LLM Detective: A Comparative Study of Supervised and Zero-Shot Methods for Detecting Texts Generated by Large Language Models

Jaouhara Zerhouni Khal
Computer Science and Engineering
Southern Univ. of Sci. and Tech.
Shenzhen, China
12211456@mail.sustech.edu.cn

Hok Layheng
Computer Science and Engineering
Southern Univ. of Sci. and Tech.
Shenzhen, China
12210736@mail.sustech.edu.cn

Harrold Tok Kwan Hang
Computer Science and Engineering
Southern Univ. of Sci. and Tech.
Shenzhen, China
12212025@mail.sustech.edu.cn

***Abstract*—The widespread adoption of large language models (LLMs) has raised concerns about the potential for generating misleading or harmful content, necessitating reliable methods to distinguish between human-written and LLM-generated texts. This study compares supervised and zero-shot approaches for detecting LLM-generated texts across English, Chinese, and multilingual datasets. We fine-tune BERT models for supervised classification and implement the FourierGPT zero-shot method, evaluating both on accuracy, precision, recall, F1 score, and AUROC. Our results demonstrate that supervised methods significantly outperform zero-shot approaches, achieving higher accuracy and AUROC across all datasets, particularly in English and Chinese contexts. However, the zero-shot method offers flexibility in scenarios with limited labeled data, showing competitive precision in Chinese and multilingual settings. These findings underscore the trade-offs between the two approaches and highlight the need for further research into robust, language-agnostic detection methods.**

***Index Terms*—Large Language Models, Text Detection, Supervised Learning, Zero-Shot Learning, BERT, FourierGPT**

I. Introduction

Large language models (LLMs) have become integral to various applications, from content generation to conversational agents. However, their widespread use has raised concerns about the potential for generating misleading or harmful content. This has necessitated the development of reliable methods to detect whether a given text was generated by an LLM or written by a human. Existing research has primarily focused on supervised learning approaches, where models are trained on labeled datasets to distinguish between human-written and LLM-generated texts [1–3]. However, supervised methods require large labeled datasets, which may not always be available, especially for languages other than English. To address this limitation, zero-shot detection methods have been proposed, leveraging the inherent properties of LLMs without requiring labeled training data [4]. In this paper, we conduct a comparative study of supervised and zero-shot methods for detecting LLM-generated texts across English, Chinese, and multilingual datasets, evaluating the performance of fine-tuned BERT models against the FourierGPT zero-shot approach.

II. Methodology

A. Datasets

We use three datasets to evaluate our methods:

- **English Dataset**: Derived from the Ghostbuster dataset, with the ’essay’ domain used for training and validation, and the ’wp’ domain as an out-of-distribution (OOD) test set.
- **Chinese Dataset**: Sourced from the Face2 dataset, using the ’news’ domain for training and validation, and the ’wiki’ domain for OOD testing.
- **Multilingual Dataset**: Combines English ’essay’ and Chinese ’news’ domains for training and validation, with English ’wp’ and Chinese ’wiki’ domains for testing.

B. Preprocessing

For supervised methods, texts are tokenized using language-specific BERT tokenizers and standardized to a maximum length of 512 tokens. To address class imbalance, where LLM-generated texts (label 1) were approximately six times more frequent than human-written texts (label 0), we employed undersampling. This involved randomly selecting a subset of the majority class to match the number of samples in the minority class, ensuring a balanced dataset for training. For zero-shot methods, we follow the FourierGPT approach [4], extracting texts from JSONL files, computing negative log-likelihood (NLL) scores with pre-trained models, and applying Fourier transforms to generate spectrum data.

C. Supervised Learning Approach

We fine-tune BERT models for classification:

- **English**: BERT-base-uncased
- **Chinese**: BERT-base-chinese
- **Multilingual**: BERT-base-multilingual-cased

Training uses the Hugging Face Transformers library with a batch size of 16, 10 epochs, and model selection based on the validation F1 score.

D. Zero-Shot Detection Approach

The zero-shot method replicates FourierGPT [4]:

1) Compute NLL scores using Mistral-7B (English) and Qwen-7B (Chinese, multilingual).
2) Normalize NLL scores with z-score.
3) Apply Fourier transforms to obtain spectrum data.
4) Classify texts using a heuristic based on the power sum of the first k frequencies, with k optimized on validation data.

E. Evaluation Metrics

Performance is assessed using:

- Accuracy
- Precision
- Recall
- F1 Score
- Area Under the ROC Curve (AUROC)

The source code for this project, including all preprocessing, training, and evaluation scripts, is publicly available at https://github.com/Layheng-Hok/LLM-Detective.

III. Results

Table I summarizes the performance of both methods across the datasets.

TABLE I: Performance Comparison

Metric	**English**	**Chinese**	**Multilingual**
Supervised			
Accuracy	0.8539	0.7665	0.5409
Precision	0.8582	0.6976	0.5215
Recall	0.8475	0.9410	0.9871
F1 Score	0.8528	0.8012	0.6825
AUROC	0.9297	0.8715	0.5509
Zero-Shot			
Accuracy	0.5317	0.5559	0.5443
Precision	0.5434	0.7182	0.6916
Recall	0.3971	0.1840	0.1598
F1 Score	0.4589	0.2929	0.2596
AUROC	0.5440	0.6502	0.6055

In the supervised learning approach, the fine-tuned BERT models achieved strong performance across all datasets. For the English dataset, the model recorded an accuracy of 0.8539 and an AUROC of 0.9297, reflecting its robust ability to distinguish between human-written and LLM-generated texts. The Chinese dataset yielded an accuracy of 0.7665 and an AUROC of 0.8715, with a notably high recall of 0.9410, indicating effective identification of LLM-generated texts. The multilingual dataset showed a lower accuracy of 0.5409 but an exceptionally high recall of 0.9871, suggesting that while it excels at detecting LLM-generated texts, it sacrifices precision, resulting in a moderate F1 score of 0.6825.

The zero-shot approach, based on the FourierGPT method, exhibited more variable performance. On the English dataset, it achieved an accuracy of 0.5317 and an AUROC of 0.5440, with a low recall of 0.3971, indicating limited effectiveness in identifying LLM-generated texts. For the Chinese dataset, the accuracy was 0.5559, with a high precision of 0.7182 but a very low recall of 0.1840, leading to a poor F1 score of 0.2929, suggesting it struggles to detect positive cases comprehensively. In the multilingual dataset, the accuracy was 0.5443, with a precision of 0.6916 but an even lower recall of 0.1598, resulting in an F1 score of 0.2596, highlighting its challenges in balancing precision and recall across diverse languages.

Comparing the two methods, the supervised approach consistently outperforms the zero-shot method in accuracy and AUROC across all datasets, benefiting from its reliance on labeled training data. The supervised method's high recall in the multilingual setting (0.9871) contrasts sharply with the zero-shot method's low recall (0.1598), underscoring the former's superiority in detecting LLM-generated texts, albeit at the cost of lower precision. The zero-shot method, while less accurate overall, offers flexibility in scenarios with limited labeled data, as evidenced by its competitive precision in the Chinese and multilingual datasets. These findings suggest that while supervised methods are preferable when resources permit, zero-shot approaches may still hold value in specific contexts. For detailed visualizations, refer to the Appendix.

IV. Future Work

While our study provides a comprehensive comparison of supervised and zero-shot methods, several avenues for future research remain. Exploring alternative zero-shot techniques, such as those based on linguistic features or stylometry, could improve performance in scenarios with limited labeled data. For instance, methods like Fast-DetectGPT, which leverages conditional probability curvature for efficient zero-shot detection [5], and GPT-who, an information density-based detector [6], offer promising directions for enhancing zero-shot performance without requiring extensive labeled datasets. Evaluating these approaches against our current FourierGPT baseline could reveal improvements in accuracy and robustness, particularly for low-resource languages. Enhancing the multilingual model's generalization across languages through cross-lingual transfer learning or language-agnostic features may address its lower accuracy. Additionally, investigating advanced preprocessing strategies, such as data augmentation, could refine both approaches. Extending this work to other languages and domains would further assess the generalizability of our findings.

V. Conclusion

This study provides a comprehensive comparison of supervised and zero-shot methods for detecting texts generated by large language models (LLMs). Our findings demonstrate that supervised approaches, leveraging fine-tuned BERT models, significantly outperform zero-shot methods in terms of accuracy and AUROC across English, Chinese, and multilingual datasets. The supervised method's ability to learn from labeled data enables it to achieve high recall, particularly in multilingual contexts, though at the expense of precision. In contrast, the zero-shot FourierGPT approach, while less accurate overall, offers a flexible alternative in scenarios where labeled data is scarce, maintaining competitive precision in Chinese and multilingual settings.

These results highlight the strengths and limitations of both detection strategies. Supervised methods excel when sufficient labeled data is available, but their reliance on such data limits their applicability in low-resource languages or domains. Zero-shot methods, though less precise, provide a viable solution in these contexts, underscoring the need for continued innovation in language-agnostic detection techniques. Future research should explore hybrid approaches that combine the strengths of both methods, as well as investigate advanced preprocessing and feature extraction techniques to enhance detection performance across diverse languages and domains. As LLMs continue to evolve, developing robust and adaptable detection methods will be critical to mitigating the risks associated with their misuse.

References

[1] N. Pangakis and S. Wolken, "Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels," 2024. [Online]. Available: https://arxiv.org/abs/2406.17633

[2] B. Alhijawi, R. Jarrar, A. AbuAlRub, and A. Bader, "Deep learning detection method for large language models-generated scientific content," *Neural Computing and Applications*, vol. 37, no. 1, p. 91–104, Nov. 2024. [Online]. Available: http://dx.doi.org/10.1007/s00521-024-10538-y

[3] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang, "Mage: Machine-generated text detection in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2305.13242

[4] Y. Xu, Y. Wang, H. An, Z. Liu, and Y. Li, "Detecting subtle differences between human and model languages using spectrum of relative likelihood," *arXiv preprint arXiv:2406.19874*, 2024.

[5] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature," in *The Twelfth International Conference on Learning Representations*, 2023.

[6] S. Venkatraman, A. Uchendu, and D. Lee, "Gpt-who: An information density-based machine-generated text detector," 2024. [Online]. Available: https://arxiv.org/abs/2310.06202

Appendix

*Supplemental Visualizations

A. Dataset Distributions

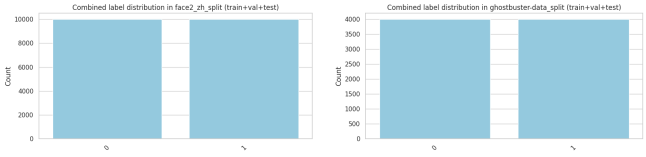

Fig. 1: Combined Label Distribution (Human vs. LLM) for Chinese (Left) and English (Right) after undersampling.

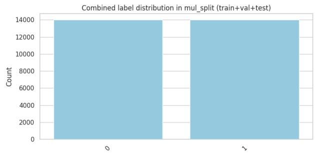

Fig. 2: Combined Label Distribution for Multilingual after undersampling.

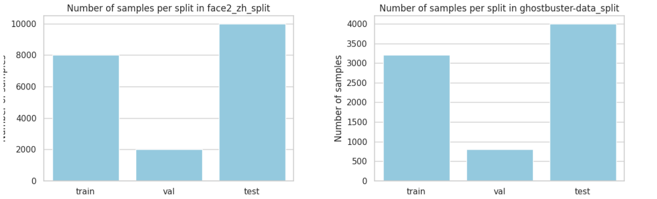

Fig. 3: Sample Counts per Split for Chinese (Left) and English (Right) after undersampling.

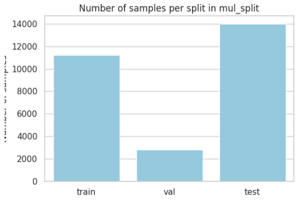

Fig. 4: Sample Counts per Split for Multilingual after undersampling.

B. Training Metrics (Supervised Learning)

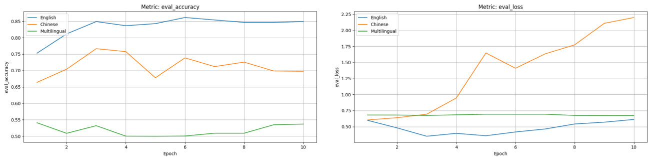

Fig. 5: Supervised Training Evaluation: (Left) Accuracy over epochs; (Right) Loss over epochs.

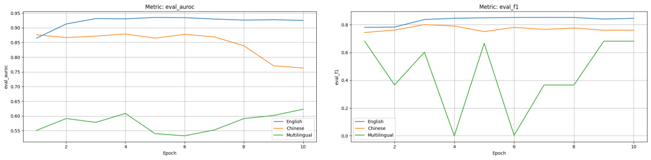

Fig. 6: Supervised Training Evaluation: (Left) AUROC over epochs; (Right) F1 Score over epochs.

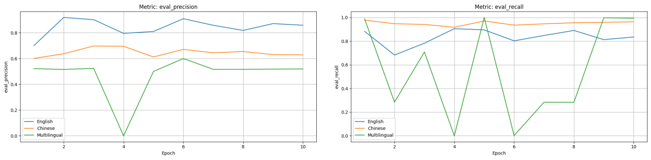

Fig. 7: Supervised Training Evaluation: (Left) Precision over epochs; (Right) Recall over epochs.

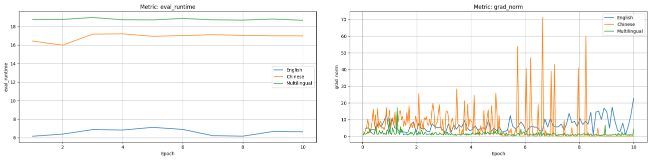

Fig. 8: Supervised Training Evaluation: (Left) Runtime per epoch; (Right) Gradient Norm per epoch.

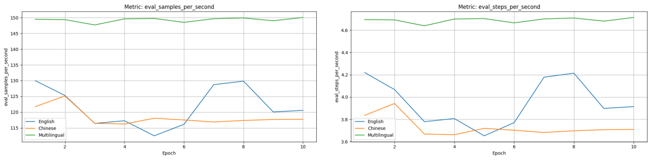

Fig. 9: Supervised Training Evaluation: (Left) Samples/sec; (Right) Steps/sec.

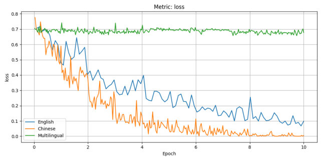

Fig. 10: Training Loss Curve (aggregated over language models).

C. Test Metrics (Supervised vs. Zero-Shot Detection)

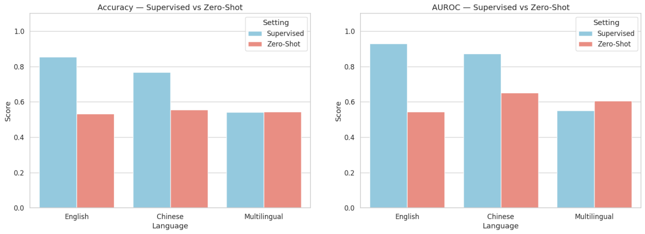

Fig. 11: Test Metrics Comparison: (Left) Accuracy; (Right) AUROC.

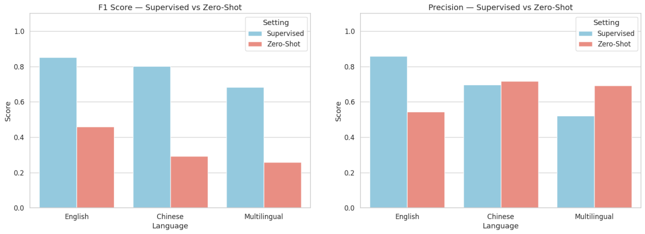

Fig. 12: Test Metrics Comparison: (Left) F1 Score; (Right) Precision.

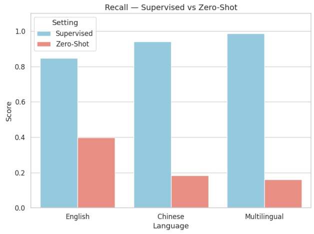

Fig. 13: Test Metrics Comparison: Recall.