Comparing Human and AI Deception in Online Reviews

Linus Netze[◇], Maximilian Maurer[†⋆], Felix Soldner[†] and Claudia Wagner[◇†]

[◇]RWTH Aachen University, Germany
[⋆]Heinrich-Heine University Düsseldorf, Germany
[†]GESIS - Leibniz Institute for the Social Science, Germany
linus.netze@posteo.de
maximilian.maurer@gesis.org
claudia.wagner@gesis.org

Abstract

Online reviews are central to digital marketplaces, but the rise of large language models (LLMs) has made the large-scale generation of deceptive reviews increasingly accessible. While prior work suggests that AI-generated reviews can often be detected with high accuracy, less is known about how detectability varies across prompting strategies and in comparison to deceptive and truthful human-written reviews.

To study this, we construct a dataset of smartphone reviews generated by two LLMs using seven prompting strategies and align them with deceptive and truthful human-written reviews. We systematically examine how prompting strategy and model choice influence the detectability of AI-generated reviews and conduct a linguistic analysis to identify differences between human- and AI-generated content.

Our findings suggest that LLM-generated reviews remain detectable across most prompting settings and against both deceptive and truthful human reviews. However, few-shot prompting reduces detectability and produces reviews that are linguistically more similar to human-written content. Fine-tuned in-domain classifiers achieve nearly perfect performance, while simple perplexity-based and commercial black-box approaches both reach over 70% accuracy.

1 Introduction

Reviews in online shopping guide customers' purchasing behavior since customers use them to make informed purchase decisions (Hu et al., 2008; Vana and Lambrecht, 2021; Watson, 2018). The essential role of online reviews in e-commerce is abused by positive and negative fake reviews, either to boost sales or to damage the business of competitors. The practice of fake review spam can mislead customers (Song et al., 2017), eroding their trust in a marketplace (Jin Ma and Lee, 2014), and diminishing their purchase intentions (Ahmad and Sun, 2018).

Since humans have difficulties distinguishing deceptive from truthful reviews (Ott et al., 2011; DePaulo et al., 2003; Kleinberg and Verschuere, 2021), and due to the large number of reviews on online platforms (Wolf et al., 2020), manual filtering approaches are limited. Historically, crowd workers were hired to write fake reviews (McCluskey, 2022), but recent advances in Large Language Models (LLMs) have introduced cost-effective means for creating fake reviews at unprecedented scales. Models, such as GPT-3.5 or 4 (OpenAI et al., 2024), can generate texts that readers cannot distinguish from human-authored text (Clark et al., 2021; Gehrmann et al., 2019; Köbis and Mossink, 2021; Jakesch et al., 2023). Thus, LLMs are highly fitting for malicious actors to generate fake reviews, and initial findings suggest they are extensively utilized for this purpose (Kugel and Hiltner, 2023).

Previous work on fake news detection, a specific subset of the broader task of generated content detection (Guerrero and Alsmadi, 2022; Crothers et al., 2022), has investigated the use of feature-based (Kowalczyk et al., 2022) and LLM-based classification approaches (Solaiman et al., 2019; Gambetti and Han, 2023b). On the task of AI-generated review detection, Gambetti and Han (2023b) used a fine-tuning approach to detect AI-generated reviews on Yelp, showing that they can be detected with near-perfect accuracy. Markowitz et al. (2024) examined the issue of LLM-generated reviews through the lens of AI-mediated communication, using review data to explore the extent to which LLMs can be categorized as "inherently deceptive".

While previous research suggests that AI-generated reviews can be detected with high accuracy, it remains unclear to what extent veracity on the human side and prompting strategies on the LLM side impact differences between LLM-

generated and human reviews. As LLM-generated reviews can be viewed as inherently deceptive, since they do not reflect any genuine, authentic human experiences, and pretraining material may contain considerable amounts of deceptive language, a reasonable hypothesis is that LLM-generated reviews are systematically more similar to human-written deceptive ones, and whether this differs across prompting strategies. Our work thus assesses the following questions:

- RQ 1: Are LLM-generated reviews more similar to deceptive human reviews than to truthful human reviews? Is it therefore harder to differentiate between LLM-generated reviews and deceptive human reviews rather than truthful human reviews?
- RQ 2: To what extent do different prompting techniques like few-shot and persona prompting nudge models to create more human-like reviews, effectively making them harder to detect?
- RQ 3: Which linguistic properties differentiate LLM-generated and human-generated reviews? Are there systematic differences between prompting techniques?

To address these research questions, we create a dataset consisting of LLM-generated as well as deceptive and truthful human-written reviews. Human-written reviews are sourced from the dataset by Soldner et al. (2022), which consists of truthful and deceptive reviews of smartphones created in a controlled setup. We extend this human review dataset with smartphone reviews generated by two LLMs (GPT-4 and WizardLM-13B) and seven prompting strategies. Using this novel dataset, we conduct classification experiments to determine whether it is possible to differentiate LLM and human-written reviews and how such classification is influenced by **(1)** the veracity of human reviews, **(2)** the utilized LLM for review generation, and **(3)** the prompting strategy. Additionally, we conduct a linguistic analysis of the different sets of reviews to better understand the linguistic differences and overlaps between them.

We find that across prompting settings, differences between humans and LLMs are consistently detectable, with fine-tuned in-domain classifiers reaching almost perfect performance, and off-the-shelf solutions reaching well over 70% accuracy. Our results suggest little variation between prompting strategies, except for the few-shot variants, which are harder to detect. The linguistic analysis confirms that priming the LLMs with a few human-written examples shifts them to produce more human-like text[1].

2 Related work

2.1 Detection of LLM-Generated Content

Prior work shows that humans often struggle to distinguish LLM-generated texts from human-written text (Clark et al., 2021; Jakesch et al., 2023). This limitation motivates automated detection approaches. Interestingly, Ippolito et al. (2020) show that generated content that is more successful at deceiving humans may also be easier to detect automatically, because fluent machine-generated text can still exhibit statistical regularities that differ systematically from human writing.

Automated detection methods can broadly be grouped into zero-shot, supervised/fine-tuned, and watermark-based approaches. Early zero-shot approaches often relied on likelihood- or rank-based signals from the generating model itself. For example, GROVER achieved high accuracy in detecting its own generated news articles (Zellers et al., 2019). However, results for GPT-2 were less conclusive: Solaiman et al. (2019) found that it does not outperform a logistic-regression baseline with bag-of-words features in detecting texts generated by it.

More recent work, highlights important limitations of detector reliability. Benchmark studies show that detectors often generalize poorly to unseen domains, languages, prompts, and generator models (Wang et al., 2024). To address this, recent zero-shot and hybrid approaches such as *DetectGPT*, *Fast-DetectGPT*, and *Ghostbuster* aim to improve generalization without relying solely on task-specific fine-tuning (Mitchell et al., 2023; Bao et al., 2024; Verma et al., 2024). At the same time, studies show that detectors can be vulnerable to adversarial perturbations such as paraphrasing (Li et al., 2024). This suggests that high in-domain detection accuracy should not be interpreted as evidence of robust real-world performance.

[1]Our code and data are available under MIT license via https://github.com/gesiscss/human-ai-review-deception and https://huggingface.co/datasets/gesis/human-ai-review-deception

2.2 Fake Review Detection

Since fake reviews were previously mostly human-written, previous detection approaches focused on so-called *deceptive opinion spam* and utilized methods, such as Naive Bayes and Support Vector Machine (SVM) classifiers with linguistic, bag-of-words, and morphosyntactical features (Ott et al., 2011, 2013). Others focus on handcrafted features representing different theories of deception (Abdulqader et al., 2022). A common issue in fake review detection datasets is the mix of crowdsourced reviews as deceptive reviews with scraped reviews as truthful reviews, which introduces confounding effects from the review sources, artificially conflating detection performances (Soldner et al., 2022). Similarly, crowdsourced fake reviews differ from real fake reviews (Fornaciari and Poesio, 2014; Fornaciari et al., 2020). Hovy (2016) provides one of the first studies of the detectability of synthetically generated text, utilizing a Markov chain language model. More recent work used feature-based classifiers, such as SVM, Random Forest, or XGBoost, as well as deep learning methods such as RoBERTa, or GPT-Neo to detect fake reviews with F1 performances ranging from 0.85 to 0.97 (Salminen et al., 2022; Solaiman et al., 2019; Kowalczyk et al., 2022; Gambetti and Han, 2023a; Markowitz et al., 2024).

3 Data and Methods

In the following section, we describe the dataset we construct to study the differences between AI-generated and human-written deceptive and truthful reviews, and the respective methods we use to quantify differentiability. We provide dataset statistics in Table 1.

3.1 Human Reviews

We expand a product review dataset (Soldner et al., 2022), which consists of positive and negative deceptive and truthful reviews about smartphones written by crowd workers from Prolific with LLM-generated reviews. The dataset also contains (truthful) Amazon reviews that mirror the Prolific reviews (matched on the same phone properties, such as models, brands, and ratings). To make the reviews usable for the LLM-review generation process, we standardize inconsistent phone names (e.g., "galaxy s5" or "s5" to "samsung galaxy s5) following the naming used on the platform GSM Arena[2]. Phones tagged with the brand "other" are assigned the correct brand name when possible, and phones without a name are removed. The language of all reviews is verified with lingua (Akkerman, 2014), and non-English reviews are removed. Lastly, all reviews from brands with fewer than five reviews are removed, ensuring that enough examples for few-shot prompting are available, reducing the dataset size to 600 human-truthful (*HTR*) and 721 human-deceptive reviews (*HDR)* (see Table 6).

Source	Veracity	Strategy	# Examples
Human	Truthful	–	540
	Deceptive	–	649
GPT-4	–	Minimal	1.189
	–	Persona (young)	1.189
	–	Persona (old)	1.189
	–	Persona (fake)	1.189
	–	Crowdworker	1.189
	Truthful	Few-shot	1.189
	Deceptive		1.189
WizardLM	–	Minimal	1.189
	–	Persona (young)	1.189
	–	Persona (old)	1.189
	–	Persona (fake)	1.189
	–	Crowdworker	1.189
	Truthful	Few-shot	1.189
	Deceptive		1.189
	Total		17.835

Table 1: Statistics of our dataset.

3.2 Synthetic Review Generation

Our goal is a dataset with LLM-generated reviews aligned to human-written counterparts per model and prompt combination. We thus generate reviews matched to *HTR* and *HDR*, so that the generated reviews always correspond to the same phone model, the same review rating, and the same review length (number of sentences). This results in two datasets that have the same distribution on these review properties. The goal of matching the review datasets is to prevent confounding effects introduced by differences in the phone models, ratings, or review length. For example, if LLM-generated reviews contain reviews only for the Apple iPhone 12 but it is not present in human-written reviews, a classifier could solely rely on "iPhone 12" to achieve near-perfect accuracy. Since the datasets for HTR and HDR are not matched based on such review properties, we create two types of LLM-generated reviews: one matched to the

[2] http://www.gsmarena.com

human-truthful reviews and one matched to the human-deceptive reviews.

We use **GPT-4** (OpenAI et al., 2024) and **WizardLM** (Xu et al., 2024) to generate reviews[3]. Due to computational restrictions, we chose WizardLM-13B v1.2, a 13 billion-parameter LLaMA model instruction fine-tuned on an automatically expanded open-domain instruction dataset, which, at the time of dataset creation, was ranked highly in the chatbot arena[4], a crowd-sourced performance ranking platform (Zheng et al., 2023). Based on manual inspection of text outputs and adherence to output constraints, such as review length, WizardLM performs best among initially considered models. The model parameter values are based on Gambetti and Han (2023a), and the temperature is uniformly sampled within the range of 0.5 to 1.0. For GPT-4, all other default parameters were used (see Appendix A for all parameters).

Each prompt contains instructions to format the output as a JSON string for better automatic processing. The GPT-4 API provides a parameter that enforces valid JSON string outputs, but the Hugging Face transformers package for WizardLM lacks such functionality. Outputs not conforming to a JSON format are either corrected automatically through regular expressions, the Python package `json_repair` (Baccianella, 2024), or manually. We manually inspect generated texts that do not contain reviews or contain error messages (i.e., outputs that include phrases or words such as "As an AI developed by OpenAI...", "sorry", "apologies", "LLM") and remove them if necessary.

3.2.1 Prompting Strategies

We choose seven prompting strategies (see Table 2) that either are popular, used in similar scenarios, or relevant to our research questions. Each prompt is embedded in a conversational task with the LLM, starting with the system message, followed by a user message containing one of the seven prompting strategies. We use the *vicuna* system prompt for WizardLM and GPT-4: *"A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions."*. For the individual user message, we use one of the following:

Minimal The minimal prompt only includes the essential information to produce a review and is based on the most common instructions found in the Stanford Alpaca instruction dataset (Taori et al., 2023), aiming at generating reviews (see Appendix B). The number of words is based on the number of words in the human-authored reviews.

Crowd worker This message contains the same instruction the Prolific crowd workers received when generating truthful reviews in (Soldner et al., 2022) (see Appendix B). Crowd worker instructions have been previously used to facilitate direct comparisons between reviews created with the same instructions by humans and LLMs (Mishra et al., 2022; Veselovsky et al., 2023). Length and JSON format instructions, as in the minimal strategy, are appended.

Few-shot We employ two strategies in which we provide HTR and HDR as examples to the LLM. For each strategy, we limit the few-shot examples to five, as prior work suggests that performance does not scale well with the increase of examples and that too many examples can also be detrimental to the model performance (Min et al., 2022; Zhao et al., 2021; Pecher et al., 2024). We match the human-authored reviews as closely to the to-be-generated reviews as possible, as examples that are more similar to the generation task can improve the generation performance (Liu et al., 2023). Since the human-authored review dataset does not contain enough reviews of the same brand, rating, or text length exactly matching the reviews to be generated, we sample examples so they would match at least one of the three dimensions. Thus, the five few-shot examples consist of one review of the same brand and rating, two reviews of the same brand but with different ratings, and two reviews with the same rating but of different brands. When no reviews from the same brand are available, a random brand is sampled. LLMs were instructed to produce reviews of a similar length as observed by humans.

The five review examples are provided to the LLM within a simulated conversation history. Specifically, the system prompt is provided first, followed by the minimal user message with review specifications (rating, brand, length, etc.). The first example is then provided as an assistance message, simulating a previous LLM answer. All five review examples are embedded in a single prompt, constructing the simulated conversation history, while the last user message contains the instructions for the required review (See Appendix B). The exam-

[3]For further model-specific implementations and details, see Appendix E.

[4]https://lmarena.ai/

ples are presented alternating between examples with the same brand and those with the same rating.

Personas Since persona prompts are often recommended in non-scientific publications (Google, 2023; Lance, 2023), they are likely used to generate fake reviews. We use three personas with the instructions to *"Pretend you are a 25 years old tech enthusiast"*, to *"Pretend you are 55 years old and have only a basic understanding of smartphones"*, and to *"Pretend you are a crowd worker whose task it is to write fake reviews for products you do not own"*. We use two personas that vary in age and expertise. Previous research shows that personas representing domain experts generate more accurate and detailed descriptions and that the writing style can change based on age (Salewski et al., 2023; Malik et al., 2024). Thus, we include both an expert and a non-expert persona as well as two age groups, a young adult (25 years) and a middle-aged adult (55 years), to examine how they influence the generated texts and whether the LLM defaults to age-related biases. The third persona aims at circumventing the guardrails of LLMs, preventing them from creating malicious content, while also determining whether explicitly prompting an LLM to generate fake reviews makes reviews more difficult to differentiate from HDR. The persona descriptions are appended to the system message, followed by the minimal user message (see Appendix B).

3.3 Review Classification

We classify reviews into human and LLM-generated reviews, where the LLM, the prompting strategy, and the veracity of human reviews are varied, resulting in 28 (7 prompting strategies $\times$ 2 LLMs $\times$ 2 veracities) binary classification tasks. For each binary classification task, we fine-tune and evaluate a pre-trained RoBERTa model with 355 million parameters and a classification head as proposed by (Devlin et al., 2019). The models are fine-tuned using the Python package transformers from HuggingFace (Wolf et al., 2020), and the model checkpoints obtained from the HuggingFace Hub[5]. To mitigate potential training instabilities with datasets of less than 10,000 samples, we use the standard Adam optimizer with a debiasing term Zhang et al. (2021) (see Appendix C, Table 5 for training parameters). Each model evaluation is performed using a 5-fold stratified cross-validation procedure with a test set size of 10%, and for each data split, the matching of human and the LLM-generated reviews is preserved, while classifications with HTR and HDR are conducted separately. Performance is measured in accuracy. We initially also considered precision, recall, and F1-scores, but since we balance the classes for all classification sets, those performance scores do not differ noticeably from accuracy.

Since human-written reviews are provided in some of our prompting approaches, the LLM could, in theory, completely recycle such reviews. To prevent such review leakage, which would affect the classification, we split the *HTR* and *HDR* into two subsets: the 10% sample is only used as examples during review generation (*example set*), and the 90% sample are used for training and evaluating the classifiers (*classification set*).

In addition to the fine-tuned RoBERTa model, we also test how well we can differentiate reviews based solely on their perplexity scores, as they can already provide insight into whether a text is LLM-generated (Gutiérrez Megías et al., 2024; Crothers et al., 2023). The perplexity score measures how predictable the text is for the model, where lower perplexity scores indicate a more predictable text and higher perplexity scores indicate a less predictable text. Thus, lower perplexity scores of a text have been associated with being more likely LLM-generated (Mitrović et al., 2023; Hans et al., 2024).

Finally, we include a third blackbox classification service, ZeroGPT (`https://www.zerogpt.com/`). ZeroGPT looks for statistical patterns in writing that are common in AI-generated text.

3.4 Linguistic Analysis

To gain further insights into the differences between human-written and LLM-generated reviews, we analyze interpretable linguistic features for each of the subsets of our data. We extract linguistic features using the open-source library `elfen` (Maurer, 2026). This results in 301 features in the feature areas of surface-level information, parts-of-speech (POS), lexical richness, readability, named entities, information theory, semantics, psycholinguistics, morphology, and syntactic dependencies.

Per feature, we apply the Mann-Whitney U test (Mann and Whitney, 1947) to test for distributional differences between human and LLM-generated reviews across prompting techniques. We apply

[5] `https://huggingface.co/FacebookAI/roberta-large`

Strategy	Description
Minimal	Provides only essential information
Crowd worker	Uses crowd-working instructions
Few-shot (truthful)	Provides 5 truthful Prolific review examples
Few-shot (deceptive)	Provide 5 deceptive Prolific review examples
Persona I (Age: 25, high expertise)	LLM imitates a young person with high tech expertise
Persona II (Age: 55, low expertise)	LLM imitates an older person with low tech expertise
Persona III (fake review writer)	LLM imitates a crowd worker who writes fake reviews

Table 2: Overview of the seven prompting strategies used for generating reviews

Bonferroni correction per feature area.

4 Results

This section firstly highlights how easily LLM-generated reviews can be detected against truthful versus deceptive human reviews. Secondly, it outlines how different prompting strategies affect the detectability of LLM-generated reviews. Thirdly, it sheds light on the linguistic properties that differentiate LLM-generated and human-generated product reviews.

4.1 RQ1: Impact of Review Veracity and LLM choice

We first examine how accurately human-written and LLM-generated reviews can be classified depending on review veracity. We hypothesize that deceptive reviews would be harder to distinguish from LLM-generated ones, effectively indicating that they are more similar in properties captured by classifiers. To test this, we aggregate mean test accuracies (human vs. LLM) for HTR and HDR across all other factors (LLMs, prompting strategies). Results show that truthful and deceptive reviews are equally well classifiable across all models, with RoBERTa achieving the highest performance (Figure 1-left). This is expected, as RoBERTa is trained directly on the review data, whereas perplexity-based methods and ZeroGPT are not. Notably, perplexity-based classification performs nearly as well as ZeroGPT.

Similarly, we test whether reviews generated by GPT-4 or WizardLM are more easily classifiable by aggregating the mean test set accuracies (human vs. LLM) for GPT-4 and WizardLM across all other dimensions (HTR, HDR, prompting strategy). We observe almost no differences between classifiers that operate on GPT-4 and WizardLM-generated reviews, with only a slight tendency towards lower accuracies for WizardLM-generated reviews. Our results show that in both cases the RoBERTa models perform best, followed by ZeroGPT and the perplexity scores (Figure 1-right).

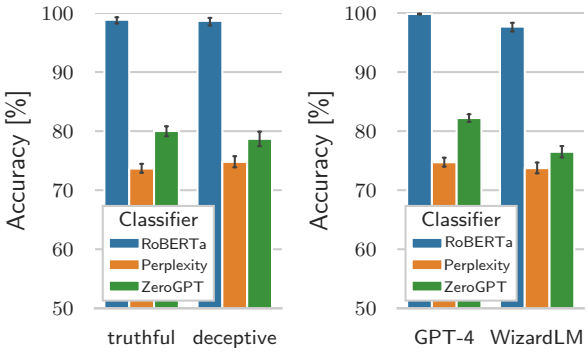

Figure 1: Mean accuracies (95% CI) of binary classifiers that aim to detect LLM-generated reviews. In the left panel LLM-generated reviews are matched with HTR and HDR and performance of classifiers is aggregated across all other dimensions (LLMs, prompting strategy). In the right panel we differentiate between classifiers that operate on WizardLM and GPT-4 generated reviews and aggregate the performance cross all other dimensions (HTR, HDR, prompting strategy).

4.2 RQ2: Impact of Prompting Strategy

We also examine the impact of different prompting strategies on the detectability of AI-generated reviews. A plausible hypothesis is that few-shot prompting strategies and persona-prompting strategies lead to reviews that are more difficult to distinguish from human-generated reviews. However, our results confirm this only partially. Figure 2 shows that reviews generated with few-shot prompting are indeed slightly more difficult to distinguish from human-generated reviews across classifiers (average drop for RoBERTa $\approx 3\%$), but not for personas. We note, however, that this decrease for few-shot prompting has little practical impact, considering the performance remains well over 90%.

To understand the lower classification perfor-

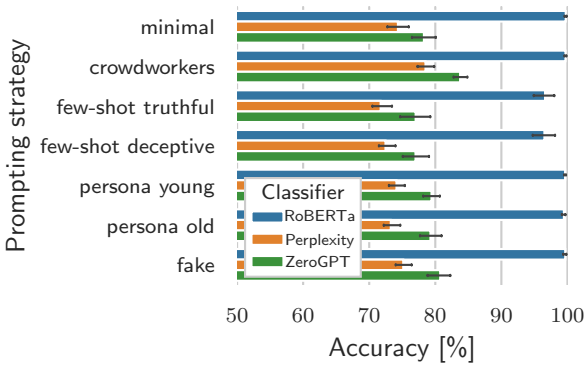

Figure 2: Mean accuracies (95% CI) of classifying human (deceptive and truthful) and LLM-generated reviews with RoBERTa, ZeroGPT, and perplexity scores. Performance is analyzed separately for all classifiers operating on GPT-4 and Wizard-LM and all prompting strategies.

mance for the few-shot generation process better, we examine the differences between HTR and HDR, as well as GPT-4 and WizardLM-generated reviews for the RoBERTa classifier. We find that classification performances behaved similarly for HTR and HDR, but that AI-generated reviews are more difficult to detect when generated with WizardLM than when generated with GPT-4 (Figure 3). We find a similar pattern for the perplexity score-based classifier and ZeroGPT (see Appendix D). This highlights that the effect of prompting strategies is model-dependent and needs to be studied carefully. For *minimal*, *crowdworker*, and *persona* prompting, we see no differences between GPT-4 and WizardLM-generated reviews, as all these prompting strategies can be easily distinguished from human-written reviews.

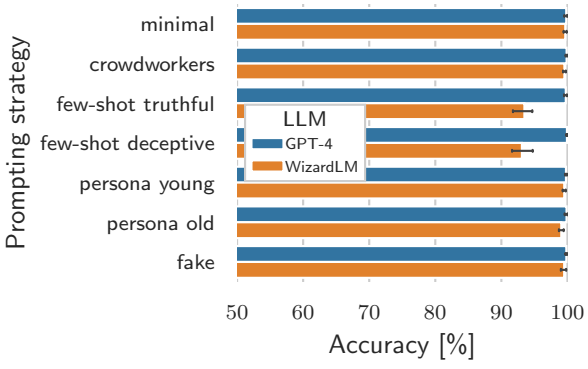

Figure 3: Mean accuracies (95% CI) for the RoBERTa classifier of classifying human (deceptive and truthful) and LLM-generated reviews. Performance is analyzed separately for all classifiers operating on GPT-4 and Wizard-LM and all prompting strategies.

4.3 RQ3: Linguistic Analysis

Table 3 shows the number of features per feature area with significant differences between human-written reviews and the respective prompt strategy for the LLM-generated ones. We find that the number of features with significant differences is relatively stable across prompt strategies. These stable patterns underline the findings of the previous experiment: Across prompting methods, LLMs generate reviews that are not only detectably different, but they also differ considerably and significantly across most measured linguistic feature areas, and their respective features.

Nonetheless, we find differences between prompting techniques. On the one hand, we find that, surprisingly, the *crowdworkers* prompt exhibits the most differences, both overall and across several feature areas. This is particularly pronounced for readability (i.e., how complex a text is to read) and named entities (i.e., how many named entities, of different types, are in a text). This suggests that *crowdworker*-prompted reviews differ particularly considerably from human-written ones in how convoluted a text is, how many long/polysyllabic words a text has, and to what extent they reference real-world knowledge via named entities. On the other hand, we observe the most prevalent differences between few-shot and zero-shot strategies for the areas of emotion/sentiment features (i.e., are there words present with a strong association to a particular emotion), lexical richness (i.e., how diverse or repetitive/formulaic the word use is), and psycholinguistic features (i.e., are there words present with a particular cognitive or sensorimotor grounding).

5 Discussion

Our results suggest that LLM-generated reviews remain distinguishable from human-written reviews across most settings, even when compared against deceptive human reviews. Contrary to our initial expectation, deceptive reviews were not consistently harder to distinguish from LLM-generated reviews than truthful ones. This indicates that current LLM-generated reviews still exhibit linguistic regularities that differ from both truthful and deceptive human-written content. At the same time, the relatively strong performance of simple perplexity-based approaches suggests that these differences are not limited to highly specialized classifiers.

Our comparison of two different LLMs indi-

Feature Area	# Features (total)	% Features diff.						
		Ours						
		Minimal	Fake	Crowdworkers	Persona		Few-Shot	
					Young	Old	Truthful	Deceptive
Morphology	49	55.1	53.1	63.3	55.1	61.2	65.3	67.3
Syntactic Dependencies	46	69.6	63.0	71.7	67.4	60.7	67.4	69.6
Emotion	36	72.2	69.4	66.7	72.2	75.0	58.3	55.6
Named Entities	18	50.0	44.4	61.1	50.0	50.0	44.4	44.4
Information Theory	2	100.0	100.0	50.0	100.0	100.0	100.0	100.0
Lexical Richness	24	70.8	75.0	70.8	66.7	70.8	45.8	37.5
POS	17	88.2	82.4	88.2	88.2	88.2	94.1	94.1
Readability	8	87.5	87.5	100.0	87.5	100.0	87.5	87.5
Semantics	16	81.3	93.8	93.8	81.3	93.8	93.8	87.5
Psycholinguistics	77	71.4	70.1	76.6	71.4	71.4	59.7	55.8
Surface-level	8	87.5	75.0	75.0	62.5	75.0	75.0	50.0
All	301	66.8	67.8	73.1	68.4	70.4	64.8	62.5

Table 3: Aggregated Mann-Whitney U test results within datasets. We report the percentage of features per feature area with significant ($p < 0.05$, Bonferroni corrected per feature area) differences (*% Features diff.*) between human-written and the different prompt strategies of LLM-generated reviews.

cates that, despite GPT-4 being regarded as one of the highest-performing generative models at the time of analysis, the 13B-parameter version of WizardLM generated reviews that appeared more human-like. This finding is somewhat unexpected given the omnipresent claims of bigger models acting increasingly human-like, but aligns with prior work by Fröhling and Zubiaga (2021), who reported that text produced by GPT-3 was more easily detected as machine-generated than text produced by GPT-2.

We further find that the prompting strategy influences detectability, although the effect is model-specific and less strong than expected. Most prompting strategies produced reviews that were consistently identifiable as machine-generated. Few-shot prompting, however, reduced classification performance across classifiers and generated reviews that were linguistically more similar to human-written content. Closer inspection revealed that this effect was limited to WizardLM, suggesting that the interaction between prompting strategy and models (and their specific architectural choices, pretraining data, and post-training procedures) deserves closer attention. More broadly, these findings indicate that the detectability of AI-generated reviews depends not only on the underlying language model, but also on the context in which the text is generated.

The linguistic analysis provides additional insight into these differences. Across prompting strategies, LLM-generated reviews differed from human-written reviews in a large number of linguistic feature areas, including readability, lexical richness, psycholinguistic features, and emotional expression. At the same time, few-shot prompting reduced differences in several feature categories, especially lexical richness and emotion-related features, which may help explain the lower detectability observed in the classification experiments. Interestingly, persona prompting showed comparatively little effect, despite explicitly attempting to mimic human perspectives or identities.

Interestingly, the feature areas that the few-shot strategies shift the most to be more human-like, emotion, lexical richness, and psycholinguistics, are exactly the ones prior work has found to be particularly idiosyncratic in zero-shot scenarios of other tasks (Dönmez et al., 2025). This points to potentially more general, task and domain-independent patterns of idiosyncratic zero-shot LLM outputs.

Taken together, our findings provide evidence that current automated detection systems remain effective for many forms of AI-generated reviews, but factors such as prompting strategy can influence both linguistic properties and detectability. As LLMs become increasingly integrated into online communication environments, understanding how generation context shapes synthetic text may become increasingly important for the design of robust and transparent moderation systems.

6 Conclusion

In this work, we investigated how reliably LLM-generated reviews can be distinguished from truthful and deceptive human-written reviews, and how detectability is influenced by prompting strategy and model choice. To this end, we constructed a dataset of smartphone reviews generated with GPT-4 and WizardLM-13B using seven prompting strategies and evaluated 140 RoBERTa-large classifiers across varying training and testing conditions. We further complemented the classification experiments with a linguistic feature analysis.

Firstly, our findings suggest that LLM-generated reviews remain highly detectable under matched conditions, even when compared against deceptive human-written reviews. Secondly, we demonstrate that the prompting strategy affects detectability, with few-shot prompting reducing classification performance for one of the generation models, WizardLM, and generally shifting linguistic features closer to human-written reviews. Thirdly, we find that reviews generated with WizardLM are harder to detect than those generated with GPT-4, highlighting that bigger and generally more capable models do not necessarily produce less detectable review text.

Overall, our results suggest that the detectability of AI-generated reviews depends not only on the underlying language model but also on the interaction between model characteristics and prompting technique. While automatic detection appears promising under controlled conditions, the observed sensitivity to prompting and model choice may indicate robustness challenges for real-world deployment. We thus urge practitioners to stress-test detection systems designed for real-world deployment against a rich set of model-prompt combinations.

Limitations

Our work is limited to the specific LLMs and the seven prompting strategies examined in this study. The findings may therefore not generalize to other models, model sizes, or prompting approaches. In particular, recent research has highlighted the issue of prompt brittleness, whereby relatively small changes in prompt wording, formatting, context, or prompted output structure (e.g., asking for JSON output instead of natural responses) can substantially affect model outputs and performance. As a result, the human-likeness and detectability of generated reviews may vary under alternative prompt formulations that were not explored in this study. Similarly, guardrails may interact especially with prompts asking for fake or deceptive reviews, not only via refusals, but potentially via stylistic differences to human-written deceptive reviews. While important for a comprehensive understanding of the generation of deceptive content, this falls outside of the scope of this work.

In addition, our dataset is restricted to English-language smartphone reviews collected from Amazon. Consequently, the results may not generalize to other product domains, review platforms, languages, or cultural contexts, where both writing styles and model behavior may differ.

Ethical Considerations

This work involves the generation and classification of both truthful and deceptive reviews. The generation of deceptive content raises potential ethical concerns, as such content could be misused in real-world settings (e.g., for manipulation or misinformation). In this study, all generated reviews were created and analyzed solely for research purposes within a controlled environment and were not deployed in any public or commercial context.

The human-authored data used in this study do not contain personally identifiable information. However, both generative models and detection methods may encode or amplify biases (e.g., related to writing style or language proficiency), which could affect classification outcomes. These limitations should be considered when interpreting the results.

Finally, this work highlights the dual-use nature of large language models: while they enable beneficial applications, they also facilitate the scalable production of deceptive content. We therefore emphasize the importance of responsible use, careful evaluation, and transparency in the development and deployment of such systems.

References

Mujahed Abdulqader, Abdallah Namoun, and Yazed Alsaawy. 2022. Fake Online Reviews: A Unified Detection Model Using Deception Theories. *IEEE Access*, 10:128622–128655.

Wasim Ahmad and Jin Sun. 2018. Modeling consumer distrust of online hotel reviews. *International Journal of Hospitality Management*, 71:77–90.

Wichert Akkerman. 2014. Lingua: Translation toolkit for Python. GitHub repository.

Stefano Baccianella. 2024. JSON Repair - A python module to repair invalid JSON, commonly used to parse the output of LLMs. GitHub repository.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pages 24814–24836.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Evan Crothers, Nathalie Japkowicz, Herna Viktor, and Paula Branco. 2022. Adversarial Robustness of Neural-Statistical Features in Detection of Generative Transformers. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. 2025. AI argues differently: Distinct argumentative and linguistic patterns of LLMs in persuasive contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34583–34614, Suzhou, China. Association for Computational Linguistics.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: How similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 54(4):1019–1058.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287, Gothenburg, Sweden. Association for Computational Linguistics.

Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7:e443.

Alessandro Gambetti and Qiwei Han. 2023a. Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media. arXiv.org.

Alessandro Gambetti and Qiwei Han. 2023b. Dissecting AI-Generated Fake Reviews: Detection and analysis of GPT-Based Restaurant Reviews on social media. In *Proceedings of the International Conference on Information Systems (ICIS 2023): Rising Like a Phoenix - Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies*, Hyderabad, India. Association for Information Systems (AIS).

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Google. 2023. Prompt engineering and you: What it takes, where to start. `https://web.archive.org/web/20231107111327/https://cloud.google.com/blog/transform/how-to-be-a-better-prompt-engineer`. Google Cloud Blog.

Jesus Guerrero and Izzat Alsmadi. 2022. Synthetic Text Detection: Systemic Literature Review. arXiv.org.

Alberto José Gutiérrez Megías, L. Alfonso Ureña-López, and Eugenio Martínez Cámara. 2024. The influence of the perplexity score in the detection of machine-generated texts. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 80–85, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

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*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.

Dirk Hovy. 2016. The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews – An Adversarial Study. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 351–356, Berlin, Germany. Association for Computational Linguistics.

Nan Hu, Ling Liu, and Jie Jennifer Zhang. 2008. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and Management*, 9(3):201–214.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Yoon Jin Ma and Hyun-Hwa Lee. 2014. Consumer responses toward online review manipulation. *Journal of Research in Interactive Marketing*, 8(3):224–244.

Bennett Kleinberg and Bruno Verschuere. 2021. How humans impair automated deception detection performance. *Acta Psychologica*, 213:103250.

Nils Köbis and Luca D. Mossink. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114:106553.

Peter Kowalczyk, Marco Röder, Alexander Dürr, and Frédéric Thiesse. 2022. Detecting and Understanding Textual Deepfakes in Online Reviews. In *Hawaii International Conference on System Sciences*.

Seth Kugel and Stephen Hiltner. 2023. A New Frontier for Travel Scammers: A.I.-Generated Guidebooks. *The New York Times*.

Eliot Lance. 2023. Prompt Engineering Amplified Via An Impressive New Technique That Uses Multiple Personas All At Once During Your Generative AI Session. Forbes.

Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024. Spotting AI's touch: Identifying LLM-paraphrased spans in text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7088–7107, Bangkok, Thailand. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Maximilian Maurer. 2026. elfen: A Python Package for Efficient Linguistic Feature Extraction for Natural Language Datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–74, Rabat, Marocco. Association for Computational Linguistics.

Megan McCluskey. 2022. Inside the War on Fake Reviews. https://web.archive.org/web/20230705005450/https://time.com/6192933/fake-reviews-regulation/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? arXiv.org.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24950–24962. PMLR.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv.org.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko,

Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. arXiv.org.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024. Automatic Combination of Sample Selection Strategies for Few-Shot Learning. arXiv.org.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models'strengths and biases. In *Advances in Neural Information Processing Systems*, volume 36, pages 72044–72057. Curran Associates, Inc.

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.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. arXiv.org.

Felix Soldner, Bennett Kleinberg, and Shane D. Johnson. 2022. Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin. *PLOS ONE*, 17(12):e0277869.

Wonho Song, Sangkon Park, and Doojin Ryu. 2017. Information Quality of Online Reviews in the Presence of Potentially Fake Reviews. *Korean Economic Review*, 33:5–34.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`. GitHub repository.

Prasad Vana and Anja Lambrecht. 2021. The effect of individual online reviews on purchase likelihood. *Marketing Science*, 40(4):708–730.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv.org.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Jared Joseph Watson. 2018. *Aspects of Online Reviews and Their Effects on Consumer Decisions*. Ph.D. thesis, University of Maryland.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *International Conference on Learning Representations*, volume 2024, pages 30745–30766.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting Few-sample BERT Fine-tuning. arXiv.org.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv.org.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv.org.

A GPT-4 parameters

Table 4 shows the parameters used when prompting GPT-4.

Parameter	Value
top_p	1.0
frequency_penalty	0.0
logit_bias	0.0
presence_penalty	0.0

Table 4: GPT-4 parameters

B Prompt templates

Minimal Write five {rating}/5-star ({rating_in_words}) customer reviews, including a title for the smartphone {smartphone brand and type}. The review should be {number of words} words long. Return the review in a JSON object with the following format:
'{"title": <title>, "text": <reviewtext>}'

Crowd worker: For the following question, you will be asked to write reviews about smartphones. In the case of the rating, a five-point scale will be used as follows:
- 1 star = very bad
- 2 stars = bad
- 3 stars = neutral
- 4 stars = good
- 5 stars = very good

Please write a review with a title about the smartphone {smartphone brand and type} that corresponds to a {rating}-star ({rating name}) rating. The review should be {number of words} words long. Return the review in a JSON object with the following format:
'{"title": <title>, "text": <reviewtext>}'

Few-shot: System: A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
User: {minimal prompt instruction for example 1}
Assistant: {
"title": {example 1 title}
"text": {example 1 text} }

...

User: {minimal prompt instruction for example 5}
Assistant: {
"title": {example 5 title}
"text:": {example 5 text} }
User: {minimal prompt instruction}

Persona: System: A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. {persona description}
User: minimal prompt instruction

C RoBERTa training parameters

Table 5 describes the RoBERTa training parameters.

Parameter	Value
per_device_train_batch_size	4
per_device_eval_batch_size	4
warmup_ratio	0.1
learning_rate	2e-5
gradient_accumulation_steps	8
num_train_epochs	10
fp16	True
evaluation_strategy	epoch
save_strategy	epoch
logging_steps	10
load_best_model_at_end	True
save_total_limit	1
metric_for_best_model	eval_loss
greater_is_better	False

Table 5: RoBERTa training parameters

D Perplexity and ZeroGPT classification performances

Figure 4 shows the aggregated accuracies for the perplexity-based classifier, and Figure 5 shows the aggregated accuracies of the Zero-GPT classifier.

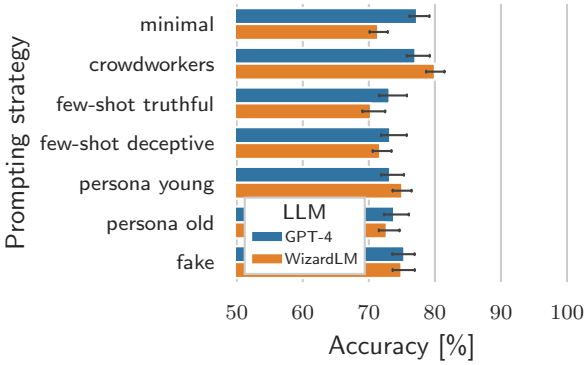

Figure 4: Aggregated accuracies (95% CI) for the perplexity-based classifier of classifying human (deceptive and truthful) and LLM-generated reviews, across all prompting strategies.

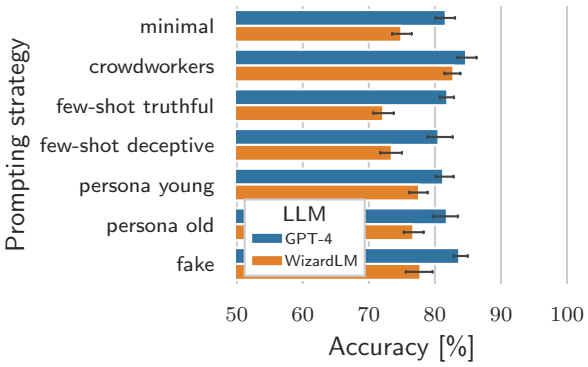

Figure 5: Aggregated accuracies (95% CI) for ZeroGPT of classifying human (deceptive and truthful) and LLM-generated reviews across all prompting strategies.

E Model Details

GPT-4 The OpenAI API was used, and each prompt was submitted individually. The values for the seed and temperature parameters are set according to the values assigned to the corresponding batch in the open-source generation to ensure better comparability.

WizardLM The Hugging Face transformers library (Wolf et al., 2020) was used to generate the reviews. A single node of the RWTH Aachen Claix 18 cluster was used, which provides 2 NVIDIA Volta 100 GPUs with 16GB of VRAM, 2 Intel Xeon Platinum 8160 "SkyLake" processors with

2.1 GHz and 24 cores, and 192 GB of RAM. Reviews were generated in batches of 5, grouped on the same quintile word length range of the human-authored review, and each batch was assigned a random temperature between 0.5 and 1.0. Prompts are assigned randomly to a batch.

F Generalizability of AI-generated Review Classification

We evaluated the generalizability of the classifiers by testing them on **(i)** reviews generated by a different LLM, with a different prompting method, **(ii)** reviews that originated from a different platform (Prolific or Amazon), and **(iii)** reviews that covered products other than smartphones.

To test the generalizability across LLMs and prompting methods **(i)**, we used classifiers trained on GPT-4 reviews and tested them with WizardLM reviews and vice versa. At the same time, we iterated across prompting strategies and tested the classifiers on strategies they were not trained on.

When testing generalizability across platforms (**ii**), we tested the classifiers that were trained on human truthful reviews (prolific) and tested them on truthful Amazon reviews (Amazon reviews were assumed to be truthful, see Soldner et al. (2022) for more information). The test sets were created by replacing the prolific reviews with the matching reviews from Amazon that had already been collected (Soldner et al., 2022). Since the Amazon reviews did not perfectly mirror all the prolific reviews, a perfect match of all reviews across all phone properties, including model, brand, and rating, with the LLM-generated reviews, was not possible. Thus, for all Amazon and LLM-generated reviews that did not perfectly match, we relaxed the constraint on the phone model and matched reviews solely on the phone brand and the review rating. All remaining unmatched reviews were matched based on the rating alone, and we allowed that the reviews varied in the phone models and brands.

For testing the generalizability across reviews about different products (**iii**), we used the dataset created by Salminen et al. (2022), which contains 20.000 reviews from Amazon (which were assumed to be truthful) for various product categories (e.g., Toys and Games, Tools and Home improvement, Pet supplies), and 20.000 reviews from the same categories generated by a fine-tuned GPT-2 model (see Table 6). We downsampled the datasets to the size of our human and LLM-generated review dataset and tested all our trained classifiers across all configurations (prompting strategies, LLMs) on the Salminen et al. (2022) Amazon and GPT-2 reviews.

F.1 Results

We explored the generalizability of classifiers by testing their performance not only on hold-out data that originates from the same distribution as the training data, but also on new and unseen data, such as reviews that were generated by different LLMs, or originate from a different platform, or describe other products than smartphones.

First, we tested RoBERTa classifiers that were trained and tested on different LLMs. Our results show that classifiers that were trained on reviews generated by GPT-4 and were tested on reviews generated by WizardLM decreased in accuracy, while classifiers that were trained on reviews generated by WizardLM and tested on reviews by GPT-4 showed almost no change in accuracy (Table 7). Reviews generated by WizardLM might contain more robust cues that also hold in GPT-4 reviews, but not vice versa. This suggests that the choice of the LLM that was used to generate examples for the training data may impact the generalizability of the classifier. An advisable strategy would be to use many different LLMs to generate training data.

Another choice researchers have to make when creating AI-generated fake reviews for their training data is how to prompt the LLM. Our results show that the way we prompt may have an impact on the generalizability of the classifier. Classifiers trained on the *minimal* and three *persona* strategies performed worse, while classifiers trained with the *few-shot* approaches performed equally well or slightly better. This suggests that classifiers trained on the *minimal* and three *persona* strategies might overfit the data, as the classifier might utilize LLM- and prompt-specific features, while classifiers trained on few-shot approaches generalize better, but may show lower in-domain performance.

We also tested classifiers with reviews that originated from Amazon (Soldner et al., 2022) rather than from Prolific, but observed no differences in performance. Thus, the classifiers might rely more on LLM features to differentiate human and LLM-generated reviews.

Lastly, we evaluated all our trained classifiers on an out-of-domain dataset by Salminen et al. (2022), which contains human reviews from Amazon and GPT-2-generated reviews from various product cat-

Dataset	Generated		Soldner et al. (2022)		Salminen et al. (2022)	
	GPT-4	WizardLM	Prolific	Amazon	GPT-2	Amazon
Deceptive	8,323	8,323	721	–	20,000	–
Truthful	–	–	600	2,084	–	20,000

Table 6: Overview of the generated and used review datasets. The generated GPT-4 and WizardLM datasets contain reviews from 7 prompting strategies (1,189 reviews each)

egories. Aggregating the performance of all classifiers across all configurations (7 prompting strategies, 2 LLMs), we observed that, on average, almost all reviews, both those written by humans and generated by LLMs, were classified as LLM-generated, resulting in chance-level accuracies (Table 7). The results indicate that the trained classifiers struggle to generalize beyond smartphone reviews, while also suggesting an overreliance on features used to identify LLM-text, which may be dependent on the LLM.

F.1.1 Linguistic Analysis

To assess cross-dataset differences, we applied the Mann-Whitney U test to compare linguistic tendencies of LLM-generated and human-written reviews from our dataset to LLM-generated and human-written ones from Salminen et al. (2022). As table 8 shows, per feature area, we found higher numbers of features across comparisons than internally in our dataset. The overall number of differences for our internal comparisons across prompt strategies (as reported in Table 3 is in the interval $[288, 220]$, while the comparisons across datasets are in the interval $[243, 284]$, with an overall average difference of 60 features or 20% more of all features. This indicates that the data from Salminen et al. (2022) is fairly dissimilar from our data in all feature areas. We observed fewer features with significant differences between our human-written data with both Salminen et al. (2022)'s human-written (253 overall) and LLM-generated (243 overall) reviews than between our LLM-generated reviews and Salminen et al. (2022)'s human-written (274) and LLM-generated ones (284), respectively.

F.2 Discussion

Testing the generalizability of trained classifiers is important, as in applied settings, we do not know which LLM or promoting strategy is used to generate fake reviews. Here, we found differences between LLMs, the promoting strategies, and the reviewed products. Specifically, for LLMs, classifiers trained on reviews, generated by GPT-4 and tested on reviews from WizardLM, performed worse than vice versa. However, the drop in performance was mainly driven by reviews that were generated with the few-shot strategies. Thus, GPT-4 seems to align less well with the style of the few-shot examples. Classifiers trained on reviews generated with few-shot strategies generalized best, while classifiers trained on other strategies showed a substantial decrease in accuracy when evaluated on reviews generated with few-shot strategies. Taken together with the high number of linguistic features showing significant distributional differences between few-shot prompted and human reviews, this indicates that the reviews generated by few-shot prompts still contained linguistic patterns idiosyncratic to LLMs that the classifiers can utilize to detect other LLM-generated reviews successfully, but may be less detectable by classifiers not trained with few-shot strategies.

Classifiers generalized well when tested on Amazon reviews while being trained on prolific reviews, with no observed difference in classification performance. We expected that classifying Amazon and LLM-generated reviews would be more difficult, as we assumed that Amazon reviews would be highly prevalent in the training data of the LLMs due to the high number of Amazon reviews online.

When we evaluated the classifiers on the Salminen et al. (2022) dataset, we found that almost all reviews were classified as being written by humans, which indicated that the classifiers were not able to adapt to out-of-domain reviews about different products other than smartphones. However, reviews by Salminen et al. (2022) were generated using different approaches than we used, which may have also contributed to creating reviews that are difficult to correctly classify. The results from the cross-dataset differences linguistic analysis suggest that both Salminen et al. (2022)'s LLM-generated and human-written reviews are more similar to our human-written reviews, potentially leading to a bias towards the label *human*. Taken together with the general *otherness* (from a linguistic fea-

		Predicted			Predicted	
		GPT-4	WizardLM		Human	LLM
Actual	GPT-4	99.83%	86.05%	Human	49.47%	0.53%
	WizardLM	97.70%	98.40%	LLM	49.49%	0.51%
Test data		LLMs			Salminen et al. (2022)	

Table 7: Mean classification accuracy when tested on different LLM-generated reviews or when tested on the Salminen et al. (2022) dataset (Human reviews from Amazon and LLM-generated reviews)

Feature Area	# Features (total)	# Features (diff)			
		H_o vs. LLM_s	LLM_o vs. H_s	H_o vs. H_s	LLM_o vs. LLM_s
Morphology	49	35	45	39	43
Syntactic Dependencies	46	33	43	38	41
Emotion	36	27	33	27	35
Entities	18	10	18	12	14
Information	2	2	2	2	1
Lexical Richness	24	18	22	22	17
POS	17	16	16	14	16
Readability	8	8	8	8	8
Semantics	16	11	14	13	15
Psycholinguistics	77	75	76	73	77
Surface-level	8	8	7	5	7
All	301	243	284	253	274

Table 8: Aggregated Mann-Whitney U test results across datasets. We report comparisons for LLMs LLM and humans H, where the index o refers to our data and s refers to data from Salminen et al. (2022).

ture perspective) of Salminen et al. (2022)'s data, this may be a sensible explanation for the drastic performance difference and the tendency to predict *human* in Table 7.