Mashee at SemEval-2024 Task 8: The Impact of Samples Quality on the Performance of In-Context Learning for Machine Text Classification

Areeg Fahad Rasheed
College of Information Engineering
Al-Nahrain University
Baghdad, Iraq
areeg.fahad@coie-nahrain.edu.iq

M. Zarkoosh
Software Engineering
Computiq
Baghdad, Iraq
m94zarkoosh@gmail.com

Abstract

Within few-shot learning, in-context learning (ICL) has become a potential method for leveraging contextual information to improve model performance on small amounts of data or in resource-constrained environments where training models on large datasets is prohibitive. However, the quality of the selected sample in a few shots severely limits the usefulness of ICL. The primary goal of this paper is to enhance the performance of evaluation metrics for in-context learning by selecting high-quality samples in few-shot learning scenarios. We employ the chi-square test to identify high-quality samples and compare the results with those obtained using low-quality samples. Our findings demonstrate that utilizing high-quality samples leads to improved performance with respect to all evaluated metrics.

1 Introduction

The advent of large language models (LLMs) like GPT-3.5 has brought about transformative capabilities, seamlessly handling tasks like question answering, essay writing, and problem-solving (Aljanabi et al., 2023; Wu et al., 2023; Rasheed et al., 2023a). However, this technological advancement necessitates careful consideration of its associated challenges. Concerns regarding the potential impact on creativity and ethical implications, particularly concerning the generation of deepfakes (Tang et al., 2023), warrant careful attention (RAYMOND, 2023). Additionally, the limitations of LLMs, including the possibility of producing erroneous information, require rigorous evaluation and verification. The substantial energy consumption required for training LLMs on massive datasets raises environmental concerns, contributing to their carbon footprint. Moreover, plagiarism issues emerge as users may misuse the generated content, either inadvertently or intentionally (Hadi et al., 2023).

Various models have been introduced in recent years designed to distinguish text generated by humans from that created by machines(Mitchell et al., 2023). Examples include GPTZero(gpt), AI Content Detector(cop), and AI Content Detector by Writer (wri) among others. Some of these models are trained on specific datasets, while others are commercially available. Designing and implementing LLMs for classification tasks requires substantial resources and computational power, which are often only accessible to institutions and governments. Therefore, various optimization models, such as LoRA (Hu et al., 2021), distillation(Hsieh et al., 2023), quantization(Dettmers et al., 2022), and in-context learning (Liu et al., 2022), have been developed to reduce the resource requirements for LLM implementation. This paper focuses on In Context Learning (ICL) (Liu et al., 2022), which utilizes the capabilities of other models to enhance their ability to classify AI-generated text.

In Context Learning (ICL) is a Natural Language Processing (NLP) technique utilized to enable Large Language Models (LLMs) to learn new tasks based on minimal examples. This technique proves powerful in scenarios where training models on extensive datasets is impractical or when there are constraints on dataset availability for a specific task. ICL operates on the premise that humans can often acquire new tasks through analogy or by observing a few examples of task performance. It can be employed without any examples and is referred to as zero-shot learning. Alternatively, if the input includes one example, it is termed one-shot learning, and if it contains more than one, it is known as few-shot learning. This paper focuses on the application of few-shot learning within the context of ICL(Ahmed and Devanbu, 2022; Kang et al., 2023).

In this study, our focus lies exclusively on few-shot learning. We present a methodology that leverages the chi-square statistic (Rasheed et al., 2023b;

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 60–63
June 20-21, 2024 ©2024 Association for Computational Linguistics

Lancaster and Seneta, 2005) to select samples for few-shot learning and evaluate its impact on the performance of a machine-generated text classification model. We work on task A English language only (Wang et al., 2024).

2 Dataset

The dataset employed for Task A comprises two main components. The first part, derived from human writing, was collected from diverse sources including WikiBidia, WikiHow, Reddit, ArXiv, and PeerRead. The second part consists of a machine-generated text produced by ChatGPT, Cohere, Dolly-v2, and BLOOMz(Muennighoff et al., 2023). For further details, please refer to the associated paper (Wang et al., 2023).

3 Chi-square

Chi-square is a statistical test used to assess the independence of two categorical variables. It calculates the difference between observed and expected frequencies of outcomes, and a larger chi-square value indicates a stronger rejection of independence. In text analysis, chi-square can be used to identify keywords that are more likely to occur in one category than another, making it useful for feature selection and text classification. We computed the chi-square values for each training sample and recorded the sample index with the highest and lowest chi-square values for both human-generated and machine-generated samples. Table I displays the index and corresponding chi-square values for each of these instances. We will use X^2 to refer to chi-square (Lancaster and Seneta, 2005).

Table 1: Indices and chi-square values for highest/lowest in human-generated and machine-generated text

Name	Index #	X^2 Value
Highest X^2 (Human)	70873	1351.59
Lowest X^2 (Human)	85726	1.21
Highest X^2 (Machine)	2426	1154.27
Lowest X^2 (Machine)	29111	0.8243

4 System overview

The system architecture is illustrated in Figure 1. The process starts with feeding the entire training dataset to a chi-square computation, where the chi-square value for each sample is calculated. Subsequently, the indices of the samples with the highest and lowest chi-square values are selected for both human-generated and machine-generated datasets using information from Table I. Next, context learning is prepared. Initially, multiple templates were tested, and the one presented in Figure 1 yielded the best results. This template is then fed with two samples: the first being the machine-generated sample with the highest chi-square value, and the second being the human-generated sample with the highest chi-square value. Due to context window size limitations, only the first 5000 characters of each sample are incorporated. This is applied to training samples exceeding 5000 characters to ensure the context learning size is not exceeded. Finally, the test sample is fed into the context-learning process. The Flan-T5 model large version is used. The results are then recorded and evaluated. The dev/test sample size was truncated to 3000. We also evaluated the system using samples with the lowest chi-square values and doing the same process.

5 Findings and Analysis

We employed the Flan-T5 Large model for both the development and testing datasets. We selected samples from both human-generated and machine-generated sources, with each sample limited to 5000 characters to avoid exceeding the token size limit. A total of four experiments were conducted. The first experiment utilized samples with high chi-square values from the development set. The second experiment focused on samples with the smallest chi-square values from the development set. The third experiment involved samples with high chi-square values from the test set. Finally, the fourth experiment utilized samples with low chi-square values from the test set. Table II presents all achieved results.

Based on the results presented in Table II, we can discuss several key points.

- The results highlight the crucial role of sample quality in the performance of in-context learning. By leveraging the chi-squared metric and prioritizing samples with high values, we essentially provide the Flan-T5 model with examples rich in diverse features. This choice enables the Flan-T5 model to learn more effectively, drawing substantial insights from the samples. Consequently, the model becomes more familiar with the provided data, ultimately enhancing its performance. In

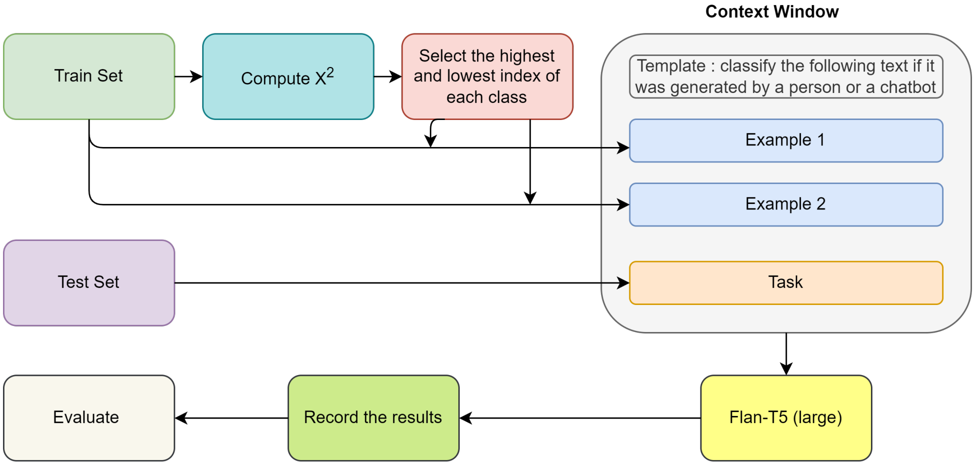
Context Window
Train Set
Compute X²
Select the highest and lowest index of each class
Template : classify the following text if it was generated by a person or a chatbot
Example 1
Example 2
Test Set
Task
Evaluate
Record the results
Flan-T5 (large)

Figure 1: Proposed System Components

Dataset	Chi Type	Recall	Precision	F1-Score	Accuracy
Dev set	Lowest	46.92	46.90	46.84	46.92
	Highest	53.76	53.76	53.74	53.76
Test set	Lowest	55.04	55.07	55.03	55.27
	Highest	58.68	58.81	58.81	**55.99**

Table 2: Experiments results

contrast, selecting samples with lower quality leads to less optimal performance. This can be noticed for both the dev and test set. The main reason behind this is that words in the sample with high chi-square values contain the most distinctive features. This is because the chi-square test assigns high values to words that are frequent within a particular class but appear less frequently in other classes.Conversely, samples with lower chi-square values likely contain more random words that appear with similar frequency across all classes. In chi-square analysis, words that appear equally or approximately equally in each class receive lower scores.

- The classification of machine-generated text represents a novel frontier in machine learning, and the availability of datasets for this task is currently limited. The dataset used in this study was generated in 2023, marking it as a recent development and underscoring the lack of established benchmarks. Models that support in-context learning have not been trained extensively on such tasks, resulting in lower accuracy when applied. While examples with high-quality data can enhance model performance, it remain below the desired threshold. Hence, it is advisable to train the model directly on the dataset rather than relying on in-context learning.

- We have utilized the Flan-T5 model; however, other models can be employed to evaluate the performance of text classification machinery. We suggest considering alternatives such as bard, Jurassic-1 Jumbo, and ChatGPT.

6 Conclusion

This work presents a system for classifying human-generated and machine-generated text. The system leverages the combined strengths of in-context learning and Chi-square analysis. Chi-square is employed to select high-quality samples from the trainin dataset for few-shot learning in the in-context learning. We implement Flan-T5 model large version for in-context learning. Evaluation using accuracy, recall, precision, and F1-score demonstrates that selecting high-quality samples improves system performance for both dev and test. Furthermore, the results indicate that relying solely on in-context learning for new tasks like machine-generated text detection yields relatively low performance.

References

Ai content detector. Accessed on March 30, 2024.

Ai content detector by writer. Accessed on March 30, 2024.

Gptzero. Accessed on March 30, 2024.

Toufique Ahmed and Premkumar Devanbu. 2022. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–5.

Mohammad Aljanabi, Mohanad Ghazi, Ahmed Hussein Ali, Saad Abas Abed, et al. 2023. Chatgpt: open possibilities. *Iraqi Journal For Computer Science and Mathematics*, 4(1):62–64.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Muhammad Usman Hadi, R Qureshi, A Shah, M Irfan, A Zafar, MB Shaikh, N Akhtar, J Wu, and S Mirjalili. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Sungmin Kang, Juyeon Yoon, and Shin Yoo. 2023. Large language models are few-shot testers: Exploring llm-based general bug reproduction. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2312–2323. IEEE.

Henry Oliver Lancaster and Eugene Seneta. 2005. Chi-square distribution. *Encyclopedia of biostatistics*, 2.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Areeg Fahad Rasheed, M Zarkoosh, Safa F Abbas, and Sana Sabah Al-Azzawi. 2023a. Arabic offensive language classification: Leveraging transformer, lstm, and svm. In *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–6. IEEE.

Areeg Fahad Rasheed, M Zarkoosh, and Sana Sabah Al-Azzawi. 2023b. The impact of feature selection on malware classification using chi-square and machine learning. In *2023 9th International Conference on Computer and Communication Engineering (ICCCE)*, pages 211–216. IEEE.

DANIEL RAYMOND. 2023. Disadvantages of large language models. Accessed on March 30, 2024.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. SemEval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.