient (MCC) (Chicco and Jurman, 2020). These measurements can show any relationship between MRs and prediction accuracy. Let Tp be true positives that are the actual positives and correctly predicted, Tn be true negatives that are the actual negatives and correctly predicted, Fp be false positives that are the actual negatives and wrongly predicted as positives, and Fn be false negatives are the actual positives and wrongly predicted negatives. ACC is calculated as Eq. (2), and MCC is calculated as Eq. (3)

$$ACC = (T_p + T_n)/(T_p + T_n + F_p + F_n) \tag{2}$$

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

3.7. Experiment design

Our experiment includes six steps as follows:

Step 1. We apply the proposed MRs to the "Source" testing dataset and generated thirteen "Follow-up" testing datasets. **Step 2.** We train the subject PLMs with a training dataset containing 45,000 Stack Overflow questions with labels. **Step 3.** PLMs make predictions on the "Source" testing dataset containing 15,000 Stack Overflow questions with labels to create source outputs. **Step 4.** PLMs make predictions on thirteen "Follow-up" testing datasets to create thirteen sets of follow-up outputs. **Step 5.** We compare source and follow-up outputs and analyse their differences in measurement. A failure is revealed if the relation is violated, which means differences occur between the outputs (Fig. 1). **Step 6.** If there is zero violation with the whole group of MRs, a non-violation region is identified for supporting simulation metamorphic testing.

Four trained PLMs were tested with one source testing dataset and thirteen follow-up testing datasets. Thus, this study had four sets of source outputs and fifty-two sets of follow-up outputs for comparison.

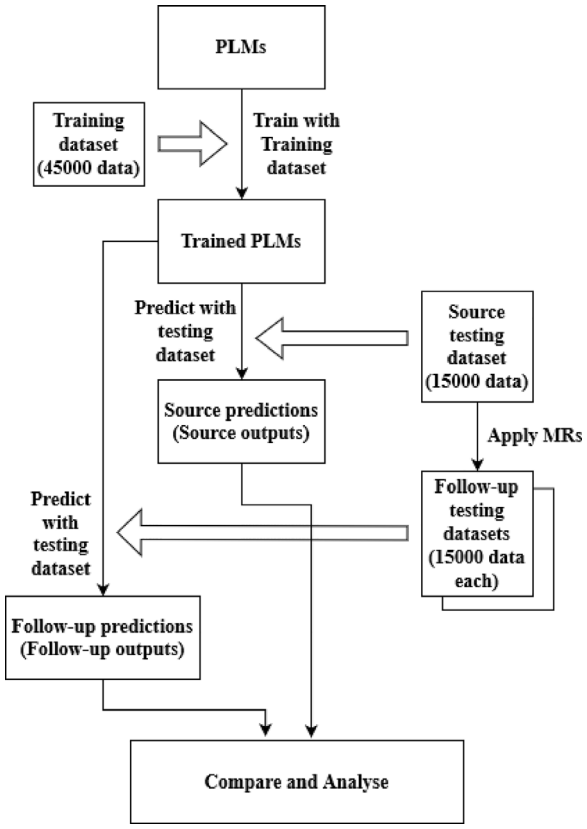

Fig. 1. Illustration of the experiment (Step 1 to 5).

Table 2
RR values of four PLMs with all MRs.

		XLNet	XLM-Roberta	GPT1	T5	Average
MRG1	MR1.1	0.41%	0.37%	**0.00%**	0.35%	0.28%
	MR1.2	0.66%	0.37%	**0.00%**	0.37%	0.35%
	MR1.3	0.97%	0.37%	**0.00%**	0.37%	0.43%
	MR1.4	1.31%	0.37%	**0.00%**	0.37%	0.51%
	MR1.5	1.57%	0.37%	**0.00%**	0.37%	0.58%
MRG2	MR2.1	0.02%	0.02%	0.04%	0.04%	0.03%
	MR2.2	0.05%	0.11%	0.39%	0.11%	0.17%
	MR2.3	0.06%	0.11%	0.40%	0.11%	0.17%
	MR2.4	0.06%	0.11%	0.40%	0.14%	0.18%
MRG3	MR3.1	24.30%	19.40%	**0.00%**	17.85%	15.39%
	MR3.2	26.95%	21.27%	**0.00%**	20.39%	17.15%
	MR3.3	27.09%	21.33%	**0.00%**	20.48%	17.23%
	MR3.4	27.11%	21.35%	**0.00%**	20.46%	17.23%

4. Result and discussion

We first studied the violations of all MRs with all subject PLMs to evaluate the effectiveness of our proposed MRGs (RQ1). Then, we analysed the zero-violation cases to seek an understanding of locating the non-violation region to support simulation MT (RQ2). Finally, we further analysed the results to examine the reasons for the violations.

4.1. Zero violation with MRG1

Table 2 shows that GPT1 had zero RR with all MRs of MRG1. Although MRG1 only duplicates the content of the title, it is effective MRG as only GPT1 achieves zero violation instead of all PLMs. Thus, MRG1 can be used as a validation tool for content quality classification on Stack Overflow problems using GPT1.

Table 3
ACC values of four PLMs with all MRs and Source.

		XLNet	XLM-Roberta	GPT1	T5
SRC		85.67%	85.86%	46.15%	86.15%
MRG1	MR1.1	85.51%	85.72%	46.15%	86.06%
	MR1.2	85.51%	85.72%	46.15%	86.06%
	MR1.3	85.50%	85.72%	46.15%	86.06%
	MR1.4	85.41%	85.72%	46.15%	86.06%
	MR1.5	85.44%	85.72%	46.15%	86.06%
MRG2	MR2.1	85.69%	85.87%	46.16%	86.16%
	MR2.2	85.67%	85.89%	46.20%	86.18%
	MR2.3	85.68%	85.89%	46.20%	86.18%
	MR2.4	85.68%	85.88%	46.20%	86.19%
MRG3	MR3.1	73.70%	77.85%	46.15%	78.37%
	MR3.2	71.57%	76.57%	46.15%	76.85%
	MR3.3	71.47%	76.65%	46.15%	76.89%
	MR3.4	71.43%	76.69%	46.15%	76.93%

Table 4
MCC values of four PLMs with all MRs and Source.

		XLNet	XLM-Roberta	GPT1	T5
SRC		78.55%	78.86%	20.27%	79.22%
MRG1	MR1.1	78.30%	78.63%	20.27%	79.09%
	MR1.2	78.28%	78.63%	20.27%	79.09%
	MR1.3	78.27%	78.63%	20.27%	79.09%
	MR1.4	78.13%	78.63%	20.27%	79.09%
	MR1.5	78.17%	78.63%	20.27%	79.09%
MRG2	MR2.1	78.58%	78.87%	20.28%	79.24%
	MR2.2	78.55%	78.90%	20.32%	79.27%
	MR2.3	78.56%	78.90%	20.32%	79.27%
	MR2.4	78.56%	78.89%	20.32%	79.29%
MRG3	MR3.1	61.22%	67.01%	20.27%	68.95%
	MR3.2	58.70%	65.14%	20.27%	67.60%
	MR3.3	58.56%	65.25%	20.27%	67.79%
	MR3.4	58.51%	65.30%	20.27%	67.85%

Table 5
Four extra MRs related to uppercase conversion.

MR group (MRG)	Descriptions
Extra MRG3: Uppercase conversion	MR3.5 — Convert 10% texts to uppercase MR3.6 — Convert 37.5% texts to uppercase MR3.7 — Convert 62.5% texts to uppercase MR3.8 — Convert 87.5% texts to uppercase

As GPT1 considers all MRs in MRG1 to be semantic equivalent, any case that violates any MRs of MRG1 can reveal a fault in the model. Therefore, this result can be interpreted as "duplicating the content of the title from one time to six times" is the non-violation region. In other words, if GPT1 finds Stack Overflow questions with duplicated content of the title (one time to six times) semantically different, this can be interpreted as a fault in the model without testing. As this region can determine whether the question is violated without testing (Ying et al., 2022), this region can support simulation metamorphic testing.

Answering RQ1 and RQ2: MRG1 can validate GPT1 in content quality classification on Stack Overflow problems. In addition, as GPT1 did not violate the whole range of MRs in MRG1, this result can create a non-violation region and prove that MRs of MRG1 can support simulation metamorphic testing for GPT1.

4.2. Zero violation with MRG3

Table 2 shows GPT1 has zero RR with all MRs of MRG3. We further tested GPT1 with another set of extra MRs related to uppercase conversion (Table 5).

GPT1 also achieved zero violations with the four extra MRs related to MRG3 (Table 6). Results show that GPT1 considered all MRs in MRG3 to be semantically equivalent, even though GPT1 is not an

Table 6
RR values of GPT1 with extra MRs related to uppercase conversion.

	GPT1
MR3.5	0.00%
MR3.6	0.00%
MR3.7	0.00%
MR3.8	0.00%

uncased version. Thus, MRG3 can be used as a validation tool for the content quality classification on Stack Overflow problems using GPT1.

GPT1 finds all MRs, including extra MRs, in MRG3 to be semantically equivalent, the results indicate no violation from 0% to 100% uppercase conversion. Any case that violated any MRs of MRG3 can reveal a fault in the model. Therefore, this result can be interpreted as "uppercase conversion from 0% to 100% of all texts" is the non-violation region. In other words, if GPT1 finds Stack Overflow questions semantically different between uppercase and lowercase format, this can be interpreted as a fault in the model without testing. As this region can determine whether the question is violated without testing (Ying et al., 2022), this region can support simulation metamorphic testing.

Answering RQ1 and RQ2: MRG3 can validate GPT1 in the content quality classification on Stack Overflow problems. In addition, as GPT1 did not violate any MR in MRG3, these results create a non-violation region and prove that MRs of MRG3 can support simulation metamorphic testing for GPT1.

4.3. Further analysis

GPT1 performance. Although GPT1 achieves zero violation with MRG1 and MRG3, Tables 3 and 4 show that the ACCs and MCCs of GPT1 are the lowest in the study. It is argued that the uni-directional model, which is GPT1, could be less useful than bi-directional models, such as XLM-Roberta, in making inferences at the sentence level or in learning common sense (Devlin et al., 2018; Zhou et al., 2020a). It is because uni-directional model, such as GPT1, uses a left-to-right architecture, where every token can only attend to its left context, whereas a bidirectional model incorporates context from both directions (Devlin et al., 2018; Zhou et al., 2020a). It can be explained by GPT1's worse performance in RR with MRG2 than with MRG1 and MRG3. MRG1 duplicates the content of the title, and MRG3 changes the case format of the texts; those MRs do not alter the left and right positions of texts in the contexts. On the other hand, MRG2 is expanding abbreviations in the title and body texts to their full forms, and these changes have a more significant impact on the position of texts in the contexts.

For example, GPT1 might find "I don't eat and don't drink" and "I do not eat and do not drink" semantically different as based on the left-to-right architecture, tokens "eat" and "drink" might be interpreted differently as the number of left tokens is changed. On the other hand, "I don't eat" and "I DON'T EAT" or "I don't eat" and "I don't eat I don't eat", tokens "eat" and "drink" might not be interpreted differently as the changes on the left have a less significant impact on the position of texts in the contexts.

Therefore, it is essential to understand the associated PLMs before selecting appropriate semantic-preserving MRGs for validation and simulation metamorphic testing. This study demonstrates that MRs, which alter the position of texts in the contexts while preserving their semantic meaning, can impact the stability of the performance of GPT1 and other uni-directional models. This makes the abbreviation replacement type of MR (MRG2) unsuitable as a validation tool for uni-directional models.

Non-zero violation. Three PLMs violated all MRs of MRG1, all PLMs violated all MRs of MRG2, and three PLMs violated all MRs of MRG3 (Table 2). MRs are supposed to be semantic-preserving; however, the results show that most PLMs do not comply and re-label the data, which indicates that PLMs considered the follow-up cases to be not semantically equivalent to source cases.

Firstly, Tables 2–4 show that XLNet, XLM-Roberta and T5's RRs of all MRs of MRG3 are over 17.85%, and their ACCs and MCCs of follow-up outputs are lower than the ACCs and MCCs of source outputs. It can be explained that these three PLMs are case-sensitive. Thus, uppercase conversion is not a suitable validation tool for XLNet, XLM-Roberta and T5.

Secondly, the average RR of MRG1 and MRG2 are less than 1%, and MRG2 has the lowest average RR in the study. However, no PLM achieved zero violation in MRG2. We reviewed the classification report of all PLMs and located the re-label cases with MRG2 of all PLMs. We further investigated those violated cases and found that all the abbreviation replacements took place in the title text or the first line of the body text. In other words, the position of abbreviation replacement in the context can affect the prediction of PLMs, which agrees with Browning and LeCun (2023) that PLMs use the statistical likelihood of words and patterns to disambiguate the sentence without the need to know the semantic content of the sentences.

These results also agree with Zhou et al. (2020a) that the common sense knowledge of PLMs remains at a surface level instead of deep semantic comprehension, and models are still behind human-level common sense reasoning. As all MRs in this study are semantic-preserving, it is not logical to find contexts semantically different due to the case format, the position of abbreviation replacements or the number of duplications of the title texts. Even though all the PLMs in this study had already trained with a training dataset containing 45,000 Stack Overflow questions with quality ratings, the levels of common sense knowledge of PLMs might not be improved due to the limitations of PLMs' capacities (Zhou et al., 2020a).

Thirdly, the results show that PLMs use statistical patterns to disambiguate the sentence, so studying those statistical patterns and exploring informative statistical features could be an alternative method to improve the performance of PLMs. Li et al. (2021) demonstrated the potential of incorporating highly informative statistical features, such as word frequency on labels, into PLMs to significantly enhance their performances. Thus, further research is necessary to explore different informative statistical features on different labels and turn them into effective representations to enhance the performance of PLMs.

Therefore, the semantic comprehension and common sense knowledge of PLMs remains at the surface level, so most of the PLMs in the study cannot achieve zero violation with the proposed semantic-preserving MRs. Further investigations on how to improve the levels of semantic comprehension and common sense knowledge of PLMs, as well as how to explore different informative statistical features and turn them into effective representations, are required to improve their performances.

5. Threats to validity

This section discusses the threats to validity that are faced by this study.

5.1. Internal validity

Internal validity refers to the degree of confidence to which the causal relationship studied in research is independent of other factors, so the correctness of the experimental environment is significant for internal validity. One internal threat to the experiment is parameter configuration, as different input parameters can cause result variations. Standard parameters were applied to all PLMs, and we built PLMs with the pretrained models provided by the machine learning community "Hugging Face" and the libraries of "Python". We also used Microsoft Excel to create the "Follow-up" testing datasets. These tools and technologies should be considered reliable since they have been widely used. White-box testing has also been conducted so the source codes

are error-free. Another concern is the measurements of this study. We focused on the relationship between MRs and measurements, and the prediction outcomes of all PLMs were listed and checked to ensure that the measurements mentioned in Section 3.6 were calculated accurately.

5.2. External validity

External validity refers to how well the outcome of a study can be expected to apply to other settings, which means how generalizable the findings are. One external threat is the generality of our approach. This study uses semantic-preserving MRs, which are generic and applicable to any experiment using the English language, so the potential threat to the external validity of this study is low. This study chose the "Stack Overflow Questions with Quality Rating" dataset of Kaggle, which has 60,000 data, and four different PLMs were employed to run trials with thirteen MRs. Thus, this study had four sets of source outputs and fifty-two sets of follow-up outputs for comparison. We believe the results should be generalized. At last, although the datasets we used for experiment assessment may vary case by case, the subject PLMs were assessed and evaluated by MT, which is a widely used and general approach. Thus, we argue that the dataset's effect on our approach's effectiveness should be small.

5.3. Construct validity

Construct validity concerns the extent to which the test or measure accurately assesses what it is supposed to. The independent variables of this study are prediction outcomes of the "Source" and "Follow-up" testing datasets of four different PLMs, which were trained by the same training dataset. The dependent variable is the differences between the independent variables, which are the differences between the prediction outcomes of "Source" and "Follow-up" testing datasets of PLMs, the RR. As this study focuses on validating PLMs, RR could show the violations. Suppose there is no violation between source prediction outcomes and follow-up prediction outcomes, which means no inconsistent prediction between them. In that case, RR is zero, and the associated PLM is insensitive to the related MR, deeming the MR suitable for validation. Conversely, violating the MR could reveal an error in PLM because the related PLM should be insensitive to the MR. That can correctly measure the intent of this study. To sum up, the construct validity of the independent variables and dependent variables of this study should be acceptable.

6. Conclusion and future direction

PLMs have been leveraged to enhance and maintain the content quality of Q&A websites; however, evaluating PLMs is a test oracle problem. This study proposed using MT to alleviate the test oracle problem, and three MRGs containing thirteen semantic-preserving MRs, which revolve around the concept of symmetry, were defined. We experimented with the proposed MRGs and four PLMs on the content quality classification of the Stack Overflow Q&A test scenario.

In conclusion, this study proved that the proposed MRs serve as effective validation tools for the content quality classification on Stack Overflow Q&A using PLMs. GPT1 did not violate MRG1 and MRG3. Additionally, the zero violations in the ranges of MRs in MRG1 and MRG3 highlighted the presence of non-violation regions, proving that the proposed MRs can support simulation metamorphic testing. Therefore, if the Stack Overflow question fails in the non-violation region and the non-violation PLM finds the question semantically different, this can be interpreted as a fault in the PLM without testing.

Despite GPT1 did not violate two MRGs, its performance in ACC and MCC rank the lowest in this study. It can be explained by the unidirectional structure of GPT1 (Devlin et al., 2018; Zhou et al., 2020a), so GPT1 only has no violation with those MRs do not alter the position of texts in the contexts.

Furthermore, most PLMs cannot achieve zero violations in this study. It can be explained by PLMs that they might use the statistical likelihood of words and patterns, instead of semantic understanding, to perform language tasks (Browning and LeCun, 2023). The common sense knowledge of PLMs remains at a surface level instead of deep semantic comprehension, which makes models behind human-level common sense reasoning (Zhou et al., 2020a). Despite training all PLMs in this study with a training dataset containing 45,000 Stack Overflow questions with quality ratings, the training did not notably improve the level of common sense of PLMs (Zhou et al., 2020a). Further investigations on improving levels of semantic comprehension and common sense knowledge of PLMs, as well as to explore informative statistical features and transform them into effective representations, are required to improve their performances.

6.1. Future direction

Some aspects could extend this research. Firstly, subsequent studies might replicate the experiment of this study on different domains, such as Q&A websites using different languages or by employing more sophisticated generative pretrained large language models, such as ChatGPT4, Meta Llama, Gemini or Mistral. Secondly, future research could investigate diverse MRs that can evaluate the common sense knowledge and semantic comprehension abilities of pretrained models. Thirdly, more studies are needed to explore finding non-violation regions in different domains to support simulation metamorphic testing. Fourthly, another valuable research focus is exploring highly informative statistical features as additional information to PLMs and experimenting with the proposed MRs or different MRs. This investigation could assist in better understanding PLMs' performances and discovering more suitable MRs for PLMs' validation tools.

6.2. Limitation

This study only experimented on a three-labelled classification problem, which might lead to a high percentage of mapping. Additionally, due to the constraints on computational resources, this study was conducted using the basic versioned PLMs that utilize a minimal number of parameters.

CRediT authorship contribution statement

Pak Yuen Patrick Chan: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jacky Keung:** Writing – review & editing, Supervision.

Declaration of generative AI in scientific writing

The authors declare that they have not used any generative artificial intelligence (AI) and AI-assisted technologies in the writing process.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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, Hong Kong (6000796, 9229109, 9229098, 9220103, 9229029).

References

Annamoradnejad, I., Habibi, J., Fazli, M., 2022. Multi-view approach to suggest moderation actions in community question answering sites. Inform. Sci. 600, 144–154.

Browning, J., LeCun, Y., 2023. Language, common sense, and the winograd schema challenge. Artificial Intelligence 104031.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Wang, Y., 2024. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. 15 (3), 1–45.

Chen, T.Y., Cheung, S.C., Yiu, S.M., 2020. Metamorphic Testing: A New Approach for Generating Next Test Cases. ArXiv.org http://dx.doi.org/10.48550/arxiv.2002.12543.

Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21 (1), 6. http://dx.doi.org/10.1186/s12864-019-6413-7.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. ArXiv preprint arXiv:1911.02116.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

Duque-Torres, A., Pfahl, D., Klammer, C., Fischer, S., 2023. Bug or not bug? Analysing the reasons behind metamorphic relation violations. In: Paper Presented at the 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering. SANER.

Ellis, J.D., Iqbal, R., Yoshimatsu, K., 2021. Verification of the neural network training process for spectrum-based chemical substructure prediction using metamorphic testing. J. Comput. Sci. 55, 101456.

Gao, S., Gao, L., Li, Q., Xu, J., 2023. Application of large language model in intelligent q & a of digital government. In: Paper Presented at the Proceedings of the 2023 2nd International Conference on Networks, Communications and Information Technology.

HuggingFace, 2024a. Google's T5 Version 1.1. Retrieved from https://huggingface.co/google/t5-v1_1-base.

HuggingFace, 2024b. OpenAI GPT 1. Retrieved from https://huggingface.co/openai-community/openai-gpt.

HuggingFace, 2024c. XLM-RoBERTa (base-sized model). Retrieved from https://huggingface.co/FacebookAI/xlm-roberta-base.

HuggingFace, 2024d. XLNet (base-sized model). Retrieved from https://huggingface.co/xlnet/xlnet-base-cased.

Kamienski, A., Hindle, A., Bezemer, C.-P., 2023. Analyzing techniques for duplicate question detection on Q & A websites for game developers. Empir. Softw. Eng. 28 (1), 17.

Li, X., Li, Z., Xie, H., Li, Q., 2021. Merging statistical feature via adaptive gate for improved text classification. In: Paper Presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Mousavi, R., Raghu, T., Frey, K., 2020. Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums. J. Manage. Inf. Syst. 37 (4), 1073–1098.

Pugh, S., Raunak, M.S., Kuhn, D.R., Kacker, R., 2019. Systematic testing of post-quantum cryptographic implementations using metamorphic testing. In: Paper Presented at the 2019 IEEE/ACM 4th International Workshop on Metamorphic Testing. MET.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (1), 5485–5551.

Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., Tonella, P., 2020. Testing machine learning based systems: a systematic mapping. Empir. Softw. Eng. 25 (6), 5193–5254.

Saha, P., Kanewala, U., 2019. Fault detection effectiveness of metamorphic relations developed for testing supervised classifiers. In: Paper Presented at the 2019 IEEE International Conference on Artificial Intelligence Testing. AITest.

Segura, S., Durán, A., Troya, J., Ruiz-Cortés, A., 2019. Metamorphic relation patterns for query-based systems. In: Paper Presented at the 2019 IEEE/ACM 4th International Workshop on Metamorphic Testing. MET.

Segura, S., Towey, D., Zhou, Z.Q., Chen, T.Y., 2018. Metamorphic testing: Testing the untestable. IEEE Softw. 37 (3), 46–53.

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

Shazeer, N., 2020. Glu variants improve transformer. ArXiv preprint arXiv:2002.05202.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Pfohl, S., 2023. Large language models encode clinical knowledge. Nature 620 (7972), 172–180.

Stacy, B., Hauzel, J., Lindvall, M., Porter, A., Pop, M., 2022. Metamorphic testing in bioinformatics software: A case study on metagenomic assembly. In: Paper Presented at the 2022 IEEE/ACM 7th International Workshop on Metamorphic Testing. MET.

Sun, C.-a., Liu, B., Fu, A., Liu, Y., Liu, H., 2022. Path-directed source test case generation and prioritization in metamorphic testing. J. Syst. Softw. 183, 111091. http://dx.doi.org/10.1016/j.jss.2021.111091.

Wen, A., Elwazir, M.Y., Moon, S., Fan, J., 2020. Adapting and evaluating a deep learning language model for clinical why-question answering. JAMIA Open 3 (1), 16–20.

Xiao, D., Z, L.I.U., Yuan, Y., Pang, Q., Wang, S., 2022. Metamorphic testing of deep learning compilers. Proc. ACM Meas. Anal. Comput. Syst. 6 (1), http://dx.doi.org/10.1145/3508035, Article 15.

Xie, X., Ho, J.W.K., Murphy, C., Kaiser, G., Xu, B., Chen, T.Y., 2011. Testing and validating machine learning classifiers by metamorphic testing. J. Syst. Softw. 84 (4), 544–558. http://dx.doi.org/10.1016/j.jss.2010.11.920.

Xie, X., Zhang, Z., Chen, T.Y., Liu, Y., Poon, P.-L., Xu, B., 2020. METTLE: A metamorphic testing approach to assessing and validating unsupervised machine learning systems. IEEE Trans. Reliab. 69 (4), 1293–1322.

Xu, B., Nguyen, T.-D., Le-Cong, T., Hoang, T., Liu, J., Kim, K., Le, B., 2023. Are we ready to embrace generative AI for software Q & A? In: Paper Presented at the 2023 38th IEEE/ACM International Conference on Automated Software Engineering. ASE.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32.

Ying, Z., Bellotti, A., Towey, D., Chen, T.Y., Zhou, Z.Q., 2022. Using metamorphic relation violation regions to support a simulation framework for the process of metamorphic testing. In: Paper Presented at the 2022 IEEE 46th Annual Computers, Software, and Applications Conference. COMPSAC.

Ying, Z., Towey, D., Bellotti, A., Zhou, Z.Q., Chen, T.Y., 2021. Preparing SQA professionals: Metamorphic relation patterns, exploration, and testing for big data. In: Proceedings of the International Conference on Open and Innovation Education. ICOIE 2021, pp. 22–30.

Zhang, M., Keung, J.W., Chen, T.Y., Xiao, Y., 2021. Validating class integration test order generation systems with metamorphic testing. Inf. Softw. Technol. 132, 106507.

Zhang, F., Liu, J., Wan, Y., Yu, X., Liu, X., Keung, J., 2023. Diverse title generation for stack overflow posts with multiple-sampling-enhanced transformer. J. Syst. Softw. 200, 111672.

Zhang, Z., Yang, J., 2024. HintMiner: Automatic question hints mining from Q & A web posts with language model via self-supervised learning. In: Paper Presented at the International Conference on Artificial Intelligence and Statistics.

Zhi, W., 2024. Multi-modal medical Q & A system. In: Paper Presented at the Proceedings of the 2024 International Conference on Computer and Multimedia Technology.

Zhou, Z.Q., Sun, L., Chen, T.Y., Towey, D., 2020a. Metamorphic relations for enhancing system understanding and use. IEEE Trans. Softw. Eng. 46 (10), 1120–1154.

Zhou, X., Zhang, Y., Cui, L., Huang, D., 2020b. Evaluating commonsense in pre-trained language models. In: Paper Presented at the Proceedings of the AAAI Conference on Artificial Intelligence.