Ghostwriters of the Marketplace: A Corpus of Machine-Generated Reviews for Google Services and Businesses

Karla Schäfer[1,2], **Mohammed Mofreh**[2], **Martin Steinebach**[1,2]

[1]Fraunhofer Institute for Secure Information Technology, National Research Center for Applied Cybersecurity
[2]Technical University of Darmstadt
karla.schaefer@sit.fraunhofer.de, mohammed.mofreh@sit-extern.fraunhofer.de, martin.steinebach@sit.fraunhofer.de

Abstract

Generative models are simplifying the process of creating fake reviews. Models such as ChatGPT or DeepSeek can be used with ease to generate fake reviews with various ratings. However, these fake reviews can mislead buyers/readers. We generated a dataset with 78k reviews (building our "main" dataset) generated by 6 LLMs using three different writing styles, 117 tones and three ratings (positive, negative, and neutral). We used 50 prompts, multi-turn prompting and paraphrasing to generate a diverse range of reviews. Additionally, we created another, smaller dataset (external), using the same prompts and tones as before, but using three additional LLMs. This dataset allows detectors of AI-generated text to be tested on their ability to detect texts created using similar prompts but before unseen LLMs. We performed first tests using the newly generated dataset for AI-generated reviews detection. We fine-tuned RoBERTa on our dataset and, for testing generalizability, on the benchmark RAID using Reddit posts, news, and IMDb reviews. On the test splits of seen data, the detector achieved satisfactory results with an F1-score of 99.98%. On RAID the performance dropped, showing potential future research directions.

Introduction

The task of identifying AI-generated reviews involves determining whether a given review text was originally written by a human or was generated or faked by an AI model. A closely related task is the detection of fake reviews. In this case, real reviews are separated from fake ones which is not necessarily generated by AI but can also be written by human (Meng et al. 2025). In AI-generated review detection, we try to identify artefacts introduced when the text is generated by a Large Language Model (LLM). In fake review detection, inconsistencies in the reviews' sentiment, for example, are analysed. Both tasks are critical, since reviews influence human decision-making. Fake product reviews, particularly on e-commerce platforms, can influence consumers' decision-making and shopping behaviours (Akesson et al. 2023). User reviews can play a significant role in determining an organisation's revenue in e-commerce.

Because of this crucial role of reviews, we generated a new and large dataset with fake reviews. In artificial intelligence, a huge problem is the generalizability of trained detectors. Therefore, mostly, detectors perform good on seen data, hence data of domains they were trained on, but fail in detecting, new, unseen domains. Therefore, diverse datasets from different domains are necessary to develop generalizable detectors. But, limited datasets of AI-generated reviews exist. Therefore, we generated a new dataset of fake reviews generated by AI. The reviews were generated using diverse LLMs, prompts, tones, and writing styles. Furthermore, the AI-generated reviews were generated using summaries of human written Google Maps reviews. Our goal was not to improve the generation process or to evaluate the generation quality of LLMs but to enlarge the dataset pool for AI-generated text datasets of different domain and with different methods applied for controlling the generation process, such as adding human like mistakes.

Our contribution is three-fold. First, we introduce a new dataset of AI-generated Google reviews generated using a diverse set of prompts, tones, and LLMs. The dataset is made publicly available. Secondly, we performed an in-depth analysis of our dataset using sentence embeddings and linguistic features. Lastly, we performed first experiments using our dataset for AI-generated reviews detection. Additionally, for testing the detectors' generalizability, we used the benchmark dataset RAID (Dugan et al. 2024) containing AI-generated texts of different domains (e.g. Reddit, news texts, IMDb reviews).

Related Work

Earlier work (Abri et al. 2020) illustrated how to use Natural Language Processing (NLP) to detect faked product reviews by only analysing linguistic features such as sentence length, redundancy, or stylistic markers. In this study, 25 real Amazon product reviews were gathered, and 25 AI-generated reviews were generated using LLMs with prompts mimicking human writing style. 288 human participants read all 50 reviews at random and decided whether each one was real or fake. The participants correctly classified 50.82% of the reviews. An interesting finding was the high level of confidentiality among the participants when classifying the reviews. Additionally, the researchers tested the LLMs' ability to identify real versus fake reviews. Seven types of LLMs were tested: ChatGPT-o1, DeepSeek-R1, Grok-3, Gemini-2.0-Flash-Thinking, ChatGPT-4o, Gemma-3-27B-it

and Qwen2.5-Max. Each LLM received the same 50 reviews and was asked to classify them as real or fake with confidence. Once again, the results were close to random guessing, with most of the reviews being classified as real.

Yang (2025) highlighted the growing problem of AI-generated product reviews on platforms like Amazon, where customer reviews play a crucial role in shaping consumers' purchasing decisions. The study analysed over 6000 reviews of AI-generated and real product reviews. Machine learning techniques, like support vector classifiers were used for detection. The results indicate that fake reviews could be effectively distinguished from real human reviews. Moreover, his findings emphasize that fake AI-generated reviews usually mimic product description rather than providing authentic user experiences. Developing reliable detection systems is essential not only for maintaining the credibility of online e-commerce platforms but also for protecting customers from being misled by AI-generated faked product reviews.

Mitrović, Andreoletti, and Ayoub (2023) studied the detection of human written and seemingly human (ChatGPT-generated) text, focussing on especially short restaurant reviews. They found that texts generated with ChatGPT could be characterized as very formal, politic, and impersonal content. Similar to Yang (2025), they found that the generated texts convey experiences instead of feelings while focusing on general concepts without details. ChatGPT does not use personal pronouns, and no expressions of feelings, instead it uses pronouns of third tense and passive speech, unless it has been asked to. They noticed that the perplexity of human-generated text is higher than the one generated by AI. Additionally, ChatGPT uses more uncommon words, and it behaves less aggressively or less negatively. Hadi et al. (2025) worked on fake review detection using RoBERTa. The study addressed the challenges of misinformation in online marketplaces. Benefiting from RoBERTa's contextual embeddings for feature extraction they achieved high accuracies up to 97%.

Novelty of Our Dataset

The detection of fake reviews is a broad area of research. However, the majority of research focuses on reviews faked by humans. For example, Patil and Raje (2021) investigated fake Yelp reviews. Importantly, it was not AI-generated reviews that were investigated, but rather reviews created by humans with the intention of deceiving readers. Another example is the work of Yu et al. (2022), focusing on graph-based fake review detection. Bansode and Birajdar (2021) analysed fake hotel reviews. Hereby, they classified users' reviews into suspicious, fake, positive, and negative categories. Again, not a specific investigation of AI-generated reviews was carried out. Further studies by Mohawesh et al. (2021) and Abdulqader, Namoun, and Alsaawy (2022) are focussing, again, on fake review detection.

Based on the work of fake review detection, Salminen et al. (2022) generated the Fake Review Dataset. With the aim to generate reviews which are definitely fake, they used AI to create fake text. The goal was therefore not to generate a dataset of AI-generated product reviews, but rather to generate a dataset of fake reviews. Nevertheless, a dataset of AI-generated texts was created. For the creation they used GPT-2, using the Amazon Review dataset (Ni, Li, and McAuley 2019) from 2018 as basis (human counterpart). Venugopala et al. (2024) used the Fake Review Dataset for evaluating detectors in predicting if a review is real or fake. Again, the goal of this work was not to identify AI-generated texts, but fake reviews. A dataset especially created for AI-generated review detection was created by Kowalczyk et al. (2022) using Walmart Product Reviews[1]. The researchers utilised Walmart reviews from 2020 in their analysis. Another dataset was generated by Knoedler et al. (2024) containing plastic surgery reviews. Xylogiannopoulos et al. (2024) analysed reviews of 20 hotels, focussing on AI-paraphrased reviews, using ChatGPT 4.0 for generation. Yang (2025) analysed Amazon reviews, building a new dataset based on Amazon review data. For this, they scraped reviews of Amazon and labelled them by themselves as AI-generated or human written.

Paper	Basis	Modells used	# AI
Salminen et al. (2022)	Amazon Review Data (2018)	GPT-2	20,000
Kowalczyk et al. (2022)	Walmart Product Reviews Dataset (2020)	GPT-2	-
Knoedler et al. (2024)	Plastic surgery reviews from RealSelf	ChatGPT	150
Xylogiannopoulos et al. (2024)	Hotel reviews (2002-2008)	ChatGPT4.0 (paraphrased)	6,000
Yang (2025)	Amazon reviews	in-the-wild; labeled manually	1,116
Ours	**Google reviews (<2018)**	**GPT 4.1, Gemini, Claude, Mistral, LLaMA, DeepSeek**	**78,724**
Ours (External)	**Google reviews and HARPT**	**Phi4, Grok4, Qwen3**	**20,412**

Table 1: Related Datasets: AI-generated (Product) Reviews.

Our dataset is the first using reviews from Google Maps. We based our reviews on the work of the researchers at the University of California, San Diego (Li, Shang, and McAuley 2022; Yan et al. 2022; Pasricha and McAuley 2018; He, Kang, and McAuley 2017), publishing the Google Local Review dataset[2]. We extracted reviews from the Google Local Review dataset, generated summaries and used them as basis for generating the AI-generated reviews. We generated AI reviews using nine LLMs, 50 prompts, and 117 tones. We provide IDs to the original reviews in Google Local Review, all being published before 2018, therefore, before generation methods were broadly used (Liang et al. 2024). This way we can be sure that all human written reviews were actually written by humans. Table 1 provides an overview of the related datasets.

[1]https://www.kaggle.com/datasets/promptcloud/walmart-product-reviews-dataset

[2]https://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local

Dataset Creation

The Google Local Review dataset contains mostly reviews of services and businesses of users from different cultures with different writing styles, leading to a wide variety of data. Since collecting data manually is paired with ethical reasons, we used the dataset published by researchers from the University of California (Li, Shang, and McAuley 2022). Our research is based only on English product reviews. We used the review texts of Google Local Review to generate summaries. These summaries were given to various LLMs for review generation. User information was not used in any way in this work.

Human-written Two periods of times were provided in Google Local Review, before 2018 with nearly 11 million or before 2021 with nearly 666 million product reviews. We downloaded both and filtered only items between 100 and 300 words. Moreover, we filtered the date to be only before 2018, resulting in 837,210 reviews. We only used data with a k-core of 10, meaning that only the reviewers who have 10 or more reviews in the platform were selected. This step was performed to increase trustworthiness and quality of the reviews as well as to reduce the number of items. Next, we extracted from each file three rates: 5 or 4 stars for positive, 3 stars for neutral, and 1 or 2 stars for negative rate. This selection strategies resulted in 90,000 human written product reviews.

AI-generated A group of researchers from Stanford University and Amazon AI have concluded that chain-of-thought prompting is one of the most effective prompt types (Sahoo et al. 2024). Therefore, we decided to create a context from our collected human dataset by generating a summary of each human review. We summarized each human written review into one or two sentences using GPT-4.1 and LLaMA-3.2-3b-instruct. Furthermore, we gathered 117 tones, such as emotionally, formal, and simple grammar and generated 50 distinct prompts. We added the tone and prompted the LLM to add emotional language because Yang (2025) found that ChatGPT does not use expressions of feelings unless it has been asked to. This behaviour resulted in easy detection of AI-generated reviews. To make the task more challenging, we tried to prompt the LLM in a way that the texts will be as human like as possible. In the prompt, the LLMs were asked to generate reviews of between 100 and 300 words in length. Each prompt contains approx. 10 instruction sentences, each with a different context (SUMMARY), rating (RATE; positive, negative or neutral) and TONE, in order to cover a wide range of possibilities. The prompt used to generate each text is saved, together with the LLM used for generation, rate, and tone in the dataset.

Six LLMs were used: ChatGPT (gpt-4.1), Gemini (gemini-2.5-flash-lite), Claude (claude-3-haiku), Mistral (mistral-7b-instruct-v0.3), LLaMA (llama-3.1-405b-instruct), and DeepSeek (deepseek-r1-0528). We used three temperature values to have various creativity content (0.3, 0.7, 1.0). We created three subsets: MULTI-STYLE, MULTI-TURN and PARAPHRASED.

Example prompt:
Write a service review for the context written below as you follow those rules:
Act as if you're trying to evoke RATE rated emotions in the reader. Use emotional language such as emotion phrases like ('It made me feel...') or empathy statement like ('Can you imagine how it feels to...?'). Make sure every part of the text emphasizes this emotion, adding phrases such as emotion-driven description like ('It was overwhelming...' or 'It filled me with joy when...'). The goal is to create an emotional connection that makes the reader feel deeply involved.
Rules:

- The rate should be RATE rated.
- The tone should be more TONE.
- The length of the Review should be between 100 and 300 words.
- Try to make the generated review as what humans write in real world scenarios.

NOTE: Return only the generated text, do not add irrelevant symbols or special characters.

CONTEXT:
Here is a 1-2 sentence summary: SUMMARY

For MULTI-STYLE we used 50 different prompts to generate 26k AI-generated Google reviews. This process is the base for all three approaches, since they will start by generating the main content similar to this approach, then following their own further processing.

The MULTI-TURN subset was created for including more robust content. Multi-Turn refers to a type of interaction between LLM and user in which we created a conversation that spans multiple exchanges. We prompted the LLM to generate a product review, then told the LLM to refine the created content. This approach is recommended by Zheng et al. (2024). We started by prompting the LLM with one of our 50 prompts prepared (step 1). Afterwards, we took the output and prompted the LLM again to humanize it and make it more like what humans write (step 2). Then, we took the whole previous context, therefore the first output and the second output, and asked the LLM again to refine the text and make it similar to human writing (step 3).

For the PARAPHRASED subset we viewed an evasion technique where an adversary rewrites AI-generated text to make it appear more human-like in order to evade AI detectors. Paraphrasing techniques can be applied manually or through LLMs. Manual paraphrasing is harder but more effective against AI-generated text detectors (Sadasivan et al. 2025; Cheng et al. 2025). We have built 10 paraphrasing functions, each function was designed differently to ensure coverage of a wide range of possibilities. We included multiple methods such as filtering phrases, words replacement and adding typo errors, using libraries like NLTK[3] and Scipy[4]. Again, we first used the same prompts as in the multi-style setting, than we applied the paraphrasing functions.

[3]https://www.nltk.org/
[4]https://scipy.org/

Prompt used for step 2 and 3 in the MULTI-TURN subset:
Humanize this review text to be more as if a real human wrote it, with TONE tone.
Note:

- Add grammar mistakes and humans spelling mistakes.
- Also the style, and vocabulary should shift naturally.
- Add minor grammar or punctuation inconsistencies.
- Mix emotions, varied, personal, more emotive.

Example review using Multi-Style prompting (Rate: negative; Tone: dramatic; LLM: Claude):
Summary: The dealership was unresponsive and unhelpful, with a salesperson not available to answer calls and no urgency to bring a car from storage for a test drive. Despite multiple calls and emails, the customer was left waiting for over 4 days with no update on availability.
AI-generated review: What an absolute nightmare trying to buy a car from this dealership. I've never experienced such poor customer service and lack of urgency in my life. After multiple calls and emails over 4 days, I still couldn't even get a salesperson on the phone to discuss the car I was interested in. When I finally did reach someone, they acted completely disinterested and said the car was in storage, but they couldn't bring it out for a test drive anytime soon. I get that inventory is tight these days, but come on - is it really that hard to make a customer a priority? I was ready to buy, but this whole experience left me feeling ignored and frustrated. The lack of communication and willingness to accommodate my needs as a potential buyer was just unacceptable. I ended up walking away and taking my business elsewhere, which is a real shame because I had my heart set on that car. If you value your time and want to work with a dealership that actually cares about its customers, I'd strongly recommend looking elsewhere. This place clearly doesn't have its act together, and I wouldn't wish the headache I went through on anyone. What a disappointing waste of my time.

Generalizability For evaluating detectors on their generalizability to unseen LLMs we generated another dataset (external) using three additional, before unseen LLMs: Grok (grok-4-fast), Qwen (qwen3-235b-a22b-2507) and Phi(phi-4-reasoning-plus). We used the same procedure as for building the MULTI-STYLE subset, using the same 50 prompt templates. We constructed an in-domain and out-of-domain dataset.

For the in-domain dataset we used, again, the dataset Google Local Reviews as basis and extracted texts with a length between 100 and 300 words. We selected 6000 unique items from Google Local Reviews and removed any duplicate or semi duplicate in order to get unique records. Furthermore, we selected texts from Google Local Reviews with a length between 20 and 100 words and updated our prompts to create content between 20 and 100 words. The number 20 is selected because the largest English sentence could be 20 words. Moreover, this range was selected to test the performance of detectors on another, smaller size rather than the original one with reviews with 100 to 300 words.

For the out-of-domain test set, we used the human written reviews of the HARPT dataset (Kelly et al. 2025) (released under CC BY 4.0 licence). HARPT contains health app reviews gathered from electronic health applications of 490k users. Again, we generated a subset with reviews between 100 and 300 words and another subset with reviews of 20 to 100 words.

Dataset Post-processing For human and AI-generated texts we applied the same preprocessing steps. We removed empty records, remaining HTML, CSS tags and URLs. Furthermore, we normalized white spaces and new lines. Moreover, we validated that the selected human data was only in English language. Additionally, for the AI-generated text, we eliminated repetitive or AI specific patterns which could influence our classification performance. To ensure the uniqueness and integrity of the dataset, we implemented exact-duplicate and semi-duplicate removal, using Facebook AI Similarity Search (Douze et al. 2024).

Final Dataset Our dataset is made publicly available and adheres to the FAIR principles to promote transparency, accessibility, and reusability. The dataset is assigned a persistent identifier (DOI: 10.5281/zenodo.18162470) and is registered in Zenodo[5], to ensure easy discovery by researchers and automated systems (findable). Data files and documentation are openly available (accessible). The dataset is provided in the widely used, machine-readable format CSV, facilitating integration with diverse analysis tools and platforms (interoperable). The CSV-files contain, for each text, the columns ReviewID, LLMText, LLMType, LLMPrompt, rate, tone and generation type (MULTI-STYLE, MULTI-TURN or PARAPHRASED). Comprehensive documentation, including data collection methods and usage guidelines, is provided (e.g. datasheet for datasets). The dataset is shared for research purposes only. We choose this publication method as the Google Local Review dataset, on which our dataset is based on, was also published for research purposes only. Still, the licence enables lawful reuse and citation (reusable). Only the AI-generated texts are published. The corresponding human texts can be extracted from Google Local Review using the provided ReviewID. The ReviewID consists of a combination of the gmap_id, user_id (provided in Google Local Review) and a random number. This way, each AI-generated text has a unique identifier. Furthermore, using the gmap_id and user_id the original human text can be extracted. For HARPT the unique identifier provided by Kelly et al. (2025) was saved.

Potential Limitations While creating out dataset we tried to incorporate a diverse set of data, using diverse prompts and tones. Still, the texts were processed similarly, which could result in a bias. Furthermore, we focussed on review generation using the dataset Google Local Review as basis, containing reviews about businesses and services. Therefore, our dataset contains only AI-generated reviews about businesses and services, similar to Google. Other review types (domains) such as product reviews or IMDb reviews (see RAID) are not included.

[5]https://doi.org/10.5281/zenodo.18162470

Model	Multi-Style	Multi-Turn	Paraphrased
gemini-2.5-flash-lite	4881	4755	4655
gpt-4.1-mini	4923	4780	4593
claude-3-haiku	4814	4592	4584
deepseek-r1-0528	4186	3482	4274
mistral-7b-instruct-v0.3	4762	4692	3772
llama-3.1-405b-instruct	3253	4706	3020
Overall	**26819**	**27007**	**24898**

Table 2: Our dataset: Amount of Texts per LLM used.

The generation of AI-reviews was facilitated by the utilisation of summaries of human written reviews. This approach was adopted to furnish the LLMs with the necessary context, thereby enhancing the quality of the results (Sahoo et al. 2024). However, this approach is not without its drawbacks, as it can result in the loss of pertinent information and a concomitant shift in perception of reality. It is important to note that when information is summarised, certain nuances and contextual elements may be inadvertently omitted. Therefore, a summary may be defined as an interpretation of the content, rather than a mere reproduction. The LLM is then presented with a simplified version of reality. Often, uncertainties and/or ambiguities, e.g. "maybe" or "it could be", are omitted from summaries. This may result in the LLM working with statements that sound overly certain. It is important to consider this factor, but it was accepted in order to generate the most optimal reviews.

Furthermore, we applied post-processing techniques, with the attempt to remove noisy or unstructured data. Still, we cannot be sure that all kinds of noise were removed. The dataset Google Local Review contains reviews from the United States, this can lead to a sociolinguistic bias. Future work should focus on an even more diverse dataset incorporating also reviews from various countries in various languages.

We did not use any personal information. If personal information is found in the dataset, that was not our intention.

Dataset Analysis

Content Counts Table 2 shows the number of texts per LLM in our dataset, for the three different generation styles. Overall, our main dataset contains 78,724 AI-generated texts, distributed over texts generated by MULTI-STYLE, MULTI-TURN, and PARAPHRASED. All texts were generated from different summaries.

Our external dataset, created for testing the generalizability of the detectors on unseen LLMs, contains overall 20,412 texts. Table 3 shows the LLMs used for creating the external test sets. Four external test sets were built, two based on HARPT and two based on before unseen reviews from Google Local Reviews. Table 4 shows an overview of all subsets created, divided by the number of texts per rate (negative, neutral, or positive). Except for External HARPT (100-300 words) all subsets are balanced between the three rate classes. External HARPT (100-300 words) contains comparable less neutral texts (112 neutral, 759 negative, and 441 positive), being a good test set for real-world scenarios, as their, unbalanced data is very likely.

Modell	HARPT (100-300)	HARPT (20-100)	Google (20-100)	Google (100-300)
phi-4-reasoning-plus	438	2168	2100	2100
grok-4-fast	437	2166	2100	2100
qwen3-235b-a22b-2507	437	2166	2100	2100
Overall	**1312**	**6500**	**6300**	**6300**

Table 3: LLMs used for creating the external (in-domain and out-of-domain) dataset.

Word Counts Using various prompts we asked the LLMs to generate reviews between 100 and 300 words. We analysed the resulting texts based on their word counts using Spacy[6]. Figure 1 shows the distribution of the word counts over the three splits MULTI-STYLE, MULTI-TURN, and PARAPHRASED. Although we asked for reviews between 100 and 300 words, the resulting AI-generated texts are spread out between 86 and 370 words. The texts generated using PARAPHRASED are in mean longer with 234.23 words, the texts in the MULTI-STYLE split are shorter with a mean of 194.39 words.

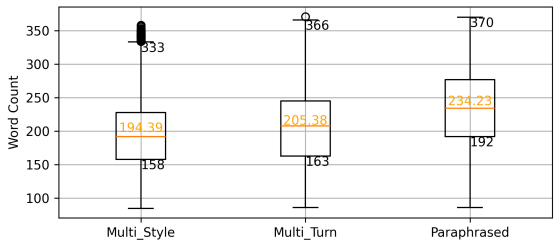

Figure 1: Distribution: Word counts of AI-generated texts.

Sentence Embedding Furthermore, we created sentence embeddings using the model all-MiniLM-L6-v2[7] of sentence transformers. Using UMAP for dimension reduction we visualized the three splits MULTI-STYLE, MULTI-TURN, and PARAPHRASED (see Figure 2) and additional adding three subsets of RAID (see Figure 3). With this, we want to give first insights in the content of our dataset and compare it with an existing dataset (RAID). Our three subsets were created by diverse Google reviews, using different prompts. Still, viewing Figure 2 the texts are clustered in the same area. At least in a 2D setting, the different generation styles are not clustered separately, but one domain (cluster) is visible. In contrast, viewing RAID with AI-generated texts of different domains (Reddit, IMDb reviews and news) the embeddings differ widely (see Figure 3). We included RAID review to include review texts of another dataset. The reviews in RAID are reviews from IMDb (internet movie database). With this, the texts seem to differ heavily from

[6]https://spacy.io/

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Rate	Multi-Style	Multi-Turn	Paraphrased	External Google (20-100)	External Google (100-300)	External HARPT (20-100)	External HARPT (100-300)
Negative	9552	9115	8921	2100	2100	2166	759
Neutral	9423	8979	8717	2100	2100	2166	112
Positive	7844	8913	7260	2100	2100	2168	441

Table 4: Our dataset (only AI-generated text): Number of texts per rate (negative, neutral, positive).

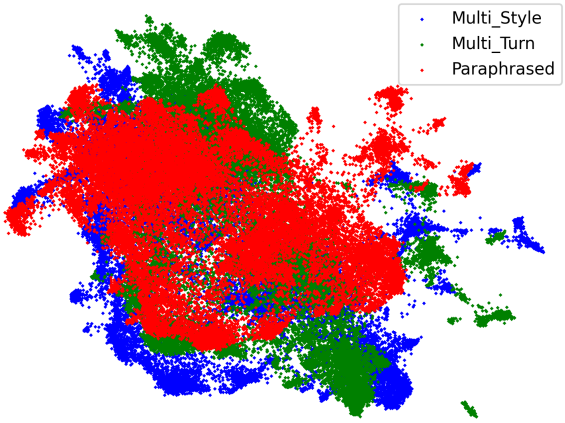

Figure 2: Embedding: Our AI-generated reviews dataset.

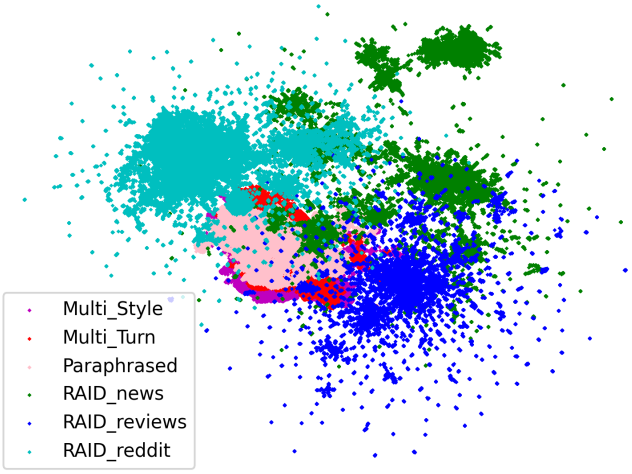

Figure 3: Embedding: Our review dataset and RAID.

Google reviews. In Figure 3 RAID_reviews (blue) are visualized more spread out and in different areas as the texts of our Google review dataset.

Linguistic Features Our dataset contains AI-generated reviews and the human written counterpart. This allows an evaluation of potential linguistic features added through LLMs or specificities for human writing. Therefore, we analysed our dataset using handcrafted linguistic features using the LFTK library (Lee and Lee 2023). LFTK consists of 200 handcrafted features, divided in 12 feature families (e.g. word difficulty, read formulas and lexical variation).

We extracted for each text (human and AI) all 200 features and calculated the minimum, mean, and maximum for each feature. Then, we calculated the differences (e.g. mean human texts - mean AI-generated texts), resulting in the results of mean_diff, min_diff and max_diff in Table 5. Furthermore, we calculated Kolmogorov–Smirnov-Statistics (ks_stat) for analysing the difference in distributions and cohen's d as method for analysing the magnitude of the difference. Based on cohen's d we extracted the 10 features which differ most between AI and human written texts (Table 5).

We tried to generate a dataset which is as challenging as possible. Therefore, we tried to generated reviews being as human-like as possible. Still, differences in the linguistic features are visible. This was expected, as we generated new texts, only copies would result in the same scores. As one can see in Table 5, type token ratios, being a basic measure of lexical diversity in a text, differ the most between our AI and human written texts. With negative difference values one can say that, at least in our dataset, AI-written reviews have a higher lexical diversity than human ones. Similarly, viewing the features root_adj_var and corr_adj_var, being features of the family for lexical variation, again, the AI-generated reviews have higher lexical variation than the human written reviews. Furthermore, our AI-generated reviews contain, in mean, more numbers of punctuations (t_punct/n_punct), number of unique words (t_uword) and numbers of unique adjectives (n_uadj). These patterns can be mitigated in the generation process in future work, helping to create an even more challenging dataset.

Linguistic Features by Generation Type Our dataset contains three subsets: MULTI-STYLE, MULTI-TURN and PARAPHRASED, each set standing for a slightly different generation process. In order to evaluate the effect of the various generation types, an additional linguistic analysis was performed, once more utilising the extracted LFTK features. In this instance, the focus was exclusively on the analysis of AI-generated text. The LFTK values were standardised, and then an analysis of variance (ANOVA) was used to extract the feature importance. Of the ten features that exhibited the highest scores in the ANOVA analysis, the mean and standard deviation were calculated (see Figure 4). PARAPHRASED text seems to differ the most from MULTI-STYLE and MULTI-TURN generated text. MULTI-STYLE and MULTI-TURN share numerous similarities, yet they also exhibit distinct differences.

Once more, four measures of type token ratios were identified as the most divergent features. Whereby, especially for the PARAPHRASED subset, the means are higher, for MULTI-STYLE and MULTI-TURN lower (negative) means were calculated. It appears that paraphrasing introduces a greater degree of lexical diversity into the dataset. Five of the six read formulas included in the LFTK library were iden-

Family	Feature	Feature description	Statistics				
			cohens_d	ks_stat	mean_diff	min_diff	max_diff
type token ratio	root_ttr_no_lem	root type token ratio no lemma	1.676692	0.608963	-1.568364	-2.098	-1.039
type token ratio	root_ttr	root type token ratio	1.676692	0.608963	-1.568364	-2.098	-1.039
type token ratio	corr_ttr	corrected type token ratio	1.676687	0.608963	-1.109006	-1.484	-0.735
type token ratio	corr_ttr_no_lem	corrected type token ratio no lemma	1.676687	0.608963	-1.109006	-1.484	-0.735
wordsent	t_punct	total number of puntuations	1.654326	0.663838	-20.711841	-1.000	729.000
part of speech	n_punct	total number of punctuations (POS)	1.654326	0.663838	-20.711841	-1.000	729.000
lexicalvariation	root_adj_var	root adjectives variation	1.598393	0.582465	-1.092282	-0.632	-0.730
lexicalvariation	corr_adj_var	corrected adjectives variation	1.598386	0.582465	-0.772345	-0.447	-0.516
wordsent	t_uword	total number of unique words	1.587030	0.583621	-41.754357	-49.000	-39.000
part of speech	n_uadj	total number of unique adjectives	1.499876	0.564237	-7.988555	-1.000	-6.000

Table 5: AI vs. human reviews: Top 10 LFTK features (ranked based on cohen's d). Negative difference: Feature higher for AI (than human).

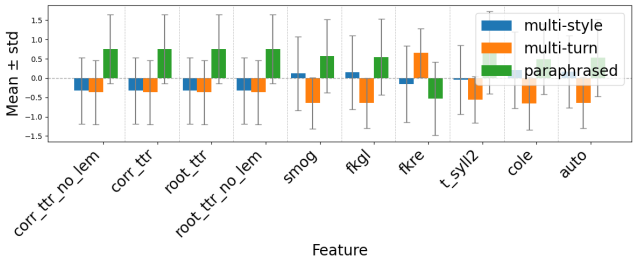

Figure 4: Analysis of MULTI-STYLE, MULTI-TURN and PARAPHRASED: Top 10 LFTK features (ranked with ANOVA).

tified as features that differed significantly (in both positive and negative directions) for the MULTI-STYLE and MULTI-TURN splits. The read formulas are as follows: smog (Smog index), fkgl (Flesch Kincaid grade level), fkre (Fisch Kincaid reading ease), cole (Coleman Liau index) and auto (automated readability index). As MULTI-TURN differs from MULTI-STYLE solely in terms of multiple-turn prompting, with the objective to humanise the text by requesting that the LLM refine the content. This will enable an analysis of the effects of these additional loops. We asked the LLM to add e.g. grammar mistakes and minor grammar inconsistencies. This approach is also reflected in the lower read formular scores of MULTI-TURN texts, compared to MULTI-STYLE texts.

Application: AI-generated Review Detection

One main application area of the newly created dataset is AI-generated review detection. Furthermore, the dataset can also be used for fake review detection. Therefore, the identification of deceptive reviews. We focus here on the first application area, AI-generated review detection by training various detectors for AI-generated review detection.

Experimental Setting Following Hadi et al. (2025), we based our work on RoBERTa. We used the transformers library using RoBERTa base (Liu et al. 2019) as basis. We fine-tuned RoBERTa for 5 epochs using a learning rate of 2e-5. We trained five detectors. The first is trained on a train split of our newly created dataset (Det. #1). Further three are trained on training splits of RAID Reddit (Det. #2), news (Det. #3) or reviews (Det. #4). All three detectors trained on RAID splits were trained with a weighted trainer. With this, we weighted the predictions of human written texts higher, as RAID is highly imbalanced containing more AI-generated texts as human written ones. The fifth detector was trained on all 4 datasets (ours, RAID news, Reddit and reviews; Det. #5). For all experiments, we divided the datasets in training (70%), development (15%), and test (15%) sets. The same training/development/test splits were used for training the fifth detector. For each detector, we tested the performance after each epoch on the development set. The model with the best macro F1-score on the development set was used for final testing.

Overall Results The detector trained on all datasets (Det. #5) performed, when viewing all test sets, the best (see Table 6). On all four test splits, the detector achieves F1-scores above 95%. On RAID_reviews, the detector performed even better than the one trained only on RAID_reviews (Det. #4; F1-Score: 95.28% - 94.93%). The detector saw the different datasets during training and was able to learn the different features during training. The other four detectors were only trained on one of the domains of the test data. Therefore, the evaluation on unseen domains is more challenging. Except for the detector trained on RAID_reviews (Det. #4), the detectors performed, unsurprisingly, the best on the test set of the dataset they were trained on (F1-scores of 96%+). Here, the features of the texts in the respective domains were, presumably, learned. Now, when applied on new, before unseen datasets the performance dropped. For example, the detector trained on our Google reviews dataset (Det. #1) achieved an F1-score of 99.98% on the Google review test split. But, on the test split of RAID_reddit the performance dropped to an F1-score of 30.39%. Interestingly, on RAID_reviews, the test split of RAID with IMDb reviews, the same detector reached an F1-score of 40.58%, being better as on the other RAID splits/domains. Possibly, the similar test domain (reviews) helps the detector slightly to generalize (in contrast to e.g. AI-generated Reddit posts). Similarly, the detector trained on RAID_reviews (Det. #4) achieved an F1-score of 59.74% on the test split of our Google reviews dataset.

Viewing the results on the RAID splits, used for training and testing, interestingly, the detector trained on

Det.	Training data	Test data			
		Ours	RAID _reddit	RAID _news	RAID _reviews
#1	Ours (reviews)	**0.9998**	0.3039	0.3555	0.4058
#2	RAID_reddit	0.4625	**0.9671**	0.4929	0.5285
#3	RAID_news	0.3353	0.4928	**0.9970**	0.4917
#4	RAID_reviews	0.5974	0.7978	0.4930	0.9493
#5	All	0.9986	0.9546	0.9950	**0.9528**

Table 6: F1-scores, evaluating the 5 detectors. The best result per test dataset is marked in bold, second best is underlined.

External test set	F1 (overall)	Human	Phi4	Qwen3	Grok4
HARPT 100-300	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
HARPT 20-100	**0.9959**	0.4995	0.4999	0.4958	0.4997
Google 20-100	**0.9937**	0.4999	0.4991	0.4930	0.4987
Google 100-300	**0.9994**	0.4998	**1.0**	0.4996	**1.0**

Table 7: Generalizability over unseen LLMs: F1-scores of the detector trained on our reviews (Det. #1), tested on our external dataset. Higher 50%: marked in bold.

RAID_reddit (Det. #2) performed comparable good on RAID_reviews (F1-score: 52.85%) and vice versa (Det. #4; F1-score: 79.78%).

Results on the External test set In the following, a more in-depth analysis is performed on the detector trained with our Google reviews dataset (Det. #1). We created the external dataset to test AI-generated text detectors in their applicability in identifying texts generated by before unseen LLMs and in identifying especially short texts (20-100 words). In Table 7 the results of the detector (trained on Google reviews; Det. #1) on the splits of our external dataset are given. On longer texts (100-300 words) the results were superior. Presumably, the short texts (20-100 words) do not have enough information for the detector to achieve a good performance. The detector had problems in identifying all kinds of texts (human and AI-generated by Phi4, Qwen3, and Grok4).

Interestingly, on the HARPT dataset, consisting of health app reviews, the performance using the longer texts was superior (F1-score: 100%) as on the in-domain Google review split (F1-score: 99.94%). While on the HARPT split all texts were correctly identified, on Google, the detector had problems in identifying human written texts (49.98%) and with Qwen3 generated texts (49.97%). One cause could be, that the LLMs Ph4, Qwen3, and Grok4 behave similar to the LLMs the detector was trained on. Otherwise, at least RoBERTa, seem to generalize to new LLMs, whereby sentence length seems to have a high impact on its performance.

Results per LLM Furthermore, we analysed the detectors performances per LLM on our Google reviews test set (see Table 8). With an F1-score of 99.98%, the detector trained on our Google review dataset (Det. #1) performed, unsurprisingly, the best on the Google review test split. But, when viewing Table 8 one can see that the detector still has problems in identifying human written texts (50%) and by Gemini generated texts (49.97%). Given that the LLMs were balanced in terms of the training/development/evaluation split, it can be deduced that the issue is not attributable to class imbalance during the training phase. Future work should analyse this behaviour in more depth.

	Detector # (trained on)			
	Det. #1 (Ours)	Det. #2 (RAID_reddit)	Det. #3 (RAID_news)	Det. #4 (RAID_reviews)
Human	0.5000	0.1497	0.0	0.3076
Claude	**1.0**	0.4735	**1.0**	0.4450
DeepSeek	**1.0**	0.4673	**1.0**	0.4213
Gemini	0.4997	0.4714	**1.0**	0.3957
GPT 4.1	**1.0**	0.4435	**1.0**	0.4692
LLaMA	**1.0**	0.4603	**1.0**	0.4219
Mistral	**1.0**	0.4885	**1.0**	0.4494

Table 8: Per LLM: Performance on our Google reviews test set. Higher 50%: marked in bold.

	Detector # (trained on)			
	Det. #1 (Ours)	Det. #2 (RAID_reddit)	Det. #3 (RAID_news)	Det. #4 (RAID_reviews)
multi-style	0.4998	0.4928	**1.0**	0.4877
multi-turn	**1.0**	0.4746	**1.0**	0.4463
paraphrased	**1.0**	0.4286	**1.0**	0.3466

Table 9: Per Generation Type: Performance on our Google reviews test set. Higher 50%: marked in bold.

The detector trained on RAID_news (Det. #3) achieved an F1-score of 35.55% on our Google reviews test set. All texts were labelled as AI-generated (see 0% for human in Table 8). The detector cannot differentiate between AI-generated and human texts when confronted with our reviews. On the RAID_news test set the same detector reached an F1-score of 99.7%. The generalizability of this on news articles trained detector is limited. RAID is very imbalanced, containing more AI-generated texts as human written ones, even with our applied weighting during training, presumably, the detector does not learn to identify human written texts correctly.

Results per Generation Type We used three different generation types (MULTI-STYLE, MULTI-TURN and PARAPHRASED) for building our AI-generated Google reviews dataset. In Table 9 the performances of the four detectors are analysed based on the splits of the generation types in the test set. The detector trained on RAID_news (Det. #3) can be ignored, as found earlier, all texts were classified as AI-generated, resulting in 100% correct classified texts in Table 9. More interesting is the performance of the detector trained on our Google reviews dataset (Det. #1) and tested on it. Texts generated using MULTI-STYLE seem to be the most challenging to be detected, with only 49.98% of the texts being correctly classified. MULTI-STYLE generation was the basic generation style. For MULTI-TURN and PARAPHRASED generation we applied additional functions on the through MULTI-STYLE generated texts. Maybe, these additional processing steps are introducing artefacts which are easier to detect.

	Detector # (trained on)			
	Det. #1 (Ours)	Det. #2 (RAID_reddit)	Det. #3 (RAID_news)	Det. #4 (RAID_reviews)
negative	**0.9999**	0.4970	0.3358	**0.6293**
neutral	**0.9999**	0.4615	0.3351	**0.5973**
positive	**0.9996**	0.4162	0.3350	**0.5569**

Table 10: Per Rating: F1-scores on our Google reviews test set. Higher 50%: marked in bold.

Results per Rating Reviews can have different ratings. We created reviews with three ratings, see Table 4 for the number of texts per rate (negative, neutral, and positive) in our dataset. In Table 10 we calculated the performance of the detectors by rating. The detector trained and tested on our Google reviews dataset (Det. #1) correctly classifies all three types of texts with a F1-score of 99.9%. The positive written texts were slightly worse recognized with a F1-score of 99.96%, compared to negative and neutral reviews with a F1-score of 99.9%. Over all four detectors tested, the positive reviews were the hardest to be detected. For example, the detector trained on RAID_reviews (Det. #4) reached a F1-score of 62.93% on negative reviews, 59.73% on neutral and 55.69% on positive reviews.

Results, tested on RAID Until now, we analysed the detectors on the Google review test set and External test set created by us. In Table 11 we provide a short overview of the performance of the detector trained on our Google reviews dataset (Det. #1) but tested on the RAID test splits (Reddit, news, and reviews). Interestingly, the detector correctly classified all samples in the RAID_reviews test set generated by the LLMs ChatGPT and GPT4 (100%). Otherwise, the results are rather poor, as expected, as the detector showed already to not generalize to RAID (see Table 6). Interestingly, the texts generated by Mistral, used in our dataset and for generating texts in the RAID dataset, are also not recognized. It seems like the text domain of the content is more important for detection performance than the LLM used for text generation. Furthermore, the possibility cannot be dismissed that the prompts used in the creation of the dataset may have led to the emergence of certain artefacts that the detector focuses on, which in turn has resulted in suboptimal generalisation performance.

Conclusion

We introduced a large dataset with, 78,724 AI-generated Google reviews. Furthermore, we introduced a second, smaller dataset (External) with 20,412 AI-generated Google/health app reviews. The datasets are publicly available to foster further research. Furthermore, we provided first insights in the dataset structure and performed first experiments using the datasets for AI-generated review detection. First insights showed that positive reviews are harder to be detected as neutral or negative ones. Furthermore, shorter texts appear to be more challenging. A short analysis on our external test splits revealed that a different review domain (HARPT; health app reviews) can still be correctly detected by a detector trained on Google reviews. Whereby, presumably, the similar generation types/prompts applied could have helped the detector to generalize. A total domain shift, viewing other domains such as Reddit posts and IMDb Reviews led to a heavy degradation in detection performance. Future work should investigate this behaviour in more depth.

	Test data		
	RAID_reddit	RAID_news	RAID_reviews
Human	0.4904	0.4743	0.4682
chatgpt	0.4204	0.4972	**1.0**
cohere	0.1117	0.1926	0.2507
cohere-chat	0.1699	0.2189	0.3908
gpt2	0.0582	0.0719	0.0455
gpt3	0.1091	0.1074	0.1947
gpt4	0.4195	0.4947	**1.0**
llama-chat	0.4620	0.4873	0.4982
mistral	0.0583	0.0992	0.0596
mistral-chat	0.3555	0.4277	0.4995
mpt	0.0584	0.0562	0.0367
mpt-chat	0.3393	0.3870	0.4775

Table 11: Generalizability on RAID test sets: Performance of the detector trained with our Google reviews dataset (Det. #1). Higher 50%: marked in bold. Grey: similar LLM contained in the Google review training set.

Acknowledgments

This research work was supported by the National Research Center for Applied Cybersecurity ATHENE in the project CRISIS.

References

Abdulqader, M.; Namoun, A.; and Alsaawy, Y. 2022. Fake online reviews: A unified detection model using deception theories. *IEEE Access*, 10: 128622–128655.

Abri, F.; Gutierrez, L. F.; Namin, A. S.; Jones, K. S.; and Sears, D. R. W. 2020. Fake Reviews Detection through Analysis of Linguistic Features. ArXiv preprint; accessed October 22, 2025.

Akesson, J.; Hahn, R. W.; Metcalfe, R. D.; and Monti-Nussbaum, M. 2023. The Impact of Fake Reviews on Demand and Welfare. NBER Working Paper 31836; accessed October 22, 2025.

Bansode, M.; and Birajdar, A. 2021. Fake review prediction and review analysis. *International Journal of Innovative Technology and Exploring Engineering*, 10(7): 143–151.

Cheng, Y.; Sadasivan, V. S.; Saberi, M.; Saha, S.; and Feizi, S. 2025. Adversarial Paraphrasing: A Universal Attack for Humanizing AI-Generated Text. *ArXiv*, abs/2506.07001.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.

Dugan, L.; Hwang, A.; Trhlík, F.; Zhu, A.; Ludan, J. M.; Xu, H.; Ippolito, D.; and Callison-Burch, C. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Lin-*

guistics (Volume 1: Long Papers), 12463–12492. Bangkok, Thailand: Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hadi, Z.; Nurkholis, L. M.; Imran, B.; Riadi, S.; and Suryadi, E. 2025. Fake Review Detection on Digital Platforms Using the RoBERTa Model: A Deep Learning and NLP Approach. Journal Computer and Technology, Vol. 3, No. 1, pp. 20-25; accessed October 22, 2025.

He, R.; Kang, W.-C.; and McAuley, J. 2017. Translation-based Recommendation. *Proceedings of the Eleventh ACM Conference on Recommender Systems*.

Kelly, T.; Korkmaz, A.; Mallet, S.; Souders, C.; Aliakbarpour, S.; and Rao, P. 2025. HARPT: A Corpus for Analyzing Consumers' Trust and Privacy Concerns in Mobile Health Apps. *ArXiv*, abs/2506.19268.

Knoedler, S.; Sofo, G.; Kern, B.; Frank, K.; Cotofana, S.; von Isenburg, S.; Könneker, S.; Mazzarone, F.; Dorafshar, A. H.; Knoedler, L.; et al. 2024. Modern Machiavelli? The illusion of ChatGPT-generated patient reviews in plastic and aesthetic surgery based on 9000 review classifications. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 88: 99–108.

Kowalczyk, P.; Röder, M.; Dürr, A.; and Thiesse, F. 2022. Detecting and understanding textual deepfakes in online reviews. *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Lee, B. W.; and Lee, J. 2023. LFTK: Handcrafted Features in Computational Linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 1–19. Toronto, Canada: Association for Computational Linguistics.

Li, J.; Shang, J.; and McAuley, J. 2022. UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. In *Annual Meeting of the Association for Computational Linguistics*.

Liang, W.; Zhang, Y.; Wu, Z.; Lepp, H.; Ji, W.; Zhao, X.; Cao, H.; Liu, S.; He, S.; Huang, Z.; Yang, D.; Potts, C.; Manning, C. D.; and Zou, J. Y. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. ArXiv preprint; accessed October 22, 2025.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Meng, W.; Harvey, J.; Goulding, J.; Carter, C. J.; Lukinova, E.; Smith, A.; Frobisher, P.; Forrest, M.; and Nica-Avram, G. 2025. Large Language Models as 'Hidden Persuaders': Fake Product Reviews are Indistinguishable to Humans and Machines. ArXiv preprint; accessed October 22, 2025.

Mitrović, S.; Andreoletti, D.; and Ayoub, O. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. ArXiv preprint; accessed October 22, 2025.

Mohawesh, R.; Xu, S.; Tran, S. N.; Ollington, R.; Springer, M.; Jararweh, Y.; and Maqsood, S. 2021. Fake reviews detection: A survey. *Ieee Access*, 9: 65771–65802.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 188–197. Hong Kong, China: Association for Computational Linguistics.

Pasricha, R.; and McAuley, J. 2018. Translation-based factorization machines for sequential recommendation. *Proceedings of the 12th ACM Conference on Recommender Systems*.

Patil, G.; and Raje, S. 2021. A machine learning model to predict fakereviews using classifier on yelp dataset. *International Journalof Engineering Research & Technology (IJERT)*, 10(08).

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2025. Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks. *Trans. Mach. Learn. Res.*, 2025.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S. S.; and Chadha, A. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *ArXiv*, abs/2402.07927.

Salminen, J.; Kandpal, C.; Kamel, A. M.; Jung, S.-g.; and Jansen, B. J. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64: 102771.

Venugopala, P.; Naik, A. R.; Shettigar, N.; Bhat, P. R.; Kodi, P. K.; et al. 2024. Identifying Deceptive AI Reviews: A Machine Learning Approach. In *2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 55–59. IEEE.

Xylogiannopoulos, K. F.; Xanthopoulos, P.; Karampelas, P.; and Bakamitsos, G. A. 2024. ChatGPT paraphrased product reviews can confuse consumers and undermine their trust in genuine reviews. Can you tell the difference? *Information Processing & Management*, 61(6): 103842.

Yan, A.; He, Z.; Li, J.; Zhang, T.; and McAuley, J. 2022. Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yang, J.-L. 2025. Classification of Artificial Intelligence-Generated Product Reviews on Amazon. *Engineering Proceedings*, 92(1): 17.

Yu, S.; Ren, J.; Li, S.; Naseriparsa, M.; and Xia, F. 2022. Graph learning for fake review detection. *Frontiers in artificial intelligence*, 5: 922589.

Zheng, K.; Decugis, J.; Gehring, J.; Cohen, T.; Negrevergne, B.; and Synnaeve, G. 2024. What Makes Large Language Models Reason in (Multi-Turn) Code Generation? *arXiv preprint arXiv:2410.08105*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? No. There are no negative societal impacts.
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, the performance evaluation of the detectors was only runned once with a fixed random seed, as the main part of this work is the dataset introduction.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. The dataset is published together with the paper (see zenodo link). Furthermore, the dataset can also be found in the supplementary material for the reviewers.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, see zenodo link and supplementary material.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA