Fraunhofer SIT at GenAI Detection Task 1: Adapter Fusion for AI-generated Text Detection

Karla Schäfer and **Martin Steinebach**
Fraunhofer SIT | ATHENE Center
Rheinstraße 75, Darmstadt, Germany
{karla.schaefer, martin.steinebach}@sit.fraunhofer.de

Abstract

The detection of AI-generated content is becoming increasingly important with the growing prevalence of tools such as ChatGPT. This paper presents our results in the GenAI Content Detection Task 1, focusing on binary English and multilingual AI-generated text detection. We trained and tested transformers, adapters and adapter fusion. In the English setting (Subtask A), the combination of our own adapter on AI-generated text detection based on RoBERTa with a task adapter on multi-genre NLI yielded a macro F1 score of 0.828 on the challenge test set, ranking us third out of 35 teams. In the multilingual setting (Subtask B), adapter fusion resulted in a deterioration of the results. Consequently, XLM-RoBERTa, fine-tuned on the training set, was employed for the final evaluation, attaining a macro F1 score of 0.7258 and ranking tenth out of 25 teams.

1 Introduction

The increased utilisation of tools such as ChatGPT has resulted in a notable rise in the prevalence of AI-generated text across all facets of modern life. In light of this, the development of detectors of AI-generated content is becoming increasingly important. The majority of research in this field focuses on the detection of AI-generated text in the English language. However, it is important to note that generation models are also capable of producing text in a multitude of languages. Consequently, the development of effective detectors that can perform well in multilingual settings is a crucial area of investigation. The binary English and multilingual machine-generated text detection challenge at the GenAI Content Detection Workshop (Wang et al., 2025) is focusing on this problem by encouraging its participants to develop detectors of AI-generated text on English (Subtask A) and multilingual (Subtask B) text. We participated in both subtasks of this challenge and trained and tested different transformers and adapters on the task of AI-generated text detection. We tested different adapter configurations and the utilisation of adapter fusion to transfer knowledge of pre-trained task adapters on AI-generated text detection.

2 Related Work

Several detectors of AI-generated text have been developed recently. For example, Abassy et al. (2024) developed a system, LLM-DetectAIve, which is capable of identifying not only text written by humans and machines, but also instances where the fact that a text was generated by a machine has been obfuscated, or cases where an LLM has been employed to enhance a text that was originally written by a human. Campino (2024) tested and trained different transformers on the detection of AI-generated text in the field of education. The transformers tested were ALBERT, BERT, ELECTRA, RoBERTa and XLNet. The results demonstrated that, without and with fine-tuning, BERT provided the best results, with superior results when fine-tuning. Other detectors were developed in challenges, such as the SemEval-2024 Task 8 challenge (Wang et al., 2024) or the PAN challenge at CLEF 2024 (Bevendorff et al., 2024). As far as we know, adapters and adapter fusion haven't been used before.

3 System Description

We participated in Subtask A and B of the GenAI Content Detection Task 1 challenge. In both settings, first, we tested different transformers, fine-tuning them on the respective training sets. In Subtask A, the English setting, we fine-tuned the transformers BERT (base and large; Devlin et al., 2019) and RoBERTa base (Liu et al., 2019). In Subtask B, using the multilingual training set, we fine-tuned XLM-RoBERTa base and large (Conneau et al., 2019). Subsequently, we trained our own task adapter on AI-generated text detection using the

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 178–183

respective datasets of the subtasks and performed adapter fusion (Pfeiffer et al., 2021) with official available pre-trained task adapters from Adapter Hub[1] (Pfeiffer et al., 2020). Adapter fusion is a method of combining the knowledge of multiple pre-trained adapters trained for different tasks.

3.1 Adapter Fusion

First, we trained our own adapters on the English and multilingual dataset (dependent on subtask). As basis, the best transformers from the fine-tuning step were used, being RoBERTa base on the English dataset and XLM-RoBERTa base on the multilingual dataset. In the English setting, different adapter configurations were tested. We tested the configurations LoRA (Hu et al., 2021), LoReFT (Wu et al., 2024) and sequential bottleneck (seq_bn; Houlsby et al., 2019). In the multilingual setting, we used sequential bottleneck as adapter configuration because of its superior performance in the English setting. With this, we trained our own AI-generated text detection adapter, called AI-Gen in the English setting and ml-AI-Gen in the multilingual setting.

A variety of task adapters were integrated and evaluated in conjunction with our internally developed adapters, AI-Gen/ml-AI-Gen. The selection of task adapters was based on an educated guess, with a particular emphasis on their suitability for the analysis of the structure and the perplexity of textual content. The incorporation of perplexity as an additional feature enabled the authors of Guo et al. (2024) to enhance the results of their AI-generated text detector. The same approach was attempted here using task adapters. The task adapters tested on the English dataset were pre-trained on the tasks semantic textual similarity[2], multi-genre NLI[3], adversarial NLI[4], linguistic acceptability[5] and machine reading comprehension[6] (Poth et al., 2021). For the multilingual setting we tested task adapters pre-trained on formality classification[7] (Krishna et al., 2020) and multilingual knowledge integration[8] (Hou et al., 2022).

[1] https://adapterhub.ml/
[2] roberta-base-pf-mrpc
[3] roberta-base-pf-mnli
[4] roberta-base-pf-anli_r3
[5] roberta-base-pf-cola
[6] roberta-base-pf-record
[7] xlm-roberta-base_formality_classify_gyafc_pfeiffer
[8] xlm-roberta-base_mlki_ep_pfeiffer

Model	macro F1	micro F1
BERT uncased base	0.806	0.815
BERT uncased large	0.792	0.808
RoBERTa base	**0.822**	**0.831**

Table 1: [Subtask A] English Transformer fine-tuned (test set: devtest)

Adapter (conf)	macro F1	micro F1
LoRA	0.729	0.768
LoReFT	0.679	0.738
seq_bn	**0.837**	**0.849**

Table 2: [Subtask A] English Adapters for different configurations, trained with RoBERTa base (test set: devtest)

3.2 Implementation Details

For the training of the transformers the learning rate was set to 2e-5. We also tested with a learning rate of 5e-5, but this resulted in overall worse scores, i.e. training RoBERTa base on the English training set the macro F1 score after 1 epoch reached 0.3844 (with learning rate 2e-5: 0.9672). For the adapter training we set the learning rate to 1e-4. In all settings truncation and padding to the max input length of the model was used. In all settings, we trained for 6 epochs. We saved and tested the models after each epoch. The model from the epoch with the best macro F1 score on the development set was used for the evaluation on the devtest set.

4 Evaluation Results

In the development phase of the challenge, the various architectural options were evaluated on the devtest set. The results of the two subtasks are presented in the following sections. Subsequently, during the final test phase, the two architectures that demonstrated optimal performance on the devtest set, were tested again.

4.1 Subtask A: English Only Data

In Subtask A, the goal was to train a detector on English data only. We first tested different transformers on the **devtest set**, fine-tuning them on the English training set. We tested BERT (base, large) and RoBERTa base, see Table 1 for the results on the devtest set. For BERT base and RoBERTa base, the optimal results on the development set were obtained after 1 epoch of fine-tuning. Consequently, we also tested smaller steps, comprising less than 1 epoch, which yielded inferior outcomes. For

Adapter Fusion	Adapter Type	macro F1	micro F1
AI-Gen+mprc	semantic textual similarity	0.799	0.813
AI-Gen+ MNLI	**multi-genre NLI**	**0.851**	**0.852**
AI-Gen+anli	adversarial NLI	0.819	0.833
AI-Gen+cola	linguistic acceptability	0.836	0.841
AI-Gen+record	machine reading comprehension	0.786	0.809
AI-Gen+MNLI+cola	(combination)	0.779	0.779

Table 3: [Subtask A] English Adapter Fusion (test set: devtest)

BERT large, the optimal results were obtained after 4 epochs fine-tuning. RoBERTa base performed the best with a macro F1 score of 0.822.

Following this, we trained an adapter based on RoBERTa base using different adapter configurations (LoRA, LoReFT and seq_bn). See Table 2 for the results. For LoRA and LoReFT the best results on the dev set were calculated after 5 epochs, for seq_bn after 2 epochs. Using the configuration sequential bottleneck (seq_bn) the resulting adapter performed the best with a macro F1 score of 0.837 and even better than RoBERTa fine-tuned (macro F1 score: 0.822). We called this adapter AI-Gen.

After training our own adapter for AI-generated text detection (AI-Gen) we used adapter fusion for testing if additional knowledge of pre-trained task adapters improve the detection performance. See Table 3 for the results. We combined our adapter AI-Gen with five different task adapters. For all combinations, the best results were calculated after 3 epochs. The combination of AI-Gen with a task adapter on multi-genre NLI (MNLI) improved the macro F1-score on the devtest set to 0.851, from a macro F1-score of 0.837 using AI-Gen alone. We also tested a combination of AI-Gen with MNLI and the second best task adapter (cola), but this worsened the macro F1 score to 0.779.

Adapter fusion of AI-Gen with MNLI was our best detector on the English dataset and therefore also applied on the final **test set** in the challenge used for ranking. On the final test set we achieved a macro F1 score of 0.828 and micro F1 score of 0.8289, ranking third in the challenge (see Table 4). Furthermore, we evaluated the performance of this detector on the different generation methods used to build the test set. In Table 5 the generation methods with the most wrongly classified labels (>40%) are presented. Overall, our English detector has the most problems detecting fakes generated using GPT4 (55.36%), Dolly (54.48%) and StableLM (52.38%). When viewing the source of the test set samples, Mixset (41.08%) and CUDRT (29.01%) stood out with the most wrongly classified samples.

Team	macro F1	micro F1
1st	0.831	0.831
2nd	0.830	0.833
Fraunhofer SIT	**0.828**	**0.829**
4th	0.819	0.822

Table 4: [Subtask A] Final Evaluation on the test set (final ranking)

Generation method	# in testset	% wrong classified
ChatGLM	2006	41.28
Baichuan	1754	49.66
Dolly	268	54.48
StableLM	252	52.48
ChatGPT-turbo	144	45.83
GPT4	112	55.36
ChatGPT	96	45.83

Table 5: [Subtask A] Performance on the English test set by generation method (% wrong classied >40%)

4.2 Subtask B: Multilingual Data

Again, we first fine-tuned and tested different transformers using the multilingual training and **devtest set**, see Table 6. The XLM-RoBERTa large model (1 epoch) achieved a macro F1 score of 0. Viewing the score files, all samples were classified as human generated. As the multilingual training set contains 90.6% English data, we also applied the RoBERTa base model from Subtask A, trained on the English

Model	macro F1	micro F1
XLM-RoBERTa base	**0.630**	**0.847**
XLM-RoBERTa large	0	0.108
RoBERTa base	0.553	0.686

Table 6: [Subtask B] Multilingual Transformer fine-tuned (test set: devtest)

Adapter Setting	Adapter Type	macro F1	micro F1
newly trained adapter (ml-AI-Gen), configuration: seq_bn		0.585	0.837
fusion: ml-AI-Gen+form_class	formality classification (form_class)	0.521	0.824
fusion: ml-AI-Gen+mlki	multilingual knowledge integration	0.392	0.433
fusion: ml-AI-Gen+form_class+mlki		0.525	0.833

Table 7: [Subtask B] Multilingual Adapter Fusion (test set: devtest)

data, resulting in a macro F1 score of 0.553. The best score was achieved with XLM-RoBERTa base fine-tuned on the multilingual dataset (6 epochs) with a macro F1 score of 0.63.

Following this, we trained our own adapter on the multilingual training data (ml-AI-Gen) using the previous best configuration, being sequential bottleneck (4 epochs). See Table 7 for the results. The trained adapter (ml-AI-Gen) performed worse on the devtest set with a macro F1-score of 0.585, compared to XLM-RoBERTa base fine-tuned (0.630). Also, the use of adapter fusion with different task adapters worsened the results. On the multilingual data, XLM-RoBERTa fine-tuned being our best detector.

Interestingly, as one can see in Figure 1, the macro F1 score on the development set exhibited superior performance during training for the models using adapter fusion.The best one being adapter fusion with ml-AI-Gen and a multilingual knowledge integration task adapter (red in Figure 1) with a macro F1 score of 0.953 after 3 epochs of training. The best macro F1 score of XLM-RoBERTa on the development set was 0.946 after 6 epochs training (blue in Figure 1).

Again, we used the best performing model on the devtest to participate in the final evaluation on the **test set** used for ranking in the competition, here being XLM-RoBERTa base fine-tuned. We achieved a macro F1 score of 0.7258 and micro F1 score of 0.7361 on the test set, ranking tenth out of 25 teams. After the challenge, because of their superior performance during training, we also checked the performance of our trained adapter fusion models on the test set (see Table 8). Still, XLM-RoBERTa base (fine-tuned) performed best. Adapter fusion didn't improve the results on both multilingual test sets.

Model	macro F1	micro F1
XLM-RoBERTa base	**0.726**	**0.736**
(f) ml-AI-Gen+form_class	0.683	0.701
(f) ml-AI-Gen+mlki	0.519	0.554
(f) ml-AI-Gen+gyafc+mlki	0.555	0.562

Table 8: [Subtask B] Evaluation on the test set (used for final ranking, (f): fusion)

5 Conclusion and Future Work

In this paper, we presented the solutions developed by our team Fraunhofer SIT for the 2024 GenAI Detection Task 1 challenge. We fine-tuned transformers and adapters, and applied adapter fusion using different task adapters for knowledge transfer. On English data, adapter fusion improved the results, resulting in our team ranking third in subtask A of the challenge. The utilisation of multilingual data did not yield enhanced outcomes in the context of adapter fusion. One potential explanation for this phenomenon is the dearth of task adapters that are accessible within the domain of multilingual data. To illustrate, the most optimal task adapter within the English setting, multi-genre NLI, is not available for multilingual data.

Limitations

We acknowledge certain limitations of our work, and intend to address these in future work. First, we used the whole training set in Subtask B, containing 90% English data. Contrarily, the test set didn't include any English samples. In future work, this dataset should be more balanced out, incorporating more data from underrepresented languages. Furthermore, adapter fusion in a more wieder experimental setup should be tested in future work, utilising a greater number of models and datasets. Additionally, the relatively short length of the texts in this dataset was not taken into account. Previous approaches, such as multiscale positive-unlabeled training (Tian et al., 2023), have demonstrated effective results on similar texts.

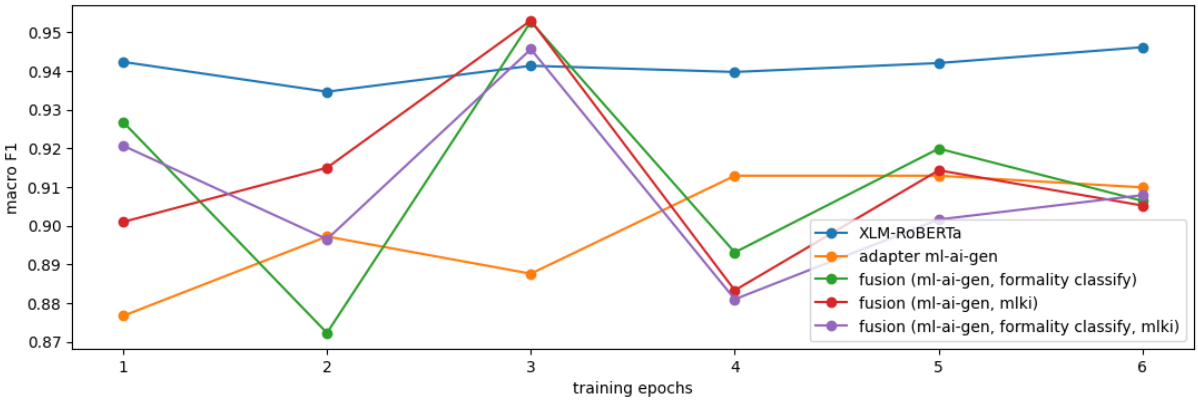

Figure 1: [Subtask B] Macro F1-Score on the dev set during training

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of ATHENE, CRISIS.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-DetectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343, Miami, Florida, USA. Association for Computational Linguistics.

Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2024. Overview of the "voight-kampff" generative ai authorship verification task at pan and eloquent 2024. In *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble, France 9 September 2024 through 12 September 2024*, volume 3740, pages 2486–2506. CEUR-WS.

José Campino. 2024. Unleashing the transformers: Nlp models detect ai writing in education. *Journal of Computers in Education*, pages 1–29.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Mingcan Guo, Zhongyuan Han, Haoyang Chen, and Jiangao Peng. 2024. A machine-generated text detection model based on text multi-feature fusion. *Working Notes of CLEF*.

Yifan Hou, Wenxiang Jiao, Meizhen Liu, Carl Allen, Zhaopeng Tu, and Mrinmaya Sachan. 2022. Adapters for enhanced modeling of multilingual knowledge and text. *arXiv preprint arXiv:2210.13617*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021.

Adapterfusion: Non-destructive task composition for transfer learning. *ArXiv*, abs/2005.00247.

Jonas Pfeiffer, Andreas Rücklé, Clifton A. Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *ArXiv*, abs/2007.07779.

Clifton Poth, Jonas Pfeiffer, Andreas R"uckl'e, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. *arXiv preprint arXiv:2305.18149*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Daniel Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *ArXiv*, abs/2404.03592.