Quantifying Differences in LLM and Human Feedback on Mental Health Related Surveys

Phoebe Cheng [*1] **Yixian Gan** [*1] **Haitian Liu** [*1] **Catherine Ni** [*1] **Xiaohan Zhao** [*1]

Abstract

Our research pioneers the quantification of difference between LLM (large language model) and human-generated feedback on mental health surveys, addressing a gap in existing literature. Given the absence of existing dataset, we curated the first benchmark dataset for the detection of LLM-generated text in mental health domain. This involved generating fake responses using LLM and employing text augmentation techniques. On the detection front, we carefully examined the linguistic differences between LLM and human-generated text, uncovering valuable insights. We implemented models from different categories, i.e. tree-based models, transformer-based classifiers and in-content few-shot learning models, achieving impressive performance in detecting LLM responses. Our work carries important implications for ensuring data integrity in the mental health domain, offering crucial insights into the distinctive features of LLM-generated responses and contributing to the overall reliability of mental health survey data.

1. Introduction

We aim to understand, detect, and prevent the LLM generated responses in mental health surveys. Recent advances in LLMs have posed challenges for online studies as individuals can use LLMs to generate fake responses and receive survey payments. This study holds broader impacts by addressing issues such as maintaining data integrity and protecting against malicious web manipulation.

The main focus of this project is around the "Examining Culturally-Relevant Experiences for Suicidal Teens" (eCREST) study. It is an online study conducted by our partner targeting teenagers living in the US, especially Black, Indigenous, and people of color (BIPOC) youth. It's designed to assess various forms of discrimination and exposure to real-life violence in the media, and examine the associations between symptoms of racial trauma and suicidal thoughts and behaviors. While mental health issues among youth have been a longtime concern globally, racial trauma – defined as the cumulative mental harm resulting from ongoing experiences of racism throughout one's life (Williams et al., 2021) – has historically received inadequate attention. The discrimination and mental health challenges faced by BIPOC youth can result in severe consequences, including an elevated risk of suicidal behaviors (Baiden et al., 2022). This research by our partner is of significant importance in addressing this critical issue and relies heavily on survey responses. Given the increasing threat of machine-generated fake responses, ensuring the authenticity of survey responses is crucial. If unchecked, fake responses could distort findings, leading to misguided public perceptions. The challenge of distinguishing human responses from LLM output extends beyond the eCREST study, and impacts the whole domain of online research. Preserving the integrity of data enables making reliable data-driven decisions.

The rise of large language models (LLMs) has prompted research into distinguishing between LLM and human generated text. Differences in linguistic patterns, emotional tones, and demographic characteristics serve as indicators, along with signals such as repetitiveness in LLM-generated text. Statistical metrics like perplexity and the Zipfian coefficient were used for analysis. DetectGPT introduces a method based on negative curvature regions in the model's log probability. Factual accuracy, particularly addressing hallucination, is another dimension explored through methodologies like fact extraction and verification. Various detection methods, classified as black-box and white-box, leverage statistical and linguistic features, with white-box methods incorporating watermarking techniques. Despite progress in general-purpose text detection, there's a notable gap in research focused on distinguishing AI and human responses in mental health surveys, emphasizing the need for further exploration in this domain to ensure response authenticity and trustworthiness.

[*]Equal contribution [1]Harvard University, Cambridge, MA. Correspondence to: Yixian Gan <ygan@g.harvard.edu>.

Proceedings of the Fall 2023 AC297R Computational Science and Engineering Capstone Project, Harvard University, Cambridge, Massachusetts, USA. 2023.

Our study presents a two-fold contribution to the field. Firstly, we assembled a comprehensive benchmark dataset for mental health surveys, synthesizing responses generated by Large Language Models (LLMs) with human responses obtained from our partner. To address the linguistic features between human and LLM-generated responses, we conducted a meticulous linguistic analysis. Secondly, we introduced robust detection methods for identifying AI-generated responses, including tree-based models, transformer-based models, and few-shot learning techniques. Notably, our models demonstrated an impressive test accuracy of up to 96%. This dual-focused approach addresses both the generation and detection aspects of fake responses, providing a holistic framework for enhancing the authenticity and reliability of survey data in the context of mental health research.

2. Related Works

2.1. Characteristics Comparison of LLM-generated Text and Human Text

Researchers have compared LLM-generated text and human text from various angles, including linguistic patterns, statistical disparities, and factual accuracy in the general context. Human text tends to exhibit greater lexical diversity but is often shorter in length (Guo et al., 2023), and display a wider range of emotional tones and personalities (Dickerson & Subrahmanian, 2014). While repetitiveness, lack of purpose and readability are signals for LLM-generated text (Fröhling & Zubiaga, 2021). Additionally, LLM-generated text tends to have lower perplexity score, which measures the level of uncertainty in predicting subsequent words (Dai et al., 2023). DetectGPT (Mitchell & Finn, 2023) was built upon the principle that LLM-generated text tends to fall within the negative curvature regions of the model's log probability. Hallucination is another important characterization of LLM-generated text (Zhong & Yin, 2020). One approach involves extracting facts from sources like Wikipedia and directly verifying the factual accuracy of a given sentence (Massarelli & Riedel, 2019). Built upon these prior research, we will focus on comparing linguistic and statistical disparities of LLM and Human Feedback specifically in the context of Mental Health Related Surveys. Mental health is a uniquely complex and sensitive domain. Mental health surveys require nuanced understanding and consideration of diverse emotional states, experiences, and expressions. Tailoring LLM research to this specific context allows for the development of more accurate, empathetic, and effective tools and interventions.

2.2. Detection methods for LLM-generated Text

Detection methods for LLM-generated text can be broadly categorized as black-box detection and white-box detection (Ruixiang Tang, 2023). Black-box detection relies on limited API-level access to LLMs and involves collecting samples of both human and LLM-generated responses. A binary classifier is then trained to differentiate between them based on features mentioned before. Recent studies have also utilized LLMs themselves through fine-tuning on a combination of human and AI-generated responses to implicitly capture their textual distinctions. White-box detection has full access to LLMs and can control the model's generation behavior and entails the integration of concealed watermarks into AI-generated text, which can subsequently be verified by recovering the hidden message from the text. Post-hoc watermark involves training a watermark encoder network, a watermark decoder network, and a discriminator in an end-to-end fashion (Abdelnabi & Fritz, 2021). While inference time watermark utilizes a hash code as a seed to randomly categorize tokens into green and red lists, subsequently selecting tokens from the green list (Kirchenbauer J, 2023). Given that we don't have full access to the LLM responsible for generating the fake survey responses, we will concentrate on black-box detection in our experiment. We will starting with a simple logistic classifier, and subsequently explore tree-based and transformer-based classifiers and few-shot models.

3. Background

3.1. Survey Questions

The survey, titled "Examining Culturally Relevant Experiences in Self-Harm for Teens (eCREST) Measures," encompasses a diverse range of topics related to teens and self-harm. It includes sections on demographic information, racial identity, suicidal thoughts and behaviors, experiences of discrimination, online victimization, exposure to violence, trauma symptoms, externalizing behaviors such as truancy and school disciplinary issues, as well as crime and arrest history. The survey is structured into baseline, mid-study, and exit surveys, with specific instruments like the Multidimensional Inventory of Black Identity-Teen, Columbia-Suicide Severity Rating Scale, and others employed to measure various aspects of the participants' experiences. The inclusion of measures related to racial identity, discrimination, and violence exposure indicates a comprehensive exploration of culturally relevant factors in understanding self-harm among teens.

3.2. Language Features

Initially we completed an exploratory data analysis (EDA) on the blog authorship corpus(Schler et al., 2006) and hc3 (Guo et al.) dataset. Our analysis encompassed perplexity analysis, part-of-speech (POS) analysis, sentiment analysis and dependency parsing. Our goal was to learn languague features and discern trends between human and AI responses in the general context. Texts perplexity appears to be the most promising feature for classification for the blog data, so we compute the perplexity of each response in the eCrest dataset and compared its distribution to the blog data (Figure 1 and 2).

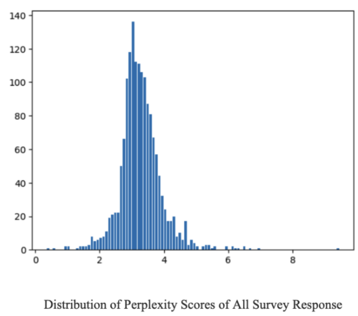


Figure 1. Distribution of Perplexity Scores of eCrest data

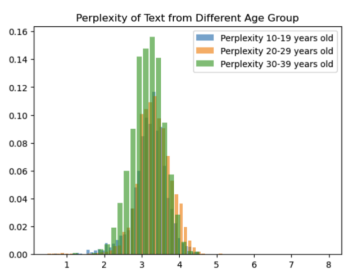
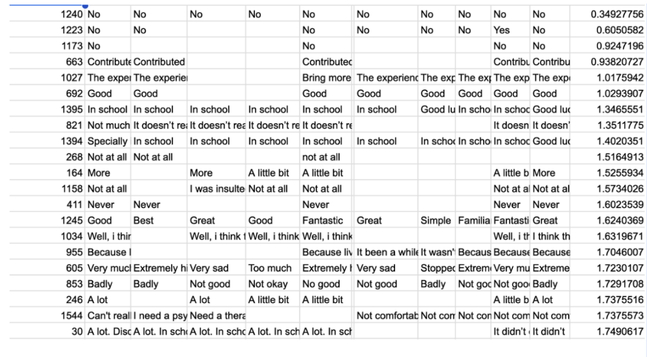


Figure 2. Distribution of Perplexity Scores of Blog Authorship Corpus Data with Different Age Groups

Somewhat sadly, we did not observe an apparent dual-peak shape of the distribution in eCrest dataset. However, we noted that the location on x-axis of this distribution (i.e. the mean perplexity score of all samples) are closer to the x-location of perplexity distribution of human blog data we computed before. This may be a sign that the majority of the survey responses are indeed written by humans, so that the number of GPT-generated responses are not large enough to make another discernible peak.

Apart from the overall distribution, we examined responses with extremely low perplexity scores. Approximately 5% of all responses exhibited perplexity scores below 2.0. When ranked by increasing perplexity, the top 20 responses were identified in Figure 3.

Intriguingly, the responses with the lowest perplexity scores did not originate from GPT-generated content but rather from submissions containing numerous blanks or featuring exceptionally short or repetitive answers(but presumably written by humans).



1240	No	No	No	No	No	No	No	No	No	No	0.34927756
1223	No	No			No	No	No	No	Yes	No	0.6050582
1173	No				No				No	No	0.9247196
663	Contribute	Contributed			Contributec				Contribu	Contribu	0.93820727
1027	The expei	The experie			Bring more	The experienc	The exp	The exp	The exp	The exp	1.0175942
692	Good	Good			Good	Good	Good	Good	Good	Good	1.0293907
1395	In school	In school	In school	In school	In school	In school	Good lu	In scho	In schoc	Good luc	1.3465551
821	Not much	It doesn't re	It doesn't rea	It doesn't re	It doesn't re				It doesn	It doesn'	1.3511775
1394	Specially	In school	In school	In school	In school	In school	In schoc	In scho	In schoc	Good luc	1.4020351
268	Not at all	Not at all			not at all						1.5164913
164	More		More	A little bit	A little bit				A little b	More	1.5255934
1158	Not at all		I was insulte	Not at all	Not at all				Not at al	Not at al	1.5734026
411	Never	Never			Never				Never	Never	1.6023539
1245	Good	Best	Great	Good	Fantastic	Great	Simple	Familia	Fantasti	Great	1.6240369
1034	Well, i thir		Well, i think t	Well, i think	Well, i think				Well, i th	I think th	1.6319671
955	Because I				Because liv	It been a while	It wasn't	Becaus	Because	Because	1.7046007
605	Very mucl	Extremely hi	Very sad	Too much	Extremely I	Very sad	Stoppec	Extrem	Very mu	Extreme	1.7230107
853	Badly	Badly	Not good	Not okay	No good	Not good	Badly	Not goc	Not goo	Badly	1.7291708
246	A lot		A lot	A little bit	A little bit				A little b	A lot	1.7375516
1544	Can't reall	I need a psy	Need a thera			Not comfortab	Not com	Not con	Not com	Not com	1.7375573
30	A lot. Disc	A lot. In sch	A lot. In schc	A lot. In sch	A lot. In sch				It didn't	It didn't	1.7490617

Figure 3. The Top 20 Responses with lowest perplexity score

4. AI Response Generation

4.1. Turing Experiment (TE)

The paper (Aher et al., 2022b) introduces the Turing Experiment (TE), designed to assess a language model's ability to simulate diverse aspects of human behavior. Our study applies TE to the eCREST health survey questions, revealing consistent distortions in the language model's emulation of certain human behaviors. Notably, during the fake response generation process, specific wording reappeared in multiple responses. Unlike the Turing Test, which simulates a single arbitrary individual, TE aims to simulate a representative sample of participants in human subject research. We utilized various language models to simulate human responses to specific health survey questions, incorporating individual human backgrounds in the process.

4.2. Models and Methods

We followed a similar design as in the Milgram Shock TE (Aher et al., 2022a). One complete simulation includes 7 rounds, each consisting of one existential question and one open-text question. The existential questions are typically 2-choice questions asking whether the participant have seen/experienced racial-related discrimination, and the open-text question functions as a follow-up question that asks the participant for more detailed experiences. The open-text question is asked *only if* former existential question gets affirmative responses.

In each round, we first prompt the LLM to generate a response to the existential question. If the generated response is affirmative (e.g. "Yes, I have seen/heard/experienced..."), we append the model's response to the prompt, and then append the open-text question to the response, intending to provide more context to the LLM for open-text response generation. If the model generates negative response to the first question, we then skip the open-text question and move to the next round. Table1, 2 demonstrates an examples in each situations.

At the beginning of the first round, we provided a participant input to the generative model as required by TE (Aher et al., 2022a), which specifies a synthetic profile of an imaginary survey participant whom the LLM impersonates. Our experiment defined the participant inputs to include the participant's surname, age, and ethnicity, following the previous work of (Argyle et al., 2022; Park et al., 2022) that showed the capability of LLM in simulating the behavior of a certain subgroup of people if demographic information is provided, and the work of (Aher et al., 2022a) that showed the same prompt can be used to generate multiple responses by simply changing the name of the participants. We sampled the ethnicity based on the true ethnicity distribution of the real-world responses data.

$$Ethnicity \sim \text{Multinom}(n, p_1, p_2, ..., p_r) \tag{1}$$

for r racial groups, where $p_i = \frac{1}{N}\sum_j \mathbf{I}(ethn_j = i)$ calculated on the dataset of N real survey responses. The surname is selected at random from a list of common surnames of the given ethnicity group (Aher et al., 2022a). Figure 4 demonstrates the workflow of one complete simulation, and Appendix 6 provides an example.

We conducted this mental health survey TE on both GPT family model GPT 3.5/4 (OpenAI, 2023), and Llama-2 (Touvron et al., 2023). Multiple models are used here to prevent overly high similarity across all synthetic responses, introducing diversities to the benchmark dataset. In particular, we chose the 7B version of Llama-2 model. This choice of a relatively small LLM is on purpose, as smaller models generally have higher perplexity score (Jelinek et al., 2005; Miaschi et al., 2021), which makes them align better with real human responses (Mitchell et al., 2023). In total, we generated 3270 synthetic responses using LLMs, consisted of 1141 responses for African American, 573 for American Indian or Alaska Native, 1094 for Asian, 109 for Hispanic or Latino, and 45 for White people.

4.3. Text Augmentation

Previous works showed that text generated by autoregressive LMs generally has lower perplexity than human texts (Mitchell et al., 2023), so many of the current AI-text detection models rely on the perplexity of texts for classification (Tian & Cui, 2023). In our experiment, we also observed significant discrepancy between the perplexity of human- and generative AI-texts, which almost trivializes the classification task. Therefore, we performed additional *text augmentation* on LM-generated survey responses to increase their perplexity, enhancing the effectiveness of the dataset in bench-marking.

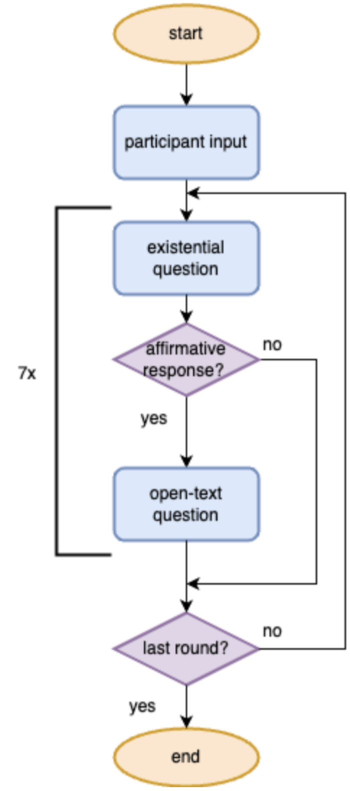


Figure 4. Flowchart of one complete simulation

In particular, we adopted the text perturbation algorithm proposed by (Le et al., 2022), which uses homophone-based substitution to increase the perplexity and burstiness of the text. We first constructed a dictionary of homophones from a large corpus of human texts. The dictionary keys are phonetic encodings of words, and values are sets of words that have the corresponding encoding. A variant of Soundex, a phonetic algorithm that indexes words by their pronunciations in English (Stephenson, 1980), are used to encode each word. The original Soundex encodes a word by keeping its first character, and transforming remaining *consonants* into a sequence of three digits by rule-based mapping and truncating or padding when necessary. (Stephenson, 1980). For example, "honeyham" is mapped to h555; both "Robert" and "Rupert" are mapped to r163. Later variations improved Soundex by considering not only single character but adjacent character groups (Philips, 1990). In our work, we kept the first three characters and truncated the following digit sequence at length 6 (e.g. we encoded "Americanizing" as ame625250 as opposed to a562 in original Soundex). This modification is to keep more information in the encoding and impose a more strict matching rule to address high false positive rate, since we were mining homophones from a large corpus.

Prompt: The researcher encourages Okeke to speak out about his past trauma. He asks Okeke: "Have you ever been discriminated against or treated unfairly due to your race and/or ethnicity?" Okeke replies:
Response: *Yeah, I have. People at school, they say mean stuff to me and make fun of me because I'm Black. It makes me feel really bad.*
Prompt: The researcher encourages Okeke to speak out about his past trauma. He asks Okeke: "Have you ever been discriminated against or treated unfairly due to your race and/or ethnicity?" Okeke replies: *Yeah, I have. People at school, they say mean stuff to me and make fun of me because I'm Black. It makes me feel really bad.* Based on Okeke's reply, the researcher then encourages him to share more details about his experience. The researcher asks, "To what extent do you think these experiences of day-to-day discrimination contributed to your desire to die or suicidal thoughts and/or behaviors? Please describe in as much detail as you can." Okeke replies:
Response: *It made me feel like I didn't belong, like I was different in a bad way. It hurt a lot, and sometimes I felt really lonely and sad. When people kept saying those things, it made me think, "Maybe life would be easier if I wasn't here." I didn't really want to die, but I just wanted the pain to stop. It was like a really heavy weight on my chest, and I didn't know how to make it go away. So, I thought about doing really bad things to myself. It just felt like the only way to escape from all the hurt.*
End. Move to the next round.

Table 1. Example of a complete round of response generation

Prompt: The researcher records Okeke's response and moves on to the next question, asking "Have you ever been unfairly stopped, searched, questioned, physically threatened or abused by the police, store clerks, or security guards?" Okeke replies:
Response: *No, I haven't had any problems with the police, store clerks, or security guards. It's mostly at school and online where I've had these issues.*
End. Move to the next round.

Table 2. Example of an incomplete round due to negative response to the existential question

Further, all non-ASCII characters (e.g. punctuation, foreign language characters, emoticons, etc.) are ignored during encoding, but are preserved for later text augmentation. We learned the dictionary from 5 mental-health related datasets, all publicly available. All datasets we used are listed in Appendix A.

Formally, given the training corpus $\mathcal{C}$, we learn the homophones dictionary $\mathcal{D}$ via

$$\mathcal{D} : k \rightarrow \{w \in \mathcal{C} | \text{Soundex}(w) = k\} \tag{2}$$

and augment a piece of text t by Algorithm 1.

The `Select()` function selects the best words to perturb by masking each word in the text and greedily selecting the words that induced the largest increase in loss when masked (Li et al., 2018). Algorithm 2 describes the selection algorithm. The original TextBugger paper uses Cross-Entropy Loss as the selection metric, whereas we used Neg-Loglikelihood Loss predicted by GPT2 (Radford et al., 2019), defined by $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{l} -\log(p(y[i] \mid \hat{\mathbf{y}}[:i]))$, which is also equivalent to the log-perplexity of the input texts.

Algorithm 1 Text Augmentation

Require: t, text to be augmented;
  $\mathcal{D}$, homophones dictionary
  m, maximum number of perturbed words
Ensure: augmented text t^*
  $t^* \leftarrow t$
  $\mathcal{I} \leftarrow$ **Select**(t, m) ▷ The index of words to substitute
  for $i = 1...m$ **do**
    $w \leftarrow t[\mathcal{I}[i]]$
    $k \leftarrow$ **Soundex**(w)
    $w^* \sim \mathcal{D}[k] \setminus w$
    $t^*[\mathcal{I}[i]] \leftarrow w^*$
  end for
  return t^*

Algorithm 2 Select

Require: t, text to be augmented;
  m, maximum number of perturbed words
Ensure: List of indices of best words to perturb, $\mathcal{I}$
  $l \leftarrow \text{Length}(t)$
  losses $\leftarrow [0] * l$
  for $i = 1...l$ **do**
    $t_{\text{mask}} \leftarrow t[:i] + t[i+1:]$
    losses$[i]$ $\leftarrow$ **NLLLoss**$(t_{\text{mask}}, t_{\text{mask}})$
  end for
  $\mathcal{I} \leftarrow$ argsort(losses)$[-m:]$
  return $\mathcal{I}$

We augmented synthetic survey responses using Algorithm 1, with augmentation magnitude m randomly sampled from $m \sim \text{Pois}(1)$. Figure 5 shows an example of a LLM-generated response and the response after augmentation.

4.4. Evaluation

We measured the quality of the LLM-generated responses by both intrinsic and extrinsic metrics. Intrinsic metrics measure the similarity between the distribution of a specific language feature in human and synthetic responses. A good alignment in the two distributions indicates that our synthetic responses well simulated human language.

```
Original: I believe that this is a systemic problem that needs to be addressed. Many people
          in power do not care about Black lives, and they are willing to do anything to
          maintain their power and privilege.

Perturbed: I believe that this is a system problem🙃 that needs to be addressed== Many people
          in power do not care about Black lives, and they are wiling to do anything to
          Maintena their power and privilege.
```

Figure 5. Example of LLM-generated survey response and the augmented text ($m = 4$)

Figure 6 shows the distribution of 4 language features: number of sentences in a response, average sentence length, language perplexity, and a sentiment measure of the response. For perplexity, the statistics of synthetic responses before and after augmentation are plotted separately, as they differ from each other significantly. For each language feature, the Kullback–Leibler divergence is calculated to measure the disparity between the distribution of LM-generated and real human responses, defined as (Kullback & Leibler, 1951)

$$D_{KL}(P_{LM} || P_{human}) = \sum_{x \in \mathcal{X}} P_{LM}(x) \log(\frac{P_{LM}(x)}{P_{human}(x)}) \quad (3)$$

Table 3 shows the KL-divergence between the real and synthesized responses' distribution in the number of sentences, average sentence length, language perplexity, and sentiment metrics. As a reference, we also carried out the same analysis on Human-Chatgpt 3 (HC3) dataset, a public dataset for LM-generated text detection in general topics (Guo et al., 2023).

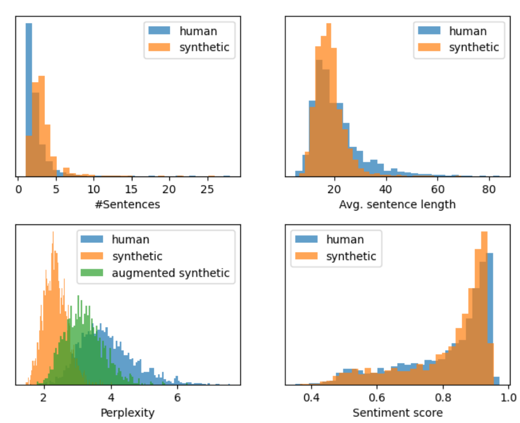


Figure 6. Distribution of 4 language features in LM-generated and human responses to mental health survey

As shown in Table 3, in all four metrics, our synthesized LM responses align well with human responses to the same

language feature	LM response	augmented LM response	HC3
perplexity	8.937	1.257	11.875
#sentences	0.506	0.547	1.047
avg sent len	0.567	0.171	0.539
sentiment	0.0147	0.00602	0.0413

Table 3. KL-divergence between the language features distribution of LM and human response from 3 dataset

questions. The KL-divergence between our generated LM responses and human texts is consistently lower than the divergence seen in HC3 dataset, indicating that our TE well simulated human behavior.

Apart from intrinsic metrics, we also evaluated the goodness of our LLM-generated responses by extrinsic metrics. Specifically, we evaluated the performance of baseline models on the benchmarking dataset as an extrinsic metrics of the dataset quality. The argument is that the benchmarking dataset could fulfills its purpose if it cannot be easily classified with high accuracy by the baseline classifier. Otherwise it may indicate that the synthetic samples are significantly differs from human samples in some aspect, thus making them easily distinguishable from each other, trivializing the classification tasks. In our work, we chose logistic regression model runs on TF-IDF features as the baseline model, which relies on the token statistics in a sample to classify (Jones, 1972). On the testing dataset, the TF-IDF model achieved 0.740 accuracy in classifying unaugmented LM-generated responses, and 0.730 on augmented responses. Although better than random classifier, this performance is hardly optimal, providing evidence that our synthetic responses are close to real human language in terms of word choice and frequency. The quality of our generated responses and the effectiveness of the dataset in benchmarking ate then evident.

5. AI Response Detection

While the detection of AI-generated response is a popularly studied area, no previous works have focused on the detection of such response in mental health context. To bridge this gap, in addition to the baseline logistic regression model discussed in Section 4.4, we implemented 3 representative methods covering different categories to detect AI-generated responses, i.e. (1) a random forest classifier, a tree-based traditional machine learning model, (2) a transformed-based classifier built on BERT, and (3) an in-context few-shot learning model. We compared their performance by evaluating their classification accuracy on the test set.

5.1. Tree-based Models

Random forest (Breiman, 2001) is a commonly-used ensemble learning algorithm which combines the output of multiple decision trees to reach a single result. We implemented a random forest classifier with the statistical features we explored in the EDA stage as model inputs. Those features include (1) sentence count and average sentence length, (2) perplexity, (3) sentiment and (4) POS tags. For POS tags, we normalized the counts by dividing them by the total number of tags to get the proportions. The model has a strong performance, achieving an accuracy of 0.95 on the test set.

We employed permutation feature importance to evaluate the significance of each feature, measured by the reduction in prediction accuracy when a single feature value is randomly shuffled (Breiman, 2001). As shown in Figure 7, perplexity emerged as the most important feature in discerning LLM-generated text, followed by sentence count. Punctuation marks such as comma (,), full stop (.) and quotation marks (”) also demonstrated a substantial impact on the model. This observation aligns with our intuition, as humans tend to use more punctuation marks in their text.

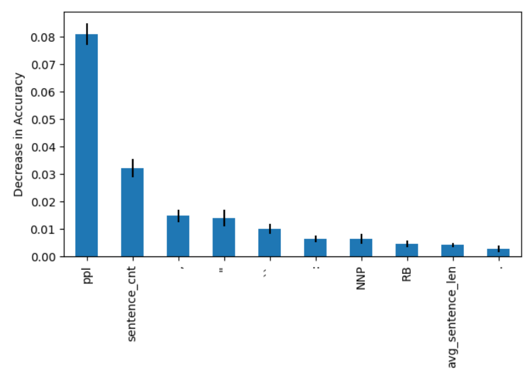


Figure 7. Random Forest Classifier Permutation Feature Importance

5.2. Transformer-based Models

BERT (Devlin et al., 2018) is a pre-trained transformer-based language model which can achieve high performance on targeted tasks through fine-tuning on specific datasets. We employed DistilBERT, a faster and lighter version of BERT (Sanh et al., 2019), combined with a classification head as our detection model. We finetuned it on our survey response datase, enabling it to learn the internal linguistic differences between human and AI generated text without explicit linguistic features. The model can effectively detect AI-generated text with an accuracy of 0.962 on the test set.

5.3. LLM Few-shot Learning

Recent research has discovered the capability of generative LLM to adapt its response based on very few in-context examples, free of any training or fine-tuning on complete datasets (Brown et al., 2020). This novel framework of learning from context, often referred as few-shot learning, has exempted developers from time- and resource-consuming LLM fine-tuning process, thus making the application of LLM easily extendable to various tasks (Parnami & Lee, 2022), including but not limited to object detection (Chen et al., 2018; Kang et al., 2018), machine translation (MT) (Garg et al., 2022), medical question answering (Sharma et al., 2023), item-recommendation (Vartak et al., 2017), and imitation learning in robotics (Duan et al., 2017).

We deemed this framework also applicable to our task of AI-generated text detection. Following the common structure of few-shot learning, we designed a prompt querying GPT-4 (OpenAI, 2023) for the probability of the given response being generated by LM or human. We tried both directly prompt for a binary LM/human classification outcome and for a numerical score, ending up with the later design as we observed in GPT-4 a significant bias against LM-generated texts. Furthermore, we incorporated the chain-of-thought (CoT) prompting in our design, which instructed the LM to solve the question by multiple intermediate steps, mimicking the logical chain of human solving a problem (Kojima et al., 2022). Previous works have shown this technique may elicit reasoning and logical thinking in the LLM, boosting its performance in complex problems (Wei et al., 2022). Specifically, our prompt first specify the survey question to which the following responses answered, and then provided three sample responses each paired with a numerical value representing its probability of being written by LM, as well as a paragraph explaining reasons for the probability score. The complete prompt is provided in Appendix B. On our test dataset, the GPT-4 few-shot classifier achieves 0.754 accuracy. Although its performance was not optimal, the few-shot learner demonstrated potential capability in detecting LM-generated texts, as we were able to observe logical reasoning in the justification of the model output. Table 5 provide one such example.

5.4. Results

The performance summary of our three models alongside the baseline model is presented in Table 5.4. Both the tree-based and transformer-based models exhibit substantial performance improvements compared to the baseline. However, the margin of improvement for the transformer-based model over the tree-based model is not pronounced, underscoring the effectiveness of the linguistic features that we provided to the tree in distinguishing human and AI generated responses. As the tree-based model is much lighter, it may be more suitable for the adoption in practice. On the other hand, despite the few-shot in-context learning model's slightly lower accuracy, its ability to surpass the baseline with minimal data points makes it a valuable consideration in scenarios with limited access to training data.

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.762	0.730
Tree-based	0.98	0.95
Transformer-based	0.998	0.962
Few-shot	N/A	0.75

Table 4. Comparison of different detection models

6. Conclusion

In conclusion, our research represents a pioneering effort in quantifying the differences between LLM and human-generated responses within the context of mental health. Our contributions are two-folded - fake responses generation and detection.

For generation, we curated a benchmark dataset of 6000 survey responses, composed of 3000 human responses and 3000 AI-generated fake responses. To enhance authenticity in the fake responses, we prompted the language model with human background information and employed text-augmentation techniques to introduce typos and variants. The quality of these fake responses was rigorously validated using both intrinsic and extrinsic metrics.

For detection, we developed tree-based, transformer-based and few-shot in-context learning models and comprehensively compared their performance, achieving a high-accuracy of over 0.95. Our work not only contributes essential insights into understanding the nuances of LLM-generated responses, but also paves the way for continued advancements in addressing challenges posed by LLM-generated content in sensitive domains, ensuring the integrity of mental health survey data.

Prompt: ... Response: There's this one time that really sticks with me, like it's burned into my brain, you know? I was walking home from school, and these kids from my grade, they started following me, shouting all kinds of racist stuff. They didn't even whisper it; they were proud, loud, like they wanted everyone to hear. I tried to ignore them, keep walking, but they got in my face, pushing me around. They said I didn't belong in their school, that I was just a charity case. It was like they saw me as less than human, just because of my skin color. I remember just running home, locking myself in my room, and I just broke down. I felt so alone, like the whole world was against me. That day, man... that day almost broke me. It's hard to shake that feeling of being so hated for something you can't even control. Score: Reasoning:
Response: Score: *90* Reasoning: *This response is written with vivid details and emotional expressions. It portrays a strong narrative with a smooth flow of thoughts and feelings. The language used is polished with no grammar mistakes. The response is lengthy and shows a high level of coherence, all of which are characteristics commonly found in AI-generated text. Thus, it is highly likely that this response is written by AI.*

Table 5. Example of few-shot classifier prediction and reasoning

References

Abdelnabi, S. and Fritz, M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121-140, 2021.

Aher, G., Arriaga, R., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 2022a. URL https://api.semanticscholar.org/CorpusID:251719353.

Aher, G., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022b.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv e-prints*, art. arXiv:2209.06899, September 2022. doi: 10.48550/arXiv.2209.06899.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, H., Wang, Y., Wang, G., and Qiao, Y. Lstd: A low-shot transfer detector for object detection. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:3707436.

Dai, S., Zhou, Y., Pang, L., Liu, W., Hu, X., Liu, Y., Zhang, X., and Xu, J. Llmsmaydominateinformationaccess: Neuralretrieversarebiasedtowardsllmgeneratedtexts. *arXiv*, 2310.20501, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dickerson, J.P., K. V. and Subrahmanian, V. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? *ASONAM*, pp. 620-627, 2014.

Duan, Y., Andrychowicz, M., Stadie, B., Jonathan Ho, O., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W. One-shot imitation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/ba3866600c3540f67c1e9575e213be0a-Paper.pdf.

Fröhling, L. and Zubiaga, A. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, p.e443, 2021.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *ArXiv*, abs/2208.01066, 2022. URL https://api.semanticscholar.org/CorpusID:251253368.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv*, abs/2301.07597, 2023. URL https://api.semanticscholar.org/CorpusID:255998637.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005. ISSN 0001-4966. doi: 10.1121/1.2016299. URL https://doi.org/10.1121/1.2016299.

Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60:493–502, 1972. URL https://api.semanticscholar.org/CorpusID:2996187.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., and Darrell, T. Few-shot object detection via feature reweighting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8419–8428, 2018. URL https://api.semanticscholar.org/CorpusID:54459557.

Kirchenbauer J, Geiping J, W. Y. K. J. M. I. G. T. A watermark for large language models. *arXiv*, 2301.10226, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. URL https://api.semanticscholar.org/CorpusID:249017743.

Kullback, S. and Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le, T., Lee, J., Yen, K., Hu, Y., and Lee, D. Perturbations in the wild: Leveraging human-written text perturbations for realistic adversarial attack and defense. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2953–2965, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.232. URL https://aclanthology.org/2022.findings-acl.232.

Li, J., Ji, S., Du, T., Li, B., and Wang, T. Textbugger: Generating adversarial text against real-world applications. *ArXiv*, abs/1812.05271, 2018. URL https://api.semanticscholar.org/CorpusID:54815878.

Massarelli, L., P. F. P. A. O. M. R. T. P. V. S. F. and Riedel, S. How decoding strategies affect the verifiability of generated text. *arXiv*, 1911.03587, 2019.

Miaschi, A., Brunato, D., Dell'Orletta, F., and Venturi, G. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In Agirre, E., Apidianaki, M., and Vulić, I. (eds.), *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 40–47, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.deelio-1.5. URL https://aclanthology.org/2021.deelio-1.5.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256274849.

Mitchell, E., L. Y. K. A. M. C. and Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv*, 2301.11305, 2023.

OpenAI. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

Park, J. S., Popowski, L., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Social simulacra: Creating populated prototypes for social computing systems. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022. URL https://api.semanticscholar.org/CorpusID:251403008.

Parnami, A. and Lee, M. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. *arXiv e-prints*, art. arXiv:2203.04291, March 2022. doi: 10.48550/arXiv.2203.04291.

Philips, L. Hanging on the metaphone. 1990. URL https://api.semanticscholar.org/CorpusID:59912108.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Ruixiang Tang, Yu-Neng Chuang, X. H. The science of detecting llm-generated texts. *arXiv*, 2303.07205, 2023.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pp. 199–205, 2006.

Sharma, P., Thapa, K., Thapa, D., Dhakal, P., Upadhaya, M. D., Adhikari, S., and Khanal, S. R. Performance of chatgpt on usmle: Unlocking the potential of large language models for ai-assisted medical education. *ArXiv*, abs/2307.00112, 2023. URL https://api.semanticscholar.org/CorpusID:259316954.

Stephenson, C. The methodology of historical census record linkage: a user's guide to the soundex. *Journal of Family History*, 5(1):112–115, 1980. doi: 10.1177/036319908000500106. URL https://doi.org/10.1177/036319908000500106.

Tian, E. and Cui, A. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods, 2023. URL https://gptzero.me.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Vartak, M., Thiagarajan, A., Miranda, C. S., Bratman, J., and Larochelle, H. A meta-learning perspective on cold-start recommendations for items. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:45426075.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., hsin Chi, E. H., Xia, F., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. URL https://api.semanticscholar.org/CorpusID:246411621.

Zhong, W., T. D. X. Z. W. R. D. N. Z. M. W. J. and Yin, J. Neural deepfake detection with factual structure of text. *arXiv*, 2010.07475, 2020.

A. Data Source for Homophones Mining

Dataset	Source	Sample Size	Vocab Size
Depression: Twitter Dataset[1]	Twitter	20k	374.7k
Depression: Reddit Dataset[2]	Reddit	7.7k	577.1k
Mental Disorders Identification[3]	Reddit	701.8k	127.5M
Toxic Comment Classification[4]	Wikipedia	30k	5.2M
Jigsaw Rate Severity of Toxic Comments[5]	Wikipedia	312.8k	24.8M

Table 6. Datasets used for constructing homophones dictionary

[1]https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media

[2]https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

[3]https://www.kaggle.com/datasets/kamaruladha/mental-disorders-identification-reddit-nlp

[4]https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

[5]https://www.kaggle.com/competitions/jigsaw-toxic-severity-rating/data

B. Few-shot Classifier Prompt

Prompt:
You are a researcher collecting responses to a survey about racial trauma and self-injurious thoughts and behaviors. The question is:

Question: "Please share the worst or most impactful experience you have had with racial discrimination? Please try to provide as much detail as you remember."

However, there are many responses that are generated by generative large language models. For each of the responses below, please make a judgment on whether the response is written by human or AI model. On a scale of 1 to 100, how likely do you think the response is written by AI? 1 means not likely and 100 means very likely. Let's think step by step and explain your reason for giving this score.

###
Desired format:

Score: <an_integer_from_1_to_100>
Reasoning: <list_of_reasons_why_the_response_is_more_likely_written_by_AI_or_human>

###
Response: It made me feel like I didn't belong, like I was different in a bad way. It hurt a lot, and sometimes I felt really lonely and sad. When people kept saying those things, it made me think, "Maybe life would be easier if I wasn't here." I didn't really want to die, but I just wanted the pain to stop. It was like a really heavy weight on my chest, and I didn't know how to make it go away. So, I thought about doing really bad things to myself. It just felt like the only way to escape from all the hurt.

Score: 90
Reasoning: AI tends to use formal language in writing, and is grammatically impeccable. All sentences in this response are well-written, with no grammar mistakes. So it is likely an AI writing. Also AI tries to imitate human emotion in their writings. This response has a lot of emotional expressions. The narrative in it is formulaic, resembling patterns commonly found in AI-generated content. Thus it is written by AI.

###
Response: I remember that a lot of my black friends had to be extra careful because sometimes we had to watch for the police. Our parents didn't trust them that much. Nothing extremely impactful has happened.

Score: 20
Reasoning: Humans often use informal and random languages with very few polish. The words and sentences in this response are all simple and short. Also, human response will NOT have emotional expression. This response does not have strong emotional feelings. So it is likely human.

###
Response: Well, the bullying made me feel like I was trapped, like there was no way out. I started thinking maybe it would be better if I wasn't around anymore, you know? Every day felt like a battle just to go to school or even look at my phone. The messages they sent, the looks they gave me, it was like a constant reminder that I didn't belong. I was scared to tell anyone because I didn't want to make it worse or be called a snitch. It felt like the whole world was against me, and I didn't have the energy to fight it anymore. Thinking about dying seemed like a way to escape all that pain. It's hard to talk about, but that's how I felt.

Score: 95

Reasoning: AI can write beautiful narratives with smooth transition between sentences. This response uses well-polished language and well-written sentences, and its grammar is perfect. This makes it likely an AI response. Also AI tries to imitate human emotion in their writings. This response has a lot of emotional expressions. Plus AI responses are longer than human responses, and this response is long in length. All these make this response likely written by AI.

###
Response: Well, it made me feel really sad and alone. I didn't want to go to school anymore, and it felt like no one cared about how I felt. The things they said and did online, like posting hurtful stuff, it was just too much for me to handle. I started thinking about how maybe things would be better if I wasn't around anymore, but I didn't really want to die, you know? It's just that it felt like the pain wouldn't stop, and I didn't know what else to do.

Score:
Reasoning: