Can You Really Trust That Review? ProtoFewRoBERTa and DetectAIRev: A Prototypical Few-Shot Method and Multi-Domain Benchmark for Detecting AI-Generated Reviews

Shifali Agrahari[1], Sujit Kumar[2], Sanasam Ranbir Singh[1]
[1]Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati, India
[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
{a.shifali, ranbir}@iitg.ac.in, kumar.sujit474@gmail.com

Abstract

Synthetic reviews mislead users and erode trust in online marketplaces, and the advent of Large Language Models (LLMs) makes detecting such AI-generated content increasingly challenging due to their human-like fluency and coherence. In the literature, LLM-generated review detection datasets are limited to one or a few domains, with reviews generated by only a few LLMs. Consequently, datasets are limited in diversity in terms of both domain coverage and review generation styles. Models trained on such datasets generalize poorly, lacking cross-model adaptation and struggling to detect diverse LLM-generated reviews in real-world, open-domain scenarios. To address this, we introduce **DetectAIRev**, a benchmark dataset for AI-generated review detection that includes human-written reviews from diverse domains and AI-generated reviews generated by various categories of LLMs. We evaluate the quality and reliability of the proposed dataset through several ablation studies and human evaluations. Furthermore, we propose an AI-generated text detection method **ProtoFewRoBERTa**, a few-shot framework that combines prototypical networks with RoBERTa embeddings, which learn discriminative features across multiple LLMs and human-written text using only a few labeled examples per class to discriminate between LLMs as the author for text author detection. We conduct our experiments on the DetectAIRev and a publicly available benchmark dataset. Our experimental results suggest that our proposed methods outperform the state-of-the-art baseline models in detecting AI-generated reviews and text detection. [1]

1 Introduction

A synthetic review is an artificially generated review created by automated systems or algorithms rather than real consumer experiences, posing risks

Figure 1: Comparison of Human-Written and LLM-Generated Reviews. The human-written review is personal and experience-driven, while the LLM-generated review is structured and feature-focused, emphasising product specifications over subjective experience.

such as distorted analytics, misrepresented opinions, erosion of trust in online platforms, and raising ethical concerns over manipulated public opinion and unfair market advantages (Liu, 2012; Kumar et al., 2025; Agrahari et al., 2025b; Kumar et al., 2024). Mitigating synthetic reviews is essential to uphold the credibility, reliability, and authenticity of digital feedback mechanisms, ensure transparency across online shopping and digital marketplaces, safeguard consumer trust in product evaluations, and promote a fair and competitive market environment (He et al., 2022; George et al., 1970; Yao et al., 2017; Sharma and Kumar, 2023; Filieri, 2016; Kaabachi et al., 2017).

In the literature, research on LLM-generated text detection has primarily focused on Wikipedia articles (Guo et al., 2023), academic essays (Peng et al., 2023), and headlines or news content (Wang et al., 2023b). However, reviews are comprehensively different from essays or Wikipedia articles due to their *subjectivity*, *brevity*, and focus on personal opinion, feedback and *sentiments*. Unlike essays, which present structured arguments and in-depth analysis, reviews are concise and often centered on specific aspects like battery life, performance, or customer service. Common words like "good", "excellent", or "disappointing" along with phrases reflecting personal experiences such as "I liked" or "I didn't enjoy" are frequently used in reviews, making them emotionally charged and evaluative. This informal, opinion-driven nature

[1]The dataset and code repository are publicly available at: https://huggingface.co/datasets/Sifi-world/DetectAIRev.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2118–2140

contrasts with the more **objective**, **fact-based** tone of essays or news articles. Furthermore, reviews are often highly context-specific, focusing on particular products or services, whereas essays and articles cover broader topics. The presence of sentiment-laden language and context-dependent evaluations presents a unique challenge in detecting Large Language Model (LLM) generated reviews, as it requires identifying these personal, subjective elements that are often absent in more formal, structured texts.

In existing literature, datasets for detecting LLM-generated reviews are limited to one or a few domains, primarily focusing on Amazon product and hotel reviews, while numerous other domains remain unexplored. For instance, (Salminen et al., 2022) curated a dataset focusing solely on Amazon product reviews, using GPT-2 and ULMFit models. Similarly, the MAiDE-up dataset (Buscaldi and Liyanage, 2024) is limited to hotel reviews, with reviews generated using GPT-3 through a few predefined prompt instructions. However, these datasets are limited in size and scope, focused on one or a few domains, with reviews generated by only a few LLMs (typically GPT-based), and limited in diversity in terms of both domain coverage and review generation styles. As a result, models trained on such datasets may generalize poorly, performing well in their training domains but struggling to detect LLM-generated reviews from unseen domains or different LLMs (Bhattacharjee et al., 2024). This undermines their robustness in real-world, open-domain scenarios, as the lack of cross-model adaptation limits their ability to detect reviews generated by LLMs and adapt to diversity in text generation patterns (Hua and Yao, 2024).

To address the limitations of existing datasets for LLM-generated review detection, we propose the **DetectAIRev** dataset, comprising reviews generated by seven different LLMs and human-written counterparts across five diverse domains. Reviews are generated using four different prompting strategies: (i) direct prompting, (ii) few-shot prompting, (iii) review replication and (iv) facet-aware prompting, resulting in a large-scale dataset that captures a broad range of generative models, facets, and linguistic variations. Diversities within our dataset enable detection systems to learn from varying text generation patterns across domains, enhance the adaptability of detection systems, and provide a comprehensive foundation for developing generalizable detectors capable of identifying AI-generated reviews in real-world settings.

Initial AI-generated text detection methods, such as watermarking-based techniques (Fu et al., 2024; Jiang et al., 2024) and statistical detection methods (Abri et al., 2020; Gehrmann et al., 2019; Su et al., 2023), were limited in performance due to their dependence on manual feature engineering and also struggled to adapt quickly to newly emerging LLMs models, reducing their long-term effectiveness. Further, recent training- and finetuning-based approaches (Zellers et al., 2019; Liu et al., 2023; Chakraborty et al., 2024) have achieved notable performance gains; however, training- and finetuning-based approaches require carefully paired training data and often fail to generalize well in Out-Of-Distribution (OOD) detection settings, largely because of their fixed binary classification framework. Unlike traditional binary detectors, the study (Guo et al., 2024) transforms AI-generated text detection as a writing style differentiation problem, treating each LLM as a distinct *Author* with a consistent stylistic signature. However, study (Guo et al., 2024) relies heavily on paired, model-specific training data and involves complex multi-objective optimization, making it resource-intensive and difficult to adapt to new LLMs. Its few-shot adaptation remains limited, and OOD generalization degrades when training data lacks stylistic diversity. To address these challenges, we frame AI-generated review detection as a writing style differentiation problem, where each LLM is treated as a distinct author with a consistent stylistic signature. Instead of traditional binary classification, our goal is to learn discriminative features across multiple LLMs and human-written text using only a few labeled examples per class. We propose *ProtoFewRoBERTa*, a few-shot framework that combines Prototypical Networks (Snell et al., 2017) with RoBERTa embeddings. In each episode, class prototypes are computed from a small support set, and query samples are classified based on their distance to these prototypes. Trained with a negative log-likelihood loss, the model efficiently distinguishes writing styles with minimal supervision and training, offering a scalable and adaptable detection approach. To comprehensively understand the patterns and challenges in detecting LLM-generated reviews, we investigate the following research question: What are the key differences between human-written and LLM-generated reviews in terms of lexical, readability, sentiment, and psycholinguistic features, and how do these

differences affect detection effectiveness?. We conduct extensive evaluations and ablation studies to assess the quality of the proposed dataset and the performance of our model. Results show that the dataset enables reliable AI-generated review detection, the proposed model consistently outperforms baseline models on both this dataset and existing benchmarks.

2 Related Work

In the literature, Several benchmark datasets exist across different domains, such as Wiki and QnA (Guo et al., 2023), ChatGPT-written abstracts (Yu et al., 2025), Applied Statistics (Salim and Hossain, 2024), M4 (Wang et al., 2023a), and Turing Bench (Uchendu et al., 2021) as detail mention in Appendix Subsec. A.3, Table 12. However, when it comes to LLM-generated review datasets, existing efforts are limited. Notably, (Salminen et al., 2022) focuses only on Amazon product reviews, using GPT-2 and ULMFit to fine-tune review generation. Likewise, the MAiDE-up dataset (Buscaldi and Liyanage, 2024) is restricted to hotel reviews, generated by GPT-3. Despite these contributions, existing datasets suffer from limited domain diversity, covering only one or two review domains. This restricts the generalization of detection models, as they struggle to adapt to different review contexts and varying LLM-generated styles. Over the last few years, numerous approaches have been proposed to tackle the task of LLM-generated text detection. Detecting machine-generated text is primarily formulated as a binary classification task (Zellers et al., 2019; Gehrmann et al., 2019; Ippolito et al., 2019) naively distinguishing between human-written and LLM-generated text. In general, there are three main approaches: the supervised methods (Agrahari et al., 2025a; Zellers et al., 2019; Zhong et al., 2020; Liu et al., 2023, 2022; Agrahari and Singh, 2025), the unsupervised ones, such as zero-shot methods (Wang et al., 2023b), and Adversarial measures on detection accuracy (Susnjak and McIntosh, 2024; Agrahari et al., 2025b), (Liang et al., 2023) especially within the education domain. (Krishna et al., 2024) train a generative model (DIPPER) to evade detection. However, most supervised methods for detecting AI-generated text require large amounts of labeled data, are computationally intensive, and often overfit to the seen training distributions, limiting their ability to generalize to new LLMs or domains. In contrast, we propose *ProtoFewRoBERTa*, a few-shot learning framework that offers a more efficient and flexible paradigm by leveraging only a few labeled examples per class. This significantly reduces training time while enhancing adaptability and generalization across diverse generative models in real-world, open-domain scenarios.

3 Proposed Dataset

This study proposes DetectAIRev, an AI-generated review detection dataset curated from human-written and LLM-generated reviews across diverse domains. We have considered human written review from several openly available review $\mathcal{R}_{\mathcal{H}}$ dataset from different domains, including **Book Reviews**[2], (McAuley et al., 2015; He and McAuley, 2016; Wan and McAuley, 2018), **E-Commerce Review**[3] (Agarap, 2018b,a), **Movie Reviews**[4] (Maas et al., 2011), **Trip Advisor Hotel Reviews**[5] (Alam et al., 2016) and **Restaurant Review** (Abri et al., 2020). Incorporating human-written reviews from multiple domains enhances robustness and generalization by capturing diverse linguistic styles, sentiment patterns, and writing structures. For example, book and movie reviews are more narrative, while hotel and restaurant reviews emphasize service and amenities, enabling the model to capture diverse linguistic styles and adapt to varied real-world scenarios. Given a human-written actual review $\mathcal{R}_H$, the objective of LLMs is to generate a corresponding review using various prompting strategies. This study considers four prompting strategies to maintain diversity in LLM-generated reviews. (i) **Zero-Shot Prompting (Direct)**: The LLM is directly prompted with an instruction (e.g., *Write a review about the product*) without any additional examples. This approach relies solely on the pretrained model to generate relevant reviews. (ii) **Few-Shot Prompting**: These examples help guide the model in generating coherent and contextually relevant reviews. The input format is defined as: $\mathcal{R}_{\mathcal{L}} = LLM(\mathcal{R}_{\mathcal{H}} \cup \mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_n)$. Where $\mathcal{R}_{\mathcal{L}}$ is the LLM-generated review, $\mathcal{E}$ denotes a few-shot example, $\cup$ denotes concatenation, and n represents the number of examples provided as few-shots. Appendix Sec. G.2 provides further details on the prompt design used in our Few-shot Prompting strategy. (iii) **Review Replication**: In

[2] Books Reviews Dataset
[3] E-Commerce Dataset
[4] IMDb Movie Reviews Repository
[5] Trip Advisor Hotel Reviews Source Repository

this method, the LLM receives a human-written review $\mathcal{R}_{\mathcal{H}}$ as input and is prompted to generate a similar review. This helps assess the ability of the model to mimic human writing styles directly: $\mathcal{R}_{\mathcal{L}} = LLM(\mathcal{R}_{\mathcal{H}})$. (iv) **Facet-aware Prompting**: This approach generates LLMs that are prompted to generate reviews based on facets of human-written reviews $\mathcal{R}_{\mathcal{H}}$. We design prompts that explicitly ask the LLM to emphasis certain aspects (e.g., *food quality*, *ambiance*, or *service in restaurant reviews*), and these aspects are extracted from a facet analysis of human-written reviews. Appendix Sec. G.3 presents the details of the facet-based prompting strategy used to generate reviews while maintaining diversity across different aspects. Facet-aware review generation aims to ensure alignment between facet emphasis in human and LLM-generated reviews. We consider seven LLMs to generate reviews using LLMs including **GPT-3** (Radford et al., 2019), **Llama 3** (Touvron et al., 2023), **Gemma** (Team et al., 2024), **Mistral** (Jiang et al., 2023), **Phi-3** (Li et al., 2023), **Qwen** (Bai et al., 2023), and **DeepSeek** (Bi et al., 2024) in this study. Table 10 provides the characteristics, versions, full descriptions and specifications of the LLMs used in this study to generate reviews. To simulate a real-world setting, we also incorporate three types of adversarial attacks: (i) alternative spelling, (ii) paraphrasing (rewrite) attacks, and (iii) misspelling, as described in (Dugan et al., 2024), with a detailed mention in Appendix Subsec. G.1. The primary motivation behind incorporating diverse domains and diverse LLMs is that each LLM exhibits unique writing and text generation styles, ensuring that the dataset captures a wide range of linguistic patterns and review contexts. This diversity in our proposed dataset, both in terms of LLMs and domains, aims to enhance the ability of the AI-generated review detection model to effectively distinguish between human-written and LLM-generated content across various real-world scenarios. Our proposed dataset, **DetectAIRev**, contains English reviews with the following columns: review text, generative model name (author name), label (LLM or human), and domain. Table 1 presents the distribution of our proposed dataset across domains and generative models.

3.1 Quality Assessment of Proposed Datasets

We perform two different types of quality and reliability evaluation: (i) We conduct a *Stylometric Feature* analysis to study the differences between human-written text and LLM-generated text in terms of lexical features, readability, sentiment, psycholinguistics, and text similarity. (ii) Human annotation and evaluation of reviews generated by LLMs are conducted to validate the quality and reliability of the DetectAIRev dataset.

Label	E-commerce	Hotel	Movie	Tourist	Restaurant
Human	19,998	10,000	10,000	7,102	8,902
LLM	53,349	25,948	27,994	26,849	45,321

Table 1: Domain-wise label distribution in the **DetectAIRev** dataset.

Metric	Human	Phi	LLaMA	DeepSeek	Mistral	Qwen	ChatGPT	Gemma
# of Words	3,548	3,400	4,327	1,978	4,215	3,846	4,102	3,213
# of Sentences	240	227	227	134	245	238	250	199
Avg. Sentence Len	14.78	14.98	19.06	14.76	17.20	16.20	16.40	10.23
# of Unique Words	1,050	914	1,437	661	1,264	1,127	1,198	689
Lexical Diversity	0.296	0.269	0.332	0.334	0.300	0.293	0.292	0.338
# of Characters	17,250	16,041	20,256	9,442	19,834	18,240	19,560	8,442
# of Punctuation	580	518	742	822	690	630	680	476

Table 2: Comparison of human-written and LLM-generated reviews based on lexical features.

3.1.1 Stylometric Feature Evaluation

We consider multiple features to compare human and LLM-generated text, including lexical, readability, sentiment, psycholinguistic, and text similarity features. We extract and calculate the textual features of our proposed dataset by following the procedure reported in previous studies (Lagutina et al., 2019; Mindner et al., 2023; Agrahari et al., 2024), to understand the lexical diversity between reviews generated by LLMs and human-written reviews. Table 2 presents key metrics, including the average number of sentences per review, quotations, and unique words per review, along with additional indicators such as the frequency of special characters and personal pronouns, which help identify conversational tones. The lexical analysis in Table 2 highlights key differences between human-written and LLM-generated reviews. From Table 2 it is evident that LLaMA produces the longest reviews with the most complex sentence structures. In contrast, DeepSeek generates the most concise outputs with higher lexical diversity and highest punctuation usage, indicating a structured writing style. Similarly, Table 2 shows that human-written reviews are short in terms of an average number of sentences compared to LLMS-generated reviews;

Table 3 compares human-written reviews and LLM-generated reviews by diverse categories of LLMs in terms of readability metrics, which assess how easy or difficult a text is to understand. Flesch-Kincaid(FK) Grade and Automated Readability Index(ARI) estimate the required school

Metric	Human	Phi	LLaMA	Qwen	Gemma	Mistral	DeepSeek	GPT
Flesch Ease	68.6	65.62	70.94	70.23	82.44	74.90	73.17	76.50
FK Grade	8.5	7.6	7.6	7.9	5.3	6.1	6.8	6.5
Gunning Fog	8.0	5.76	7.03	7.41	5.94	5.44	6.19	6.7
Dale-Chall	1.26	0.88	1.09	1.26	1.09	0.86	1.02	1.12
SMOG	11.0	10.9	10.8	10.7	8.9	9.5	9.9	9.2
ARI	11.2	8.8	9.4	9.9	6.1	7.3	8.1	7.8
Yule's K	79.77	192.65	145.35	209.24	130.02	158.33	182.72	160.5

Table 3: Comparison of readability and vocabulary metrics between human-written and LLM-generated reviews (details in Appendix B and (Mindner et al., 2023)).

Sentiment	LLaMA	GPT-3	Mistral	Human	Gemma	Phi	Qwen	DeepSeek
Positive	40.2%	50.3%	37.5%	**57.1%**	42.8%	35.6%	43.4%	47.3%
Neutral	45.1%	38.2%	47.0%	35.8%	43.5%	48.0%	35.4%	42.5%
Negative	14.7%	11.5%	15.5%	**9.1%**	13.7%	16.4%	**8.4%**	11.3%

Table 4: Sentiment feature analysis based on the ratio of positive, neutral, and negative reviews in human-written and LLM-generated reviews across different models.

grade level to understand the text. Gunning Fog and SMOG focus on complex words to determine readability. Dale-Chall considers familiar words for clarity. Yule's K measures vocabulary richness. Together,these metrics help evaluate the level of text complexity to read and understand text written by human and LLM-generated reviews. Intreprebility of range of these matrices mention in Appendix Sec. B, Table 13.

The readability scores presented in Tabl 3 demonstrate that LLM-generated reviews are generally more formal, readable, simpler, and less complex than human-written reviews. This observation is primarily based on the higher Flesch Ease scores of LLM-generated content, particularly from models like Gemma and GPT 3, compared to human-written reviews. These higher scores indicate that LLM-generated texts are easier to read and understand. Additionally, the lower FK Grade and SMOG scores of LLM-generated reviews, especially those generated by Gemma, indicate that these texts are significantly more straightforward and more accessible. In contrast, human-written texts tend to be more sophisticated, requiring a higher educational level for comprehension. Furthermore, human-written reviews exhibit greater lexical diversity and complexity, as indicated by the lower Yule's K score compared to most LLM-generated reviews, especially those generated by Qwen and DeepSeek, which have higher Yule's K scores. This difference highlights that human-authored texts contain a richer and more varied vocabulary. The Gunning Fog index further supports this, as human-written texts show higher scores, reflecting the use of more complex words and sentence structures compared to LLM-generated content, which tends to have lower scores. Such observations from Table 3 suggest

Feature	Human	Phi	LLaMA	DeepSeek	Mistral	Qwen	ChatGPT	Gemma
Cognitive Words	12.2	10.9	12.5	12.3	11.8	9.6	**13.5**	13.0
Emotional Words	9.5	9.4	8.7	**9.6**	9.1	6.7	8.8	8.2

Table 5: Psycholinguistic feature analysis (LIWC) (Pennebaker et al., 2001) of human-written and LLM-generated reviews.

that while LLM-generated reviews might appear more straightforward and readable due to their higher readability scores and lower complexity measures, the simplicity and reduced lexical diversity make them more distinguishable from human-authored content. This distinction is especially apparent when examining Dale-Chall scores, where LLM-generated texts, particularly from Mistral, tend to score lower, indicating simpler and more familiar word usage. In contrast, human-written reviews score higher, reflecting the use of less common vocabulary. The reduced lexical richness and simplicity of LLM-generated reviews could serve as potential features for detecting AI-generated content. These differences have significant implications for the detection of AI-generated reviews, as they highlight specific linguistic features that can help differentiate between human and LLM-generated texts.

Sentiment regarding specific aspects of reviews is a key indicator of human satisfaction. Therefore, comparing LLM-generated and human-written reviews in terms of sentiment helps evaluate whether LLMs can write reviews similar to human written review sentiment levels related to products. Table 4 presents the comparison between reviews generated by LLMs and human-written reviews in terms of sentiment distribution. From Table 4, it is apparent that human reviews exhibit the highest percentage of positive sentiment, suggesting a more favorable tone in human-written reviews. GPT-3 also produces a high proportion of positive reviews. In contrast, Phi and Mistral generate the lowest positive sentiment scores, with higher neutral or negative tendencies, reflecting a less pronounced emotional tone in the review generated by Phi and Mistral. Negative sentiment is relatively low across all sources, with human reviews showing the least negative sentiments, while Phi exhibits the highest negative sentiment proportion. From such observation, we conclude that there are significant differences in sentiment tendencies between AI-generated and human-written reviews.

Table 5 presents the psycholinguistic feature capture how language reflects cognitive and emotional states. Tools like LIWC (Linguistic Inquiry and

Agreement Metric	Fluency	Origin	Coherence	Factuality
Krippendorff's Alpha (α)(Krippendorff, 2011)	0.769	0.671	0.535	0.776
Fleiss' Kappa (κ) (Fleiss, 1971)	0.719	0.573	0.512	0.563

Table 6: Krippendorff's Alpha and Fleiss' Kappa scores for the annotation scheme used in human evaluation across four metrics, as described in Table 11.

Word Count) (Pennebaker et al., 2001) analyze word categories, such as pronoun usage, cognitive processes (e.g., "think", "know"), and emotional expressions analysis across different models and human-written texts. It highlights the differences in cognitive and emotional language. Notably, GPT-3 and Gemma exhibit higher use of cognitive words, suggesting a more analytical tone. Emotional word usage varies, with human reviews demonstrating the highest proportion, reinforcing their natural expressiveness. Conversely, Qwen shows the least emotional engagement, potentially indicating a more.

Metric	Human	Phi	LLaMA	DeepSeek	Mistral	Qwen	ChatGPT	GPT	Gemma
Coherence S	0.876	0.848	**0.891**	0.686	0.872	0.860	0.870	0.888	0.787
Anger	21	24	**47**	20	30	26	28	27	14
Anticipation	110	100	**129**	62	120	115	118	119	61
Disgust	5	3	**26**	13	10	8	9	7	6
Fear	18	14	**46**	16	20	17	19	21	17
Joy	160	151	**192**	72	170	165	168	175	67
Sadness	20	17	**64**	**111**	30	22	25	23	20
Surprise	40	36	**75**	43	45	42	44	48	24
Trust	170	162	**192**	133	175	168	172	178	78

Table 7: Psychological and cognitive factor analysis in human-written and LLM-generated reviews across models using NRC Emotion Lexicon (Mohammad and Turney, 2013).

Table 7 highlights key psychological and cognitive aspects of text generated by different models compared to human-written content. Notably, Llama and GPT-3 exhibit the highest coherence scores, suggesting strong logical flow in their text. Emotional expressions vary significantly across models Llama and Mistral show higher levels of anger and sadness, whereas GPT-3 and Gemma use more joyful and trust-related language. DeepSeek indicating a weaker structural flow. Although human generated review are avg in all the matrics. These differences in sentiment and coherence suggest variations in how each model balances factual accuracy with emotional tone in review generation. The Bilingual Evaluation Understudy (BLEU) Score (Papineni et al., 2002) and the Metric for Evaluation of Translation with Explicit Ordering (METEOR) Score (Banerjee and Lavie, 2005) are used to analyze the relationship between human-written reviews and those generated by LLMs. Appendix Figure 2 presents that GPT-3 review are indicating strong similarity to human-written text in term of lexical overall. DeepSeek and Qwen also achieve relatively high scores, while Phi and Mistral demonstrate the lowest alignment with human text. These results highlight differences in text fluency and word choice precision across models.

3.1.2 Human Evaluation of Dataset

We conducted human annotation and evaluation of the proposed datasets to assess their quality and reliability. We created a subset of the datasets consisting of eight hundred samples by randomly selecting one hundred reviews generated by each LLM and assigning them to four independent annotators. Each annotator was asked to evaluate the reviews based on Fluency, Origin, Coherence, and Factuality, following the scoring criteria defined in Table 11. Fluency evaluates the grammatical correctness and readability of a review, while Origin distinguishes whether it is human-written or LLM-generated. Coherence measures the logical flow and consistency of ideas, and Factuality assesses the accuracy and truthfulness of the review content. To ensure high-quality annotations, we selected annotators with a computer science and linguistics research background, including research scholars. Next, we measured annotator agreement on the scores assigned for Fluency, Origin, Coherence, and Factuality using Krippendorff's alpha (α) (Krippendorff, 2011) and Fleiss' kappa (κ) (Fleiss, 1971). Table 6 presents Krippendorff's alpha and Fleiss' kappa scores for inter-annotator agreement on our proposed datasets. Fluency had the highest agreement, indicating the grammatical correctness and readability of LLM-generated reviews. For origin, the inter-annotator agreement (Krippendorff's = 0.671; Fleiss' = 0.573) indicates moderate agreement among annotators regarding the origin of a review, i.e., whether it is generated by an LLM or written by a human. It highlights the challenge of distinguishing LLM-generated text from human-written reviews. Coherence showed moderate agreement, because LLM-generated reviews, while syntactically fluent, often lack deeper discourse-level consistency, leading to divergent interpretations among annotators. Factuality had

the lowest agreement, reflecting the difficulty of verifying factual claims. which is expected given the categorical nature and inherent subjectivity of factuality judgments. These results suggest that distinguishing AI-generated content remains challenging, emphasizing the need for refined evaluation criteria or automated assessment methods.

4 Proposed Method

This study proposes **ProtoFewRoBERTa**, A Prototypical Network-based Few-shot Learning Framework Leveraging RoBERTa (Liu, 2019) representations for AI-generated text detection. Our proposed ProtoFewRoBERTa model, inspired by Prototypical Networks (Snell et al., 2017), reformulates the approach for multi-class classification of human- and LLM-generated text using RoBERTa representations. The proposed ProtoFewRoBERTa model constructs prototypical representations for each author (human and LLMs) using few-shot example texts written or generated by the author (human and LLMs). Given a dataset $\mathcal{A}$ consisting of text samples $\mathcal{X}_i$ and their corresponding author labels $\mathcal{Y}_i$, i.e., $\mathcal{A} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, where each $Y_i \in \{1, \ldots, k\}$, we first select the Support and Query set. In each training episode, we consider all k classes from the dataset $\mathcal{A}$ and for each class $c \in \{1, \ldots, K\}$, we randomly sample p texts from $\mathcal{A}$ to form the support set and support set as defined by by the equations 1 below.

$$\mathcal{S} = \bigcup_{c=1}^{k} \{(\mathcal{X}_i^c, \mathcal{Y}_i^c)\}_{i=1}^{p} \tag{1}$$

Similarly, we consider all K classes from the dataset $\mathcal{A}$ and for each class $c \in \{1, \ldots, k\}$, we randomly sample q texts from $\mathcal{A}$ to form the query set as defined by the equations 2 below.

$$\mathcal{Q} = \bigcup_{c=1}^{K} \{(\mathcal{X}_j^c, \mathcal{Y}_j^c)\}_{j=1}^{q} \tag{2}$$

Text samples for the support and query sets are selected from each class by ensuring the two sets are disjoint, i.e., $\mathcal{S} \cap \mathcal{Q} = \emptyset$. Subsequently, for each class c in the support set $\mathcal{S}$, we compute the class representation vector (prototype) $\mathbf{a}_c$ by averaging the encoded representations of all samples belonging to class c in the support set $\mathcal{S}$, as defined:

$$\mathbf{a}_c = \frac{1}{p} \sum_{(\mathcal{X}_i, \mathcal{Y}_i) \in \mathcal{S}_c} g_{\theta_p}\big(f_{\theta_e}(\mathcal{X}_i)\big), \tag{3}$$

where $f_{\theta_e}(\cdot)$ is the RoBERTa encoder producing contextualized text representations, $g_{\theta_p}(\cdot)$ is the projection function that maps these representations into the prototypical embedding space, and $\mathcal{S}_c$ refers to the subset of support examples for class c. We use RoBERTa (Liu, 2019) as the text encoder. Given an input text $\mathcal{X}_i$, the encoder $f_{\theta_e}(\cdot)$ produces a contextual representation $\mathbf{x}_i$ using RoBERTa, and the projection network $g_{\theta_p}(\cdot)$ maps it to the metric space, resulting in the final embedding $\mathbf{x}'_i = g_{\theta_p}(f_{\theta_e}(\mathcal{X}_i))$. Intuitively, the class representation vector (prototype) $\mathbf{a}_c$ can be viewed as capturing the signature or writing style of the author (either an LLM or a human), as it is estimated by averaging the encoded representations of a few examples belonging to class c. Next, we estimate the distance d_{jk} between each query encoding and

Algorithm 1. ProtoFewRoBERTa: Episodic Training and Inference via Prototypical Networks for AI-Generated Review and Text Detection

Input: Dataset $\mathcal{A} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ with K classes; p: samples per class in the support set; q: samples per class in the query set; f_{θ_e}: RoBERTa encoder; g_{θ_p}: projection network mapping encoded text to embedding space.
Objective: Learn parameters $\Theta = \{\theta_e, \theta_p\}$ that minimize episodic classification loss $\mathcal{L}$.

— Training Phase —
for each training episode **do**
  Initialize support $\mathcal{S} \leftarrow \emptyset$, query $\mathcal{Q} \leftarrow \emptyset$
  for $c = 1$ to K **do** ▷ Construct class-specific support and query sets
    $\mathcal{S}_c \leftarrow \textsc{RandomSample}(\mathcal{A}_c, p)$; $\mathcal{Q}_c \leftarrow \textsc{RandomSample}(\mathcal{A}_c \setminus \mathcal{S}_c, q)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_c$; $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{Q}_c$
  end for
  for $c = 1$ to K **do** ▷ Compute class prototypes
    $\mathbf{a}_c \leftarrow \frac{1}{p} \sum_{(\mathcal{X}_i, \mathcal{Y}_i) \in \mathcal{S}_c} g_{\theta_p}(f_{\theta_e}(\mathcal{X}_i))$
  end for
  Initialize $\mathcal{L} \leftarrow 0$
  for each $(x_j, y_j) \in \mathcal{Q}$ **do** ▷ Compute episodic loss
    for $c = 1$ to K **do**
      $d_{jc} \leftarrow \|g_{\theta_p}(f_{\theta_e}(x_j)) - \mathbf{a}_c\|_2^2$
    end for
    $p(y = c \mid x_j) \leftarrow \dfrac{\exp(-d_{jc})}{\sum_{k'=1}^{K} \exp(-d_{jk'})}$
    $\mathcal{L} \leftarrow \mathcal{L} - \log p(y_j \mid x_j)$
  end for
  Update $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$ using AdamW optimizer
end for
— Inference Phase —
Input: Unseen support set $\mathcal{S}' = \{\mathcal{S}'_1, \ldots, \mathcal{S}'_K\}$ with few labeled examples per class
for $c = 1$ to K **do**
  $\mathbf{a}'_c \leftarrow \frac{1}{|\mathcal{S}'_c|} \sum_{(\mathcal{X}_i, \mathcal{Y}_i) \in \mathcal{S}'_c} g_{\theta_p}(f_{\theta_e}(\mathcal{X}_i))$
end for
for each unseen query x_q **do**
  for $c = 1$ to K **do**
    $d_{qc} \leftarrow \|g_{\theta_p}(f_{\theta_e}(x_q)) - \mathbf{a}'_c\|_2^2$
  end for
  Assign $\hat{y}_q \leftarrow \arg\min_c(d_{qc})$ ▷ nearest-prototype classification
end for

Detector	DetectAIRev			SemEval			Product Review			Restaurant Review			Hotel Review			DeepFake		
	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI
GTLR	0.477	0.438	0.455	0.501	0.460	0.479	0.496	0.331	0.596	0.486	0.348	0.575	0.491	0.319	0.592	0.395	0.368	0.420
OpenAI-det	0.641	0.598	0.621	0.672	0.582	0.601	0.536	0.255	0.663	0.486	0.344	0.650	0.605	0.389	0.707	0.340	0.265	0.401
DetectGPT	0.604	0.562	0.571	0.646	0.590	0.613	0.632	0.324	0.683	0.564	0.434	0.635	0.536	0.345	0.734	0.389	0.324	0.535
Radar	0.737	0.629	0.661	0.728	0.612	0.644	0.605	0.352	0.716	0.660	0.493	0.744	0.558	0.691	0.220	0.475	0.193	0.610
Binoculars	0.892	0.882	0.890	0.872	0.869	0.860	0.506	0.453	0.672	0.504	0.403	0.670	0.500	0.333	0.660	0.650	0.534	0.706
DeTeCtive	0.905	0.919	0.898	0.918	0.910	0.899	0.875	0.880	0.923	0.919	0.834	0.912	0.893	0.812	0.923	0.945	0.903	0.934
RoBERTa	0.793	0.701	0.716	0.781	0.698	0.711	0.827	0.654	0.723	0.756	0.743	0.742	0.758	0.767	0.870	0.548	0.540	0.560
RoBERTa + Stylometric	0.828	0.762	0.770	0.794	0.722	0.732	0.973	0.970	0.970	0.780	0.790	0.840	0.881	0.729	0.834	0.933	0.870	0.950
ProtoFewRoBERTa (Feat)	0.972	**0.970**	0.971	**0.967**	**0.951**	0.947	0.976	0.915	**0.952**	0.956	**0.917**	0.955	0.928	0.891	0.928	0.932	0.913	0.954
ProtoFewRoBERTa	**0.983**	0.970	**0.981**	0.961	0.948	**0.957**	**0.982**	**0.922**	0.952	**0.966**	0.865	**0.965**	**0.959**	**0.932**	**0.962**	**0.980**	**0.972**	**0.984**

Table 8: Performance (Accuracy, F1-Human, F1-AI) of baseline methods and proposed method (ProtoFewRoBERTa) on proposed dataset (DetectAIRev) and other existing datasets.

each class representation vector (prototype) $\mathbf{a}_c$ by the equation defined below.

$$d_{jk} = \left\| f_\theta(X_j^{(Q)}) - \mathbf{a}_k \right\|_2^2 \tag{4}$$

where $X_j^{(Q)}$ denotes the j^{th} text in the query set Q, and $f_\theta(\cdot)$ represents the encoder defined earlier in Equation 3. Subsequently, we estimate the probability of a query instance X_j belonging to class k by applying a softmax function over the negative distances d_{jk} as defined below.

$$p(y = k \mid X_j^{(Q)}) = \frac{\exp(-d_{jk})}{\sum_{k'} \exp(-d_{jk'})} \tag{5}$$

Next, we estimate the loss for each episode by applying the negative log-likelihood loss over the query set $\mathcal{Q}$. The episodic loss function as follows:

$$\mathcal{L} = -\sum_{(x_j, y_j) \in \mathcal{Q}} \log p(y_j \mid x_j) \tag{6}$$

The primary objective of the loss function $\mathcal{L}$ is to encourage the embedding of each query sample $X_j^{(Q)}$ to be close to the class representation vector (prototype) $\mathbf{a}_c$ of its corresponding class, while simultaneously pushing it farther away from the class representation vector (prototype) $\mathbf{a}_c$ of other classes. Our proposed method, ProtoFewRoBERTa, minimises the negative log-likelihood loss function $\mathcal{L}$ to learn the parameters of the projection layer and fine-tune the parameters of the RoBERTa model used for text encoding.

5 Experimental Results and discussion

5.1 Experimental Setup

We study the performance of the proposed and baseline models on **DetectAIRev** dataset and the SemEval-2024 Task 8 Monolingual dataset (Wang et al., 2024). Appendix Subsec. E and Table 16 present the descriptions and characteristics of the SemEval-2024 Task 8 Monolingual dataset (Wang et al., 2024). This study considers Accuracy and class-wise F1-score as the performance metrics. We consider baseline models from diverse categories to ensure a comprehensive evaluation and fair comparison of our proposed methods. This study considers statistical method GTLR (Zellers et al., 2019), supervised detectors such as OpenAI-detector and DetectGPT (Guo et al., 2023), the zero-shot detector-based method Binoculars (Hans et al., 2024), the adversarial learning-based method Radar (Hu et al., 2023), and the multi-class contrastive learning-based detector DeTeCtive (Guo et al., 2024) as baseline models. Furthermore, we fine-tune RoBERTa (Liu, 2019) for AI-generated review detection and incorporate stylometric and other linguistic features, as detailed in the corresponding Subsec. 3.1.1, to analyze their influence on detection performance. Appendix Subsec. F presents the fusion methods adopted to combine the RoBERTa encodings with stylometric and other linguistic features for AI-generated content detection. Table 10 in Appendix Subsec. C presents the details of the experimental hyperparameter.

5.2 Results and Discussions

Table 8, presents the performance of the proposed and baseline models on the DetectAIRev, SemEval Monolingual (Wang et al., 2024) Product Review (Salminen et al., 2022), Restaurant Review (Gambetti and Han, 2023), Hotel Review (Buscaldi and Liyanage, 2024) and DeepFake (Fagni et al., 2021) datasets. From Table 8, it is apparent that our proposed method **ProtoFewRoBERTa** outperforms recent state-of-the-art baseline models from literature across datasets in detecting AI-generated reviews and text. From Table 8, comparing the performance of RoBERTa with RoBERTa + Stylometric Fea-

Model	Human	GPT	LLaMA	Phi	Gemma	DeepSeek	Mistral	Qwen	Gemini	ChatGPT
F1	0.923	**0.945**	0.935	0.932	**0.945**	0.925	**0.945**	0.941	0.923	0.912

Table 9: Presents class-wise F1-scores of ProtoFewRoBERTa in the author-detection task, showing consistent performance across human and various LLM-generated text sources (author), including GPT, LLaMA, Phi, Gemma, DeepSeek, Mistral, Qwen, Gemini, and ChatGPT as author.

tures shows that incorporating stylometric features helps capture linguistic style cues and improves the performance of AI-generated review detection models. Compared to the multi-level contrastive learning approach of DeTeCtive (Guo et al., 2024), ProtoFewRoBERTa demonstrates clear performance advantages across both datasets. This improvement stems from prototype-based episodic training of ProtoFewRoBERTa, which captures fine-grained stylistic differences between text written by humans and generated by LLMs without relying on extensive contrastive objectives or heavily paired training data, addressing a key limitation of DeTeCtive in out-of-distribution scenarios. Furthermore, DeTeCtive (Guo et al., 2024) requires large, stylistically diverse labelled datasets from multiple LLMs to learn contrastive representations. In contrast, ProtoFewRoBERTa can learn prototypes from few labelled examples per class (few-shot), reducing data preparation overhead and annotation costs. The DeTeCtive (Guo et al., 2024) model requires retraining or fine-tuning whenever a new LLM style emerges. ProtoFewRoBERTa computes new prototypes for text generated by new LLMs (author) using minimal samples, making it highly scalable and future-proof for rapidly evolving LLMs. While DeTeCtive improves OOD detection compared to other baseline models in the literature, its generalization still degrades if new writing styles differ significantly from those in the training set. In comparison, the prototype space of ProtoFewRoBERTa naturally separates diverse styles, maintaining robust out-of-domain generalization with minimal recalibration. Furthermore, Table 9 presents the class-wise F1-scores of the ProtoFewRoBERTa model across human and LLM-generated text categories, underscoring the robustness of ProtoFewRoBERTa in author detection capability in distinguishing stylistic and semantic cues specific to different LLMs (author).

Ablation Analysis: We extend our experiments for deeper analysis (details in Appendix Sec. H). The qualitative error analysis (Appendix Subsec. H.1, Table 18) highlights three main error types: (i) **Generic brevity**, (ii) **Polished rewrites**, and (iii) **Noisy text**. Feature importance analysis (Appendix Subsec. H.2, Table 19) shows that emotional and experiential cues are informative but limited individually; combining stylometric and RoBERTa embeddings yields the best performance. Dataset preprocessing (Appendix Subsec. H.3, Table 20) slightly reduces scores but confirms robustness against non-standard English. Adversarial evaluation (Appendix Subsec. H.4, Table 21) shows ProtoFewRoBERTa remains resilient under semantic perturbations. Also, Extend setups (Appendix Subsec. H.5, Table 22) reveal that while weighted strategies improve interpretability, averaging achieves the highest overall performance.

6 Conclusion and Future Work

This paper proposes an AI-generated review detection dataset, DetectAIRev, where seven different LLMs generate reviews across five domains. We conduct several analyses and evaluations to assess the reliability of the proposed dataset and demonstrate that it is suitable for training an AI-generated review detection model across diverse domains and reviews generated by various LLMs. Furthermore, this study proposes an AI-generated text detection method *ProtoFewRoBERTa*, a few-shot framework that combines prototypical networks with RoBERTa embeddings, to learn discriminative features across multiple LLMs and human-written text using only a few labeled examples per class to discriminate between LLMs as the author for text author detection. We conduct our experiments on our proposed **DetectAIRev** dataset and other existing datasets, and our experimental results suggest that *ProtoFewRoBERTa* outperformed the state-of-the-art methods from the literature across the dataset. Furthermore, *ProtoFewRoBERTa* is easily scalable for detecting text generated by new authors (LLMs not included in the training dataset) and remains future-proof for rapidly evolving LLMs. To adapt the model to a new author or LLM, *ProtoFewRoBERTa* only requires estimating prototypes from a few labelled examples (generated by new LLMs), thereby reducing data preparation overhead and retraining of the model to accommodate detecting the text generated by LLMs not in training. This study identifies AI-generated review detection in low-resource languages and multilingual setups for future work.

7 Limitations

While DetectAIRev represents a significant advancement in AI-generated review detection across multiple domains, several limitations remain that warrant discussion. First, the current version of the dataset is limited to English-language reviews, which restricts the model's generalizability to multilingual or low-resource language scenarios. Extending the dataset to include diverse linguistic settings is crucial for evaluating cross-lingual robustness and supporting broader applicability. Second, although *ProtoFewRoBERTa* leverages few-shot learning for adaptability to unseen classes (i.e., text written by new LLMs), its performance may degrade when support examples are extremely sparse or noisy. Addressing this requires more robust prototype construction methods or hybrid techniques incorporating uncertainty modelling. Third, the model assumes clearly defined author categories. In real-world scenarios, however, writing styles may lie on a continuum or involve hybrid human-AI collaboration, making discrete classification more challenging. Lastly, while RoBERTa serves as a strong encoder backbone in our framework, its effectiveness may not generalise uniformly across other transformer-based encoders. A systematic evaluation of alternative backbones (e.g., BERT, DeBERTa, XLNet) is necessary to understand their impact on detection performance.

8 Ethical Considerations

Human-written reviews used to curate the DetectAIRev dataset are publicly available for academic research and development. *DetectAIRev* is curated based on reviews from publicly available datasets, including Book Reviews (Wan and McAuley, 2018), E-Commerce Reviews (Agarap, 2018a), Movie Reviews (Maas et al., 2011), TripAdvisor Hotel Reviews (Alam et al., 2016), and Restaurant Reviews (Abri et al., 2020). The LLM-generated content was produced via controlled prompting of large language models. No personally identifiable information (PII) was used or collected during the dataset construction process. All human-authored and machine-generated content is utilized solely for research purposes to advance the development of AI-generated text detection systems.

References

Majd AbedRabbo, Cathryn Hart, Fiona Ellis-Chadwick, and Zeina AlMalak. 2022. Towards rebuilding the highstreet: Learning from customers' town centre shopping journeys. *Journal of Retailing and Consumer Services*, 64:102772.

Faranak Abri, Luis Felipe Gutiérrez, Akbar Siami Namin, Keith S Jones, and David RW Sears. 2020. Linguistic features for detecting fake reviews. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 352–359. IEEE.

Abien Fred Agarap. 2018a. Afagarap/ecommerce-reviews-analysis: v0.1.0-alpha.

Abien Fred Agarap. 2018b. Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn). *arXiv preprint arXiv:1805.03687*.

Shifali Agrahari, Samridhi Bisht, and Ranbir Singh Sanasam. 2024. Text authorship attribution: Stylometric insights into human and llm-generated text. In *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD)*, pages 344–346.

Shifali Agrahari, Subhashi Jayant, Saurabh Kumar, and Sanasam Ranbir Singh. 2025a. Essaydetect at genai detection task 2: Guardians of academic integrity: Multilingual detection of ai-generated essays. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 299–306.

Shifali Agrahari, Prabhat Mishra, and Sujit Kumar. 2025b. Random at genai detection task 3: A hybrid approach to cross-domain detection of machine-generated text with adversarial attack mitigation. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 365–370.

Shifali Agrahari and Sanasam Ranbir Singh. 2025. Osint at genai detection task 1: Multilingual mgt detection: Leveraging cross-lingual adaptation for robust llms text identification. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 184–190.

Md Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Eagle: A domain generalization framework for ai-generated text detection. *arXiv preprint arXiv:2403.15690*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Davide Buscaldi and Vijini Liyanage. 2024. Detecting ai-enhanced opinion spambots: a study on llm-generated hotel reviews. In *Proceedings of the Seventh Workshop on e-Commerce and NLP@ LREC-COLING 2024*, pages 74–78.

Souradip Chakraborty, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2024. Position: On the possibilities of ai-generated text detection. In *Forty-first International Conference on Machine Learning*.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Raffaele Filieri. 2016. What makes an online consumer review trustworthy? *Annals of Tourism Research*, 58:46–64.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yu Fu, Deyi Xiong, and Yue Dong. 2024. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18003–18011.

Alessandro Gambetti and Qiwei Han. 2023. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Ackerloff George and 1 others. 1970. The market for lemons: Quality uncertainty and the market mechanism.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems*, 37:88320–88347.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17519–17537.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Sherry He, Brett Hollenbeck, and Davide Proserpio. 2022. The market for fake reviews. *Marketing Science*, 41(5):896–921.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.

Haowei Hua and Co-Jiayu Yao. 2024. Investigating generative ai models and detection techniques: impacts of tokenization and dataset size on identification of ai-generated text. *Frontiers in Artificial Intelligence*, 7:1469197.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, and Neil Zhenqiang Gong. 2024. Watermark-based attribution of ai-generated content. *arXiv preprint arXiv:2404.04254*.

Souheila Kaabachi, Selima Ben Mrad, and Maria Petrescu. 2017. Consumer initial trust toward internet-only banks in france. *International Journal of Bank Marketing*, 35(6):903–924.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection

through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. *arXiv preprint arXiv:2502.10003*.

Sujit Kumar, Monika Singh, Abhishek Ranjan, Tanveen Tanveen, and Sanasam Ranbir Singh. 2024. Prompt-based masked language modeling for numerical reasoning. *ACM Transactions on Intelligent Systems and Technology*.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Bing Liu. 2012. Opinion spam detection. In *Sentiment Analysis and Opinion Mining*, pages 113–125. Springer.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xinlin Peng, Ying Zhou, Ben He, Le Sun, and Yingfei Sun. 2023. Hidding the ghostwriters: An adversarial evaluation of ai-generated student essay detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10406–10419.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Md Shahidul Salim and Sk Imran Hossain. 2024. An applied statistics dataset for human vs ai-generated answer classification. *Data in Brief*, 54:110240.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Sunil Sharma and Satish Kumar. 2023. Insights into the impact of online product reviews on consumer purchasing decisions: A survey-based analysis of brands' response strategies. *Scholedge International Journal of Management & Development*, 10(1).

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Teo Susnjak and Timothy R McIntosh. 2024. Chatgpt: The end of online exam integrity? *Education Sciences*, 14(6):656.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, and 1 others. 2024. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2023a. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023b. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *ArXiv*, abs/2306.07401.

Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. 2017. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 1143–1158.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2025. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *IEEE Transactions on Big Data*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475*.

Appendix

A Supplementary Details of the Proposed Dataset

A.1 Diversity and Configuration of Large Language Models for Review Generation

To explore the diversity of generated reviews, we employ a wide range of large language models (LLMs) that vary in size, architecture, and capability. As shown in Table 10, these models differ in parameter count (ranging from 7B to 175B), context length, number of transformer layers, and attention heads. This architectural diversity enables a systematic analysis of how model complexity influences the quality and detectability of generated reviews.

A.2 Human Annotation and Evaluation Framework for the DetectAIRev Dataset

To assess the quality of reviews generated by LLMs in our proposed *DetectAIRev* dataset, we conducted human evaluations using four key metrics: **Fluency**, **Origin**, **Coherence**, and **Factuality**. We created a subset of the datasets consisting of eight hundred samples by randomly selecting one hundred reviews generated by each LLM and assigning them to four independent annotators. We recruited two annotators with an engineering background and two with a linguistics background to ensure high-quality annotations. Each annotator was asked to evaluate the reviews based on Fluency, Origin, Coherence, and Factuality, following the scoring criteria and dimension scales defined in Table 11. Annotators assigned scores following detailed labelling guidelines (Table 11), enabling fine-grained judgment of linguistic quality and realism while supporting a comprehensive analysis of the content generated by LLMs overall quality and

Model Name	Abbreviation	Parameters (Billion)	Context Length (Tokens)	Layers	Attention Heads
GPT-3	*GPT-3*	175B	2048	96	96
Meta-Llama-3.1-70B	*Llama*	70B	8192	80	64
Gemma-2-27B	*Gemma*	27B	8192	64	48
Mistral-Nemo-Base-2407	*Mistral*	13B	8192	80	40
Phi-3-Medium-4K	*Phi*	14B	4096	48	32
Qwen2.5-VL-7B	*Qwen*	7B	4096	32	16
DeepSeek-R1-GGUF	*DeepSeek*	16B	8192	60	32

Table 10: Comparison of different models used for LLM-generated reviews based on key parameters and hyperparameters.

Metric	Labels
Fluency	3 – The review contains no grammatical errors, and its meaning can be understood. 2 – The review contains some grammatical errors but is still understandable. 1 – The review contains many grammatical errors and cannot be understood.
Origin	1 – The review has characteristics of a human-written review. 0 – The review has characteristics of an LLM-generated review.
Coherence	1 – The review is logically consistent, with ideas flowing naturally and connecting smoothly. 0 – The review lacks logical consistency, with disjointed ideas or poor transitions between sentences.
Factuality	5 – The review is completely accurate and truthful. 4 – The review is mostly accurate but may leave out some details. 3 – The review is mostly accurate, but contains some unclear or vague statements. 2 – The review has noticeable inaccuracies or misleading statements. 1 – The review is mostly incorrect or irrelevant to the topic.

Table 11: Annotation scheme for human evaluation metrics.

its similarity to human writing. We measured inter-annotator agreement on the scores assigned for Fluency, Origin, Coherence, and Factuality using Krippendorff's alpha and Fleiss' kappa. Table 6 presents Krippendorff's alpha and Fleiss' kappa scores for inter-annotator agreement on eight hundred samples. The alpha values indicate the consistency with which annotators applied the scoring criteria, with values closer to 1 reflecting higher agreement. Accordingly, our reported Krippendorff's alpha values quantify the reliability of the annotators' dimension-wise scores and confirm that the evaluation results are trustworthy. At the same time, Krippendorff's alpha accounts for disagreement expected by chance and applies to both ordinal and categorical scales. Fleiss' kappa, in contrast, measures the degree of agreement among multiple annotators specifically for categorical ratings. The *Fluency* of text measures the grammatical correctness, readability, and natural flow of generated text. In contrast, *Coherence* evaluates the logical consistency and contextual connectivity of ideas across sentences within a review. Fluency evaluates the grammatical correctness and readability of a review, while Origin distinguishes whether it is human-written or LLM-generated. Coherence measures the logical flow and consistency of ideas, and Factuality assesses the accuracy and truthfulness of the review content.

Origin distinguishes whether a review is human-written or LLM-generated. Therefore, annotators assign a score of 0 if the review has characteristics of an LLM-generated review or generated by LLMs, and 1 if they find it to be written by a human (refer Table 11). The inter-annotator agreement for Origin is Krippendorff's Alpha = 0.671 and Fleiss' Kappa = 0.573 (in Table 6), indicating moderate agreement among annotators regarding the origin of a review, i.e., whether it is generated by an LLM or written by a human. Therefore, the LLM-generated reviews in our dataset are on par with human-written reviews, confirming that our proposed dataset is reliable for training models in AI-generated review detection. Consequently, given a review text, it is often difficult for annotators to determine whether it is LLM-generated or human-written. As a result, annotators sometimes rated the same review slightly differently, leading to only moderate inter-annotator agreement for the Origin metric. This further confirms that distinguishing between LLM-generated and human-written reviews is not trivial. From these observa-

Corpus Name	Size (Train)	Source	Language	Domain
HC3 (Guo et al., 2023)	84K	GPT-3	En	Q/A, Computer Science
M4 (Wang et al., 2023a)	147K	Multiple LLMs	Multiple	General
GPT-2 Output (Radford et al., 2019)	250K	GPT-2	En	Web
GPABenchmark (Liu et al., 2023)	1,200K	GPT-3	En	Scientific Writing (SW)
Deepfake (AbedRabbo et al., 2022)	319K	GPT, LLaMA, etc.	En	News, Q/A, etc.
TuringBench (Uchendu et al., 2021)	200K	GPT, Fair, etc.	En	News articles
DetectAIRev Dataset (Ours)	**256K**	LLaMA, Mistral, Deepseek, Phi, Gemma, Qwen, GPT-3	En	Reviews (Book, Movie, Hotel, Restaurant, E-commerce, etc.)

Table 12: Comparison of AI-generated text detection datasets from litrature and our proposed dataset **DetectAIRev Dataset**.

tions, we conclude that our proposed dataset provides a realistic and reliable benchmark for training models to detect AI-generated reviews.

Factuality assesses the accuracy and truthfulness of the review content, referring to the degree to which a review is accurate and reliable. Annotators were asked to rate factuality on a scale of 1 to 5, where 5 indicates that the review is completely accurate and truthful and 1 indicates that the review is mostly incorrect or irrelevant to the topic. Table 11 presents the scoring criteria used to assess the factuality of reviews generated by LLMs. However, people expressed their opinions, feedback, and experiences about a product, location, hotel, or restaurant food item. Therefore, factuality assessment can depend on the experience and domain knowledge of annotators; for instance, subtle product details or contextual nuances may be interpreted differently by different annotators. Therefore, annotators rated each review differently based on their prior experience and knowledge. In Table 6, Krippendorff's Alpha (0.776) indicates substantial consistency among annotators, while Fleiss' Kappa (0.563) reflects moderate agreement, which is expected given the categorical nature and inherent subjectivity of factuality judgments. For example, consider the review text: "I loved dining at SeaBreeze Restaurant. Their grilled salmon was fresh and delicious, and the menu said it was caught locally that morning." Here, annotators may differ in terms of factuality because verifying whether the salmon was actually *caught locally that morning* depends on external knowledge. Some annotators may treat the review as mostly accurate (focusing on the true parts, such as the quality of food), while others may penalise the potentially incorrect claim about the source of the salmon. So, the inter-annotator agreement score for factuality is moderate in our proposed dataset; however, this truly reflects the nature of reviews, where factual accuracy often depends on context and can be interpreted differently by annotators.

A.3 Comparison of the Proposed DetectAIRev Dataset with Existing AI-Generated Text Detection Datasets

To demonstrate the distinct breadth, relevance, and novelty of our proposed dataset, we perform a comparative analysis with existing AI-generated text detection datasets from the literature, as summarised in Table 12. Existing AI-generated text detection datasets in the literature primarily focus on domains such as question answering, scientific writing, and news articles, and often include AI-generated text produced by only a few notable LLMs, such as GPT. In contrast, our **DetectAIRev** dataset comprises both human-authored and LLM-generated texts spanning diverse domains, including book, e-commerce, movie, hotel, travel, and restaurant reviews. It further incorporates AI-generated reviews produced by a broader range of recent LLMs, including LLaMA, Mistral, DeepSeek, Phi, Gemma, Qwen, and GPT-3. This diversity establishes our dataset as a realistic and comprehensive benchmark for AI-generated review detection, enabling more reliable evaluation and generalization of detection models across diverse domains and writing styles.

B Feature Extraction and Estimation for Linguistic, Readability, and Psychological Features

Linguistic, readability, and psychological features were computed using standard NLP and readability formulas as mentioned in the literature (Mindner et al., 2023; Agrahari et al., 2024). Linguistic metrics such as word, sentence, and character counts were extracted using library NLTK and spaCy. Also, readability measures are formula-based metrics: Flesch Reading Ease (FRE) (Fleiss, 1971) as defined in Eq. 7 and Flesch–Kincaid Grade Level (FKGL) (Eq. 8 used word, sentence, and syllable

ratios.

$$FRE = 206.835 - 1.015 \times \frac{\text{\# Words}}{\text{\# Sentences}} - 84.6 \times \frac{\text{\# Syllables}}{\text{\# Words}} \quad (7)$$

$$FKGL = 0.39 \times \frac{\text{\# Words}}{\text{\# Sentences}} + 11.8 \times \frac{\text{\# Syllables}}{\text{\# Words}} - 15.59 \quad (8)$$

Gunning Fog Index considered complex words as defined 9; Similarly formular for all other such as Dale–Chall and SMOG Index estimated difficulty based on unfamiliar and polysyllabic words; Automated Readability Index (ARI) used character and word length; and Yule's K captured lexical richness from word frequency distributions.

$$GFI = 0.4 \times \left(\frac{\text{\# Words}}{\text{\# Sentences}} + 100 \times \frac{\text{\# Complex Words}}{\text{\# Words}} \right) \quad (9)$$

Table 13 lists the range and interpretation of all readability metrics. On the other hand, psychological features were derived using VADER and TextBlob for sentiment polarity and LIWC (Pennebaker et al., 2001) or NRC Emotion Lexicon (Mohammad and Turney, 2013). for emotional and cognitive categories. Finally, coherence was measured via cosine similarity between consecutive sentence embeddings from Sentence-BERT (Zhang* et al., 2020), reflecting logical flow and semantic consistency. The extraction and analysis of these features help in evaluating content authenticity and distinguishing between human and LLM-generated reviews effectively.

C Hyperparameters and Training Configuration

We train the proposed **ProtoFewRoBERTa** model within an episodic meta-learning framework designed for few-shot classification. Each episode samples a subset of classes and constructs corresponding support and query sets. The model employs **RoBERTa** as the base encoder, followed by a trainable projection layer to obtain task-adaptive embeddings. Class prototypes are computed from support samples, and query samples are classified based on their proximity to the nearest prototype in the embedding space using the squared Euclidean distance. Model parameters are optimized using the AdamW optimizer with a negative log-likelihood loss over the query set. The key hyperparameters and training configurations are summarized in Table 14.

Metric	Range	Interpretation
Flesch Reading Ease (FRE)	90–100	Very easy (e.g., children's books)
	70–89	Easy to read (e.g., general news)
	50–69	Fairly difficult (e.g., academic articles)
	30–49	Difficult (e.g., technical writing)
	0–29	Very difficult (e.g., legal text)
Flesch-Kincaid Grade Level (FK Grade)	0–5	Very easy (elementary level)
	6–8	Fairly easy (middle school)
	9–12	Standard difficulty (high school)
	13+	Difficult (college and above)
Gunning Fog Index	6–8	Easy to read (general audience)
	9–12	Moderately difficult (high school level)
	13+	Difficult (college level and above)
Dale-Chall Readability Score	4.9 or lower	Easily understood by 4th graders
	5.0–9.9	Standard difficulty (grades 5–12)
	10+	Difficult (college level)
SMOG Index	1–6	Easy to read (elementary to middle school)
	7–9	Fairly difficult (high school)
	10+	Difficult (college level and above)
Automated Readability Index (ARI)	1–5	Very easy (elementary school)
	6–8	Fairly easy (middle school)
	9–12	Standard difficulty (high school)
	13+	Difficult (college level and above)
Yule's K	Low	Simple, repetitive text
	Medium	Moderate lexical variety
	High	Complex text with diverse vocabulary

Table 13: Interpretation of Readability Metrics

D Effect of Support and Query Set Sizes on the Performance of ProtoFewRoBERTa

To study the influence of the number of support samples per class (K) and the number of query samples per class (Q) on the performance of **ProtoFewRoBERTa**, we conduct an empirical evaluation over two key hyperparameters: (i) the number of support samples per class (K) and (ii) the number of query samples per class (Q). We evaluated the performance of **ProtoFewRoBERTa** by varying the number of support samples per class (K) across 4, 8, 16, 32, and 64, and the number of query samples per class (Q) across 10, 15, 20, and 30, to examine their impact on overall model effectiveness. Table 15 presents the performance of the proposed **ProtoFewRoBERTa** model across different values of K and Q. Table 15 reveals that ProtoFewRoBERTa achieves its best performance at K = 8 and Q = 15, with the highest accuracy and F1 scores across both human and AI-generated review classes. Increasing K or Q beyond K = 8 and Q = 15 does not lead to further improvements and, in fact, slightly degrades performance, likely due to redundancy and reduced episode diversity. Conversely, smaller K and Q values provide insufficient support and query representation, limiting

Parameter	Value / Description
Base Encoder	RoBERTa-base
Projection Layer	Trainable linear layer
Few-shot Classes per Episode (N)	8
Support Samples per Class (K)	8
Query Samples per Class (Q)	15
Optimizer	AdamW
Learning Rate	2e-5
Weight Decay	0.01
Batch Size (episodes)	100
Dropout Rate	0.1 (on projection layer)
Sequence Length	256
Training Episodes	1000
Loss Function	Episodic Negative Log-Likelihood
Distance Metric	Squared Euclidean Distance
Hardware	NVIDIA A100 (40GB GPU)
Train:Val:Test	70:10:20

Table 14: Hyperparameters and training settings for the proposed ProtoFewRoBERTa model.

Para.	Value	Accuracy	F1-Human	F1-AI
K-shot	4	0.93	0.92	0.93
	8	**0.98**	**0.97**	**0.98**
	16	0.94	0.93	0.93
	32	0.93	0.93	0.93
	64	0.92	0.91	0.91
N-Query	10	0.95	0.93	0.97
	15	**0.98**	**0.97**	**0.98**
	20	0.92	0.92	0.92
	30	0.93	0.92	0.92
	40	0.93	0.92	0.92

Table 15: Performance across varying support and query set sizes.

generalisation. From such observations, we conclude that a moderate number of support and query samples per class (K = 8, Q = 15) offers the optimal trade-off between representational diversity and learning stability.

E Description of the SemEval Dataset

In SemEval 2024 Task 8 (Subtask A) (Wang et al., 2024) (monolingual track), the focus is on English-only texts aimed at distinguishing machine-generated from human-written content. The dataset, summarised in Table 16, includes statistics across multiple text generators, domains, and data splits. The training set covers five domains—Wikipedia, WikiHow, Reddit, arXiv, and PeerRead—comprising 56,400 machine-generated and 63,351 human-written texts. The development (Dev) set introduces BLOOMz as an unseen generator, with 2,500 machine-generated and 2,500 human-written samples. The test set utilises OUTFOX (Koike et al., 2024) as the surprising domain and GPT-4 as the surprising generator, comprising a total of 18,000 machine-generated and 16,272 human-written texts. This configuration ensures a diverse and challenging evaluation environment for detecting AI-generated content. Our proposed Dataset and Code repository are publicly available at the https://huggingface.co/datasets/Sifi-world/DetectAIRev and https://github.com/sifii/Detect-AI-Generated-Reviews-ProtoFewRoBERTa-and-DetectAIRev.

F Fusion of Stylometric Features with RoBERTa

We integrate stylometric and linguistic features into the RoBERTa-based classification framework using a feature fusion strategy. We first extract the stylometric and linguistic features described in Section 3.1.1, encompassing lexical, readability, sentiment, psycholinguistic, and similarity-based metrics. Next, these features are aggregated into a fixed-length feature vector that represents each input text. While training RoBERTa with feature fusion, each input review is passed through RoBERTa to obtain the [CLS] token representation. Subsequently, the [CLS] token representation is concatenated with the external, handcrafted, fixed-length feature vector. Next, the concatenated feature vector is passed through a fully connected neural network layer, enabling the model to learn from both contextual embeddings and interpretable stylometric cues jointly. This fusion approach leverages the deep language understanding capabilities of RoBERTa alongside the interpretability and discriminative power of handcrafted features, resulting in improved performance in distinguishing AI-generated reviews from human-written ones. As evident in Table 8, integrating stylometric features with RoBERTa leads to a clear improvement over using RoBERTa alone. This enhancement demonstrates that incorporating linguistic and stylistic cues such as lexical richness, readability, and emotional tone provides complementary information that strengthens the ability of the model to distinguish between human-written and LLM-generated reviews. The fusion of deep contextual embeddings with interpretable handcrafted features, therefore, results in a more robust and explainable detection framework.

G Comprehensive Review Dataset Design: Human, Facet-Aware, and Adversarial Perspectives

To produce high-quality LLM-generated reviews that closely align with human-authored content, we adopted a human-aligned prompting strategy

Split	davinci-003	ChatGPT	Cohere	Dolly-v2	BLOOMz	GPT-4	Machine Total	Human Total
Train								
Wikipedia	3,000	2,995	2,336	2,702	–	–	11,033	14,497
Wikihow	3,000	3,000	3,000	3,000	–	–	12,000	15,499
Reddit	3,000	3,000	3,000	3,000	–	–	12,000	15,500
arXiv	2,999	3,000	3,000	3,000	–	–	11,999	15,498
PeerRead	2,344	2,344	2,342	2,344	–	–	9,374	2,357
Dev								
Wikipedia	–	–	–	–	500	–	500	500
Wikihow	–	–	–	–	500	–	500	500
Reddit	–	–	–	–	500	–	500	500
arXiv	–	–	–	–	500	–	500	500
PeerRead	–	–	–	–	500	–	500	500
Test								
Outfox	3,000	3,000	3,000	3,000	3,000	3,000	18,000	16,272

Table 16: Subtask A: Monolingual Binary Classification. Data statistics for Train/Dev/Test splits across various models and sources.

across five domains: E-Commerce, Hotel, Movie, Book, and Restaurant. For E-Commerce, we used the Women's E-Commerce Clothing Reviews dataset and designed prompts like storytelling, comparative, and use-case specific reviews (e.g., a 34-year-old reviewing a "Classic Fit Jacket" rated 4 stars). Hotel reviews were based on TripAdvisor data, with prompts focusing on experiential narratives and occasion-based stays (e.g., a honeymoon highlighting luxury and service). Movie prompts leveraged IMDb-style emotional and critic-style reviews (e.g., describing acting and visuals without naming the movie). For Book reviews, we chose emotionally rich 4–5 star Amazon reviews and crafted recommendation prompts (e.g., praising character development despite a low rating). Restaurant reviews, drawn from short user entries, were generated using expressive, emotional, and comparative prompts (e.g., poetic descriptions of ambiance or flavor comparisons). This prompting framework emphasized emotional realism, personalization, and stylistic diversity to mirror human review patterns across all domains.

G.1 Adversarial Attacks for Realistic Review Generation

To enhance the realism and robustness of the `DetectAIRev` dataset, we introduce adversarial perturbations to a subset of LLM-generated reviews. These perturbations simulate strategies commonly used to evade AI-generated text detection systems. Inspired by techniques outlined in Dugan et al. (2024), we apply the following three types of adversarial attacks:

- **Alternative Spelling**: Common words are replaced with regionally or phonetically equivalent variants (e.g., "color" → "colour", "favorite" → "favourite") to test sensitivity to orthographic variants.
- **Paraphrasing (Rewrite) Attack**: Sentences are rephrased using automatic paraphrasing tools or prompt-based rewriting to retain semantic meaning while altering syntax and style. This attack challenges the detector's ability to generalize beyond surface-level phrasing.
- **Misspelling Attack**: Intentionally introduced spelling errors (e.g., "excellent" → "excellant", "battery" → "batery") are used to simulate noisy user input and test the detector's robustness to typographical noise.

Each attack type was applied to a random subset of LLM-generated reviews across different domains and models. These adversarial variants are included in the dataset with corresponding metadata flags to support targeted evaluation and robustness testing of detection models.

G.2 Prompting Strategies and Evaluation

To generate LLM-based reviews closely aligned with human-written reviews, we designed multi-

ple prompting strategies—namely zero-shot, few-shot, replication-based, and facet-guided prompting. Each approach was applied to generate domain-specific reviews (E-commerce, Hotel, Book, Movie, and Restaurant), and their outputs were evaluated against human-written reviews using BLEU and METEOR scores.

Zero-Shot Prompting. In the zero-shot setting, the LLM was only provided with a single task instruction without examples. For instance, in the hotel domain, the prompt was:

```
"Generate a realistic review where a {Age}-year-old user shares their experience using the product '{Title}', rated {Rating} stars. Focus on durability and value."
"Use review/score and summary by {profileName} to recommend {Book Title}. Mention plot richness, price, and time."
```

While effective at producing grammatically sound outputs, these reviews tended to be generic, lacking the stylistic nuance and domain-specific flair present in human reviews.

Few-Shot Prompting. This strategy involved providing 5–6 diverse human-written examples before the instruction. These examples were rotated every 100 generations to promote diversity. A representative few-shot hotel review prompt looked like:

```
Example 1: "The location was perfect and the staff were incredibly helpful. The breakfast buffet was a highlight of my trip."
Example 2: "Our room had a fantastic view of the skyline and the amenities exceeded expectations."
Now, write a similar 5-star hotel review based on a comfortable stay, emphasizing service and cleanliness.
```

Few-shot prompting produced outputs that were significantly more human-like in tone, often replicating patterns of authentic customer expression.

Replication-Based Prompting. In this setting, we selected high-quality human-written reviews and instructed the LLM to mimic their tone, structure, and stylistic richness while altering the content. For example:

```
Here is a {Human written} hotel review: "From the check-in to check-out, everything was seamless. The concierge made sure we had local maps and restaurant tips."
Now, write a new review about a different hotel experience that matches this style and structure.
```

This approach improved coherence and paragraph structuring in generated reviews and helped simulate human discourse organization.

Facet-Guided Prompting. To further enrich the generation process, we incorporated key review facets such as sentiment, aspects (e.g., service, price, ambiance), and user profiles (e.g., occasion, age). An example from the restaurant domain is:

```
"Write a detailed restaurant review covering the following facets: facets 1, facets 2, and facets 3."
```

These facet-driven prompts ensured higher content grounding and emotional resonance, key to aligning with human preferences.

Evaluation Using BLEU and METEOR. To quantify the similarity between human-written and LLM-generated reviews, we computed BLEU and METEOR scores across domains. As shown in Figure 2, The llm generation achieved BLEU and METEOR scores closest to human reference reviews. In summary, the use of few-shot and facet-guided prompting yielded the most human-aligned reviews, capturing not only the lexical patterns but also the emotional and structural authenticity of real reviews. These strategies proved essential in bridging the gap between synthetic and human-authored review generation.

G.3 Facet-Aware Review Generation

This section elaborates on the facet-aware review generation strategy used in our dataset. By guiding Large Language Models (LLMs) to focus on specific aspects (or facets) extracted from human-written reviews, we enhance the alignment and realism of the generated content. To emulate the natural focus of human reviewers, we performed facet analysis on each human-written review $\mathcal{R}_{\mathcal{H}}$. Facets refer to the key aspects frequently discussed by users in reviews. These include domain-specific elements such as "food quality" or "room cleanliness" as mention in Table 17. We extracted candidate facets using a two-step approach:

Domain	Facet 1	Facet 2	Facet 3	Facet 4	Facet 5
Restaurant	Food Quality	Service	Ambiance	Price	Location
Hotel	Cleanliness	Staff Behavior	Amenities	Location	Check-in Process
Movie	Plot	Acting	Cinematography	Soundtrack	Pacing
Book	Writing Style	Characters	Storyline	Themes	Length
E-Commerce	Product Quality	Delivery Time	Price	Packaging	Return Policy

Table 17: Top-5 most frequently mentioned facets in human-written reviews ($\mathcal{R}_{\mathcal{H}}$) across different domains.

- **Linguistic Phrase Extraction:** We applied noun phrase chunking using spaCy to identify candidate aspect terms from each review.
- **TF-IDF Ranking and Filtering:** Domain-specific high-TF-IDF terms were retained as dominant facets. Manual curation ensured aspect relevance.

Facet-aware prompting plays a crucial role in ensuring the semantic fidelity and human-likeness of LLM-generated reviews. By incorporating domain-specific aspect control, we ensure that LLM reviews are more nuanced and comparable to real-world human-authored content, thereby enriching the dataset and enhancing the robustness of downstream detection tasks.

H Comprehensive Evaluation: Ablation Studies and Error Analysis

H.1 Qualitative Error Analysis

To better understand the limitations of *ProtoFewRoBERTa*, we conducted a qualitative error analysis on misclassified examples from the test set. Table 18 presents representative cases where the model failed to correctly identify the origin of a review (i.e., human or LLM-generated).

We identify three key error patterns: (i) **Generic brevity** – short, templated human reviews often resemble LLM outputs; (ii) **Polished rewrites** – formal or paraphrased human reviews are misclassified due to surface style; (iii) **Noisy text** – ungrammatical or informal language in human reviews is mistaken for LLM-generated content.

H.2 Evaluating the Importance of Stylometric Features in AI-Generated Review Detection

To analyse the importance of stylometric features, as discussed in Subsection 3.1.1, we conducted post-hoc analyses and empirical evaluations to identify which features are most influential for AI-generated review detection and how they contribute to distinguishing human-written from LLM-generated reviews. As shown in Table 19, We consider the following characteristic in our empirical evaluations: (i) Lexical + Length, (ii) Stylometry + Readability, (iii) Spelling + Noise, (iv) POS + Syntax, (v) Discourse Markers, (vi) Redundancy+Entropy, (vii) LM Stats (PPL+ Burgundiness) (viii) Sentiment/Subjectivity (ix) Emotional words + Personal experience phrases (x) All (features only) and (xi) Hybrid (RoBERTa + All) list each feature one by one and add a line about the feature. From Table 19, it is evident that while features such as emotional words and personal experience phrases help distinguish shorter human reviews, their individual predictive power remains limited. The hybrid model, which combines RoBERTa embeddings with stylometric feature types, achieves the best overall performance, highlighting the complementary nature of linguistic features and transformer-based embeddings.

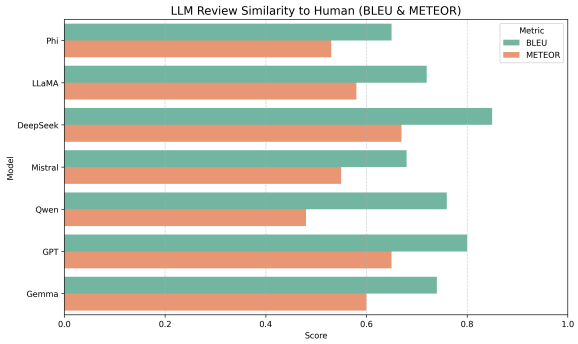

Figure 2: BLEU (green) and METEOR (orange) scores between human and each LLM, showing that DeepSeek-generated reviews are most similar to human-written ones.

True	Review Snippet	Predicte	Likely Cause
Human	*"Great value. Worked as expected. Would buy again."*	LLM	Too short / templated phrasing
Human	*"This product delivered on all fronts—excellent quality and exceptional value."*	LLM	Adversarial rewrite mimics LLM style
LLM	*"The aesthetics of the device blend seamlessly with modern home decor."*	Human	Subjective tone / stylistic personalization
LLM	*"It's okay, nothing special. Gets the job done."*	Human	Flat tone; resembles minimal user feedback
Human	*"Gud prodct. Wurks gr8! Thnx."*	LLM	Non-standard spelling; perceived as synthetic
Human	*"One of the most beautiful place in Nepal. It is heaven human."*	LLM	Ungrammatical phrasing; perceived as unnatural
Human	*"I love this dress and want it in every color."*	LLM	Common LLM-generation pattern for positive sentiment

Table 18: Representative error cases and analysis from ProtoFewRoBERTa.

Embedding / Features	Accuracy	F1-H	F1-AI
Lexical + Length	0.5523	0.56	0.55
Stylometry + Readability	0.5580	0.55	0.53
Spelling and Noise	0.5210	0.53	0.51
POS / Syntax	0.5820	0.59	0.58
Discourse Markers	0.5765	0.58	0.57
Redundancy and Entropy	0.5380	0.54	0.53
LM Stats (PPL and Burstiness)	0.6123	0.59	0.60
Sentiment and Subjectivity	0.5715	0.57	0.57
Emotional Words and Phrases	0.5272	0.53	0.52
All (Features Only)	0.5625	0.55	0.58
Hybrid (RoBERTa + All)	**0.9576**	**0.95**	**0.96**

Table 19: Performance comparison of different embedding and feature sets for AI-generated text detection.

H.3 Bias and Generalization Challenges across English Variants

To evaluate the robustness of our model against linguistic normalisation and spelling variation, we conducted an empirical analysis examining the effect of preprocessing on detection performance. This analysis is crucial for assessing whether standardisation and the inclusion of non-standard English forms influence the ability of the model to fairly and consistently distinguish human-written from AI-generated reviews. We incorporated a preprocessing step into the proposed DetectAIRev dataset. Specifically, we normalized British spellings to American spellings (e.g., "colour" $\rightarrow$ "color", "favourite" $\rightarrow$ "favorite"), and flagged simplified non-standard spellings (e.g., "Gud", "luv", "plz", "coz"). Additionally, we included a small number of non-standard English samples in training for regularization. The dataset was tested with our proposed method as shown in Table 20.

Dataset	Acc	F1-H	F1-AI
DetectAIRev	0.983	0.970	0.981
DetectAIRev (after preprocessing)	0.971	0.964	0.973

Table 20: Comparison of ProtoFewRoBERTa performance before and after text standardisation and normalisation on the DetectAIRev dataset.

Table 20 reveals that preprocessing the DetectAIRev dataset leads to only a marginal decrease in performance across all metrics. Notably, the results demonstrate that the model does not exhibit misclassification bias toward non-standard English (e.g., simplified spellings such as "Gud prodct").

H.4 Robustness of ProtoFewRoBERTa under Deep Semantic Adversarial Attacks

To investigate the robustness of ProtoFewRoBERTa against adversarial attacks, we manually constructed two deep semantic adversarial examples by rewriting existing reviews to maintain their original sentiment and semantic content while introducing substantial stylistic variations. For instance, simple declarative sentences were transformed into complex or exclamatory ones; synonyms, and phrase-level variations. Table 21 presents the performance of our proposed model, **ProtoFewRoBERTa**, trained on the **DetectAIRev** dataset and evaluated on the newly curated deep semantic adversarial test set to assess its robustness against semantic and stylistic perturbations. Table 21 presents the performance of our proposed model, **ProtoFewRoBERTa**, trained on the **DetectAIRev** dataset and evaluated on the newly curated deep semantic adversarial test set to assess its robustness in detecting AI-generated reviews exhibiting semantic and stylistic perturbations.

Dataset	Accuracy	F1-Human	F1-AI
DetectAIRev	0.983	0.970	0.980
Add Deep Adversarial	0.895	0.840	0.880

Table 21: ProtoFewRoBERTa performance under deep semantic adversarial examples.

Table 21 reveals that ProtoFewRoBERTa exhibits notable robustness in detecting AI-generated reviews under deep semantic adversarial conditions. Although its performance shows a relative decline when faced with stylistic and structural perturbations, the model consistently preserves its ability to capture core sentiment cues. This observation suggests that the model generalises effectively to linguistically diverse inputs and remains resilient against semantically preserved adversarial manipulations.

H.5 Incorporating Review-Specific Cues into the Prototype Formation Process of ProtoFewRoBERTa

Understanding the subtle linguistic characteristics that distinguish human-written from LLM-generated reviews requires more than semantic representation alone. Conventional prototype formation in few-shot frameworks, such as ProtoFewRoBERTa, often relies on uniform averaging of embeddings, which may overlook stylistic signals uniquely indicative of review authenticity, including emotional tone, personal pronoun use, or informal phrasing. To address this limitation, we enhance ProtoFewRoBERTa by integrating lightweight, review-specific cues into prototype

computation, which enables the model to assign greater importance to linguistic traits that more accurately capture the writing patterns of genuine human reviews versus synthetic ones, thereby enhancing interpretability and robustness.

We further implemented a Feature Weighted Prototype (FWP) aggregation strategy under four different settings:

(i) ProtoFewRoBERTa (average): This setup computes class prototypes by taking the simple mean of RoBERTa-derived embeddings across all support samples, treating each review equally without incorporating any stylistic or feature-based weighting.

(ii) ProtoFewRoBERTa+FWP(rule-based): The objective of this step is to move beyond averaging all review embeddings equally. Instead, we assign higher weights to reviews that exhibit typical review cues, such as first-person phrases or emotional words, so that class prototypes better reflect realistic review characteristics. Each review x is represented in two complementary ways. First, a RoBERTa embedding e(x) captures its semantic meaning in a dense vector space. Second, a lightweight cue vector z(x) encodes interpretable surface features, including the frequency of first-person words, emotional word density, the use of connectors or templates, n-gram repetition, and the rate of typos or out-of-vocabulary tokens. These features are standardized (z-scored) so that they can be consistently compared. Rule-based weighting. We then compute a weight for each review using simple rule-based functions. For human-written reviews, the weight is increased when first-person words or typos appear, but decreased when connectors or repeated phrases dominate. Conversely, for LLM-generated reviews, the weight is increased when connectors and repetitions are frequent, but decreased when typos or first-person usage is high. To ensure stability, the weights are clipped between 0.5 and 2.0.

(iii) ProtoFewRoBERTa + FWP (learned): learned parameter-based feature weighting: To compute a prototype for each class (human or LLM), we start with the support set of examples belonging to that class. Each example is first mapped into an embedding space using RoBERTa, and then multiplied by a weight that reflects its linguistic cues. Mathematically, the prototype is obtained by taking the weighted sum of the embeddings for all examples in the class and normalizing it by the total weight. This ensures that more representative samples contribute more strongly to the prototype than less representative ones. Weight Estimation: The weight wc(x) for each support example is estimated using lightweight linguistic cues extracted from the review text. Specifically, we compute a small feature vector that includes interpretable statistics such as the rate of first-person pronouns, emotional word density, the use of connectors/templates, n-gram repetition, and the frequency of typos or out-of-vocabulary tokens. These features are standardized (z-scored) and combined through simple rule-based functions: For human-written reviews weights are increased by first-person usage and typos but decreased by excessive connectors or repetition. For LLM-generated reviews weights are increased by connectors and repetition but decreased by first-person usage or typos. To avoid extreme scaling, the final weight is clipped within a stability range (e.g., 0.5 to 2.0). In this way, the weight reflects how representative a sample is of its class, and more typical reviews contribute more strongly to the prototype.

(iv) ProtoFewRoBERTa + FWP + length-norm: feature weighting with length normalization for short texts: Instead of using only handcrafted rules, we also learn per-class scorers directly from the linguistic cues. For each review, the cue vector is combined with a set of learned parameters to produce a score. This score is then transformed into a positive weight, which is used in the prototype computation. The parameters for this scorer are trained jointly with the encoder in an episodic framework, using query cross-entropy loss. To ensure stability and interpretability, cue features are z-scored, and embeddings are L2-normalized. The number of cues is kept small (8) to maintain efficiency. For short texts, we apply length normalization so that very short reviews do not dominate the prototype computation.

The results in Table 22 demonstrate that while the simple averaging baseline of ProtoFewRoBERTa achieves the highest overall accuracy, incorporating feature-weighted prototypes improves interpretability and maintains competitive performance. The learned feature-weighted variant outperforms the rule-based approach, indicating that the model benefits from adaptively learning the relative importance of review-specific cues. Furthermore, adding length normalization yields a balanced trade-off between accuracy and class-wise F1 scores, suggesting that normalization helps stabilize

Method	Acc	F1-H	F1-AI
ProtoFewRoBERTa (average)	**0.9830**	**0.970**	**0.980**
ProtoFewRoBERTa + FWP (rule-based)	0.9355	0.940	0.950
ProtoFewRoBERTa + FWP (learned)	0.9555	0.940	0.960
ProtoFewRoBERTa + FWP + length-norm	0.9643	0.960	0.980

Table 22: Comparison of ProtoFewRoBERTa and its variants with Feature Weighting Prototype (FWP) and length normalization on short texts.

prototype representations for shorter reviews without compromising detection precision.

H.6 Analysis of Cross-Domain Robustness of ProtoFewRoBERTa across Unseen Review Domains

Evaluating model robustness beyond the training domain is essential for understanding real-world generalization. In practical deployment scenarios, AI-generated review detection systems must operate reliably across diverse domains and writing contexts that differ from their training data. To assess this cross-domain adaptability, we examine how models trained on our proposed dataset, DetectAIRev, perform when exposed to previously unseen domains. Specifically, we evaluate our methods on the Product Review and Hotel Review datasets, which serve exclusively as test sets to simulate domain shift and linguistic variability.

Domain	Acc	F1-H	F1-AI
SemEval	0.71	0.65	0.80
Product Review	0.70	0.68	0.71
Hotel Review	0.88	0.88	0.88

Table 23: Performance across different domains when trained on the proposed dataset.

Table 23 presents the performance of our proposed model ProtoFewRoBERTa when trained on the proposed DetectAIRev dataset and evaluated on the Product Review and Hotel Review datasets. The performance of the proposed method in Table 23 indicates that models trained on the DetectAIRev dataset demonstrate strong generalisation capability to the Hotel Review domain, achieving high accuracy and balanced F1 scores across human and AI-generated classes. In contrast, performance on the Product Review domain declines moderately, suggesting sensitivity to domain-specific linguistic variations such as descriptive product terminology and informal phrasing. From such observations, we conclude that while the proposed model generalizes effectively to semantically similar domains, its robustness can be further improved for domains exhibiting greater lexical and stylistic divergence from the training distribution.