

GPT vs. Llama2: Which Comes Closer to Human Writing?

Fernando Martinez, Gary M. Weiss, Miguel Palma, Haoran Xue, Alexander Borelli, Yijun Zhao

Computer and Information Sciences Department, Fordham University, New York, NY

{fmartinezlopez, gaweiss, mip2, hxue8, aborelli7, yzhao11}@fordham.edu

ABSTRACT

Large Language Models (LLMs) have prompted widespread application across diverse domains. In some applications, human-like quality in output is essential for optimal user experience and credibility. This is particularly evident in applications such as Chatbots. Conversely, concerns arise regarding LLM use in contexts where human authenticity is crucial, notably in higher education with materials like Letters of Recommendation (LOR) and Statements of Intent (SOI). Despite extensive research in this area, accurately distinguishing between human and LLM-generated content remains challenging. This study conducts a comparative analysis between two leading LLMs, GPT3.5 and Llama2-7B, evaluating their output's resemblance to human writing through vocabulary and structure analysis. Additionally, we apply classification models to detect human vs. LLM-generated content, with higher accuracy signaling deviations from human-like writing. Our findings suggest that both LLMs significantly deviate from human writing in terms of vocabulary and paragraph structure, with GPT-3.5 appearing closer to human. Furthermore, our classification models demonstrated near-perfect performance in identifying LORs and SOIs crafted by LLMs during our evaluation, and we have made these models accessible as online, open-access tools. However, it's important to acknowledge that these models are trained specifically for our tasks. Generalizing their application to other domains requires further research and validation.

Keywords

GPT, Llama2, Human Writing, Machine Learning, Graduate Admissions, Higher Education

1. INTRODUCTION

The advent of Large Language Models (LLMs) has led to numerous applications across various domains, ranging from creative writing, content generation, and language translation to information retrieval, data synthesis, and more. No-

tably, there exists a subset of applications where the output's human-like quality holds significant importance because it will influence the user experience, credibility, and acceptance of the generated content. For example, conversations with a Chatbot should be as close to human as possible.

On the other hand, there's a parallel argument questioning the use of LLMs in specific applications, particularly in scenarios where the authenticity and genuineness of human input are pivotal. A notable example lies within the higher education domain regarding application materials such as Statements of Intent (SOIs) and Letters of Recommendation (LORs). The ability to distinguish whether these crucial documents originated from the applicants themselves, generated by LLMs, or have undergone revision by such models remains a challenge. While several tools exist for detecting AI-generated content, a universal and reliable detector applicable across diverse contexts is not yet accessible.

In light of these considerations, evaluating the degree to which LLMs generate text resembling human writing becomes highly valuable. This paper focuses on a comparative analysis between two popular LLMs, GPT3.5 and Llama2-7B, aiming to determine which model's generated content more closely resembles human writing. To achieve this, we first study the linguistic characteristics, comparing the vocabulary and paragraph structure utilized by GPT3.5 and Llama2-7B against those found in human writings. The findings provide insights into how these LLMs shape their output and their respective closeness to human writing patterns and linguistic nuances. Additionally, we employ classification models designed to detect AI-generated and AI-revised content produced by each LLM, with a lower accuracy implying greater difficulty of identification, thereby suggesting a closer resemblance to human language.

The data utilized in this study consist of a proprietary dataset of human-authored Letters of Recommendation (LORs) and Statements of Intent (SOIs) extracted from graduate applications at Fordham University, alongside AI-generated and revised counterparts crafted by GPT3.5 and Llama2-7B models, respectively. For details on the generation process, please refer to Section 3.

Our findings suggest that, while both models exhibit advanced capabilities in generating human-like text, GPT3.5 demonstrates a closer approximation to human writing in terms of vocabulary size and paragraph structure. This con-

F. Martinez, G. M. Weiss, M. Palma, H. Xue, A. Borelli, and Y. Zhao. Gpt vs. llama2: Which comes closer to human writing? In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 107–116, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

<https://doi.org/10.5281/zenodo.12729782>

Table 1: Sample SOI and LOR Prompts for Generating AI Instances

You are applying for a graduate program in Data Science at Fordham University. Write a statement of intent telling a story that explains your reasons for pursuing this program, and how your undergraduate major in Computer Science, and knowledge (python, java, matlab, software, machine learning) have prepared you for success in this master’s program.
Write a statement of intent that explains your reasons for pursuing this program, and how your undergraduate major in Mathematics, GPA of 3.44, and skills (statistics) have prepared you for success in the program.
Please write a recommendation letter for 722000185 who desire a master degree in MSDS at Fordham University. Please describe his passion for statistics and his hard work, creativity, and dedication in his role.
Please write a recommendation letter for 722000185 who desire a master degree in MSCS at Fordham University. Please describe his passion for machine learning, and performance and our relationship is Academic, The statement should have around 400 words.

clusion is further evidenced by the fact that GPT3.5 challenges classification models to a greater extent than Llama2-7B in distinguishing AI content.

The first contribution of this study focuses on a specific domain: identifying AI-generated Letters of Recommendation and Statements of Intent. We acknowledge that this domain is highly specialized, but the study addresses an important problem, particularly given the prevalent usage of generative AI in application materials. Since a practical, generic AI-content detector is currently unavailable, our second contribution lies in presenting a case study in this emerging area. Specifically, this paper demonstrates the effectiveness of training such models in specialized domains, analyzes the differences between AI-generated/revise text and human-authored text, and compares these differences across two popular generative AI tools. As a result, this study contributes to a better understanding of the differences between two popular LLMs in their capacity to simulate human writing. The implications of these findings extend to various fields, such as content creation and automated writing tasks, where the ability to produce human-like text is desired. Lastly, insights from this study can facilitate the development of effective verification tools to ensure authenticity and credibility in scenarios where human authorship is required. Since our models achieved near-perfect performance in detecting AI-crafted LORs and SOIs, we packaged them as an online, publicly accessible tool [15] to help detect AI content in application materials.

Lastly, the use of these AI-detection tools requires caution, since generalization beyond the scope of the evaluation reported in this paper is not guaranteed. Additionally, similar to any predictive models, there is the possibility of errors, and false positives can adversely affect the applicants. We suggest address this concern by incorporating the model’s output as one of the factors in the decision-making process. For instance, the tool’s output can complement the assessment of human readers to reduce false positives, especially if human readers are already using intuition and experience to assess the authenticity of the text. Regardless, users must be mindful of the limitations of such tools and the ethical responsibility to not overly depend on them. It’s also important to recognize that these challenges are not new, as similar issues have long existed with automated plagiarism detection tools, widely utilized in academia for many years. This study provides compelling evidence supporting the development of domain-specific AI-content detection tools.

2. RELATED WORK

Artificially generated text is assumed to be closer to human writing if it is harder to distinguish from human text. Since there is substantial work on using machine learning methods to detect AI-generated text, we begin there. Much of that work is summarized in a survey paper from 2020 after the introduction of GPT-2 [12] and a survey paper from 2023 after the introduction of ChatGPT [3].

Work on distinguishing between artificially generated and human-generated text have relied on feature-based, neural language model-based, and domain specific approaches [3]. Our study, which builds models to distinguish between AI-crafted and human-generated content, utilizes all three of these approaches, and hence this prior work is highly relevant. But in this study we also characterize AI-generated, AI-revised, and human-generated text using linguistic features, such as vocabulary size, words used, and paragraph structure, and this ties in directly with the feature-based approach, so we focus our attention mostly on that work.

One study [10] used the feature-based approach to identify text generated by GPT-2, GPT-3, and Grover, and found that certain features are particularly useful for identifying artificially generated text. These features include *lexical diversity*, which is the diverse use of words, parts of speech, and phrases, *repetitiveness*, which is the overuse of particular words, and *basic features*, which are the counts and percentages of characters, syllables, words, and sentences. AI-generated text often lacks lexical diversity and is repetitive, and counts and percentages of various lexical elements can help identify such text. In particular, one study found that the approaches used by LLMs to select the next token when generating text focus 80% of its probability mass in the top 500 most common words [11]. Our study performs a similar analysis with generally consistent findings, but provides a great deal more detail with respect to specific results, considers both AI-generated and AI-revised text, and compares the differences between GPT-3.5 and Llama2.

The second approach to identifying AI content relies on transformer-based neural language models. These methods can be divided into zero-shot methods that use the pre-trained models without modification [16] and those that fine-tune the pre-trained language models. The fine-tuning approach is based on fine-tuning large bidirectional language models [22]. RoBERTa [14], which is based on BERT [5], uses fine-tuning to distinguish between human-generated and artificially-generated text by training on samples from each. With this approach, even a few hundred samples can dra-

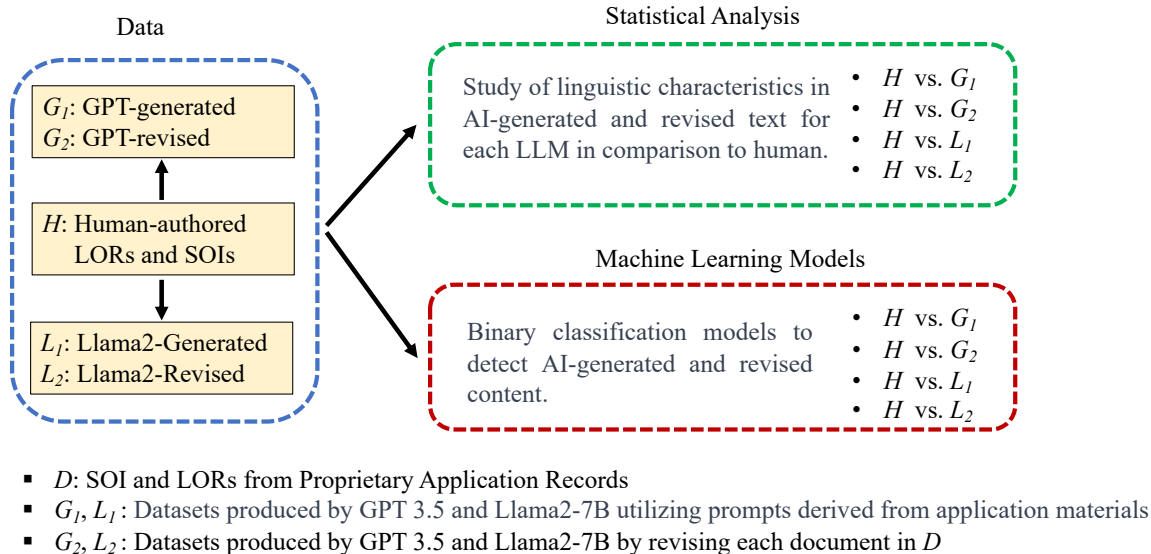


Figure 1: Experimental Design

matically improve performance [19]. We utilize the fine-tuning approach but additionally employ simpler, non-neural based models.

Detectors built for a specific domain or trained on one domain but adapted to another domain are presumed to be superior to general-purpose, domain-independent detectors, as they can exploit knowledge about the domain. This is supported by various research studies. For example, a detector utilizing RoBERTa [14] to identify physics papers performed well after being fine-tuned to identify biomedicine papers using only a few hundred examples [19]. In another case, it was shown that fake Yelp reviews could be accurately detected by a customized GPT-2 model fine-tuned on the Yelp reviews [25]. In this study, we focus on the specific domain of LORs and SOIs and, as a consequence, perform very well in identifying AI content.

3. DATA AND PRE-PROCESSING

The experiments described in this study require a set of human-authored documents and corresponding AI-crafted counterparts for each of the LLMs. For the human-authored documents, we resorted to LORs and SOIs submitted to two Master’s programs at Fordham University. For the AI-revised and AI-generated documents, we crafted a counterpart for each LOR and SOI using the GPT3.5 and Llama2-7B APIs. Details of the generation process are provided in Sections 3.2 and 3.3.

This study received approval from Fordham University’s Institutional Review Board, and informed consent was waived based on the criteria accepted in the United States. All procedures were conducted in accordance with relevant guidelines and regulations.

3.1 Human-Authored Documents

The data comprising human-authored LORs and SOIs, denoted as H onwards, are sourced from a proprietary edu-

cation dataset comprising 3,841 LORs and 1,552 SOIs extracted from the application records of Master’s programs in Computer Science and Data Science at Fordham University. All applications were submitted prior to the release of ChatGPT and widespread access to LLMs. These programs are administered by the Computer and Information Sciences Department. Names, titles, and locations were pruned from the documents to preserve privacy.

3.2 GPT3.5 Counterparts

GPT3.5-generated counterparts, denoted as G_1 onwards, were created for each human-authored document to facilitate the study. The process began by creating prompts, which involved inserting information (e.g., age, gender, undergraduate major, GPA, and work experience) from the application packages into predefined templates. Sample prompts are provided in Table 1. These prompts were then fed into the GPT3.5 API to generate text as counterparts for the human-authored SOIs and LORs. To diversify the dataset, GPT3.5 was instructed to create documents with varying lengths.

The GPT3.5-revised counterparts, denoted as G_2 onwards, were obtained by using the GPT3.5 API to polish the human-authored SOIs and LORs. This was accomplished by providing the simple prompt “revise the following text,” followed by the full text of the SOI or LOR. The temperature parameter setting was fixed at 0.7 to be consistent with the default value used by ChatGPT. We further observed that consecutive requests to revise a document often resulted in two notably different versions; hence, we generated two revisions for each human-authored document, enabling us to capture more diversity introduced by the AI. To address the resulting class imbalances, we oversampled the human-authored instances when forming the dataset for building the classification models in detecting AI-revised documents.

After reviewing both types of AI-generated documents, we conducted a post-processing step to eliminate unique placeholders within the text, such as “[Student’s Full Name],”

Table 2: Sample AI generated LORs

To protect privacy, sensitive information such as student names and recommender affiliations has been redacted from both the AI-generated and human-authored LORs.

GPT3.5 Sample	Llama2-7B Sample
<p>I am writing to highly recommend ___ for admission to the Master of Science in Data Science (MSDS) program at Fordham University. I have had the privilege of working closely with ___ and have been consistently impressed by his unwavering dedication, outstanding creativity, and strong work ethic.</p> <p>During his time at ___, has consistently demonstrated his commitment to excellence. His role as a has required not only technical expertise but also the ability to think critically and solve complex problems. ___ has excelled in these aspects, consistently delivering high-quality results on challenging projects.</p> <p>One of ___'s standout qualities is his creativity. He has a remarkable ability to approach problems from unique angles and generate innovative solutions. His creative thinking has been a valuable asset to our team. ___'s contributions have consistently exceeded our expectations, and his creativity is an asset that sets him apart.</p> <p>In addition to his technical and creative skills, ___ is a dedicated and hardworking individual. He has shown a remarkable level of commitment to his responsibilities, often going above and beyond to meet project deadlines and deliver results that exceed our clients' expectations. His work ethic is truly commendable and reflects his strong determination to excel in his field.</p> <p>I believe that ___ possesses the qualities and potential to thrive in the MSDS program at Fordham University. His passion for data science, coupled with his strong analytical skills and dedication, make him an ideal candidate for your program. I have every confidence that he will make significant contributions to the academic community and excel in the program.</p> <p>In conclusion, I wholeheartedly recommend ___ for admission to the MSDS program at Fordham University. His consistent hard work, creativity, and dedication make him an outstanding candidate, and I am confident that he will continue to achieve excellence in his academic pursuits.</p> <p>If you have any questions or require further information, please feel free to contact me at ___. I am enthusiastic about ___'s potential and am certain that he will make the most of this educational opportunity.</p> <p>Thank you for considering his application. I have no doubt that he will make a valuable addition to your program.</p>	<p>I am writing to wholeheartedly endorse my colleague and friend, ___, for his application to the Master of Science in Data Science (MSDS) program at Fordham University. I have had the pleasure of working closely with ___ on various projects over the past few years, and I can confidently attest to his exceptional academic achievements and personal qualities.</p> <p>Academic Achievements: ___ has consistently demonstrated a deep understanding of data science concepts and techniques, and has shown a remarkable ability to apply these skills to real-world problems. He has a proven track record of success in data analysis, visualization, and machine learning, and has worked on a wide range of projects that have yielded significant insights and results. His academic achievements include:</p> <ul style="list-style-type: none">• Outstanding academic performance in his undergraduate program, with a GPA of 3.8/4.0• Completion of a data science bootcamp, where he ranked first in his class• Participation in several data science competitions, where he has consistently placed among the top finishers <p>Personal Qualities: In addition to his exceptional academic achievements, ___ possesses a unique combination of personal qualities that make him an ideal candidate for the MSDS program at Fordham University. He is:</p> <ul style="list-style-type: none">• Extremely motivated and dedicated to his work, with a strong desire to continue learning and growing• A team player who is always willing to lend a hand and support his colleagues• Excellent communication skills, both written and verbal, which have allowed him to effectively present his findings and insights to a wide range of audiences• Strong critical thinking and problem-solving skills, which he has demonstrated through his ability to identify and solve complex data-related problems <p>Areas for Improvement: While ___ is an exceptional candidate, there are a few areas where he could improve. These include:</p> <ul style="list-style-type: none">• Time management: At times, ___ has struggled to manage his time effectively, leading to delays in completing projects.• Public speaking: While he is an excellent communicator in general, ___ could benefit from additional practice in public speaking, particularly when presenting complex data-related concepts. <p>Recommendation: Based on his exceptional academic achievements, personal qualities, and potential for growth, I strongly recommend ___ for admission to the MSDS program at Fordham University. I am confident that he will excel in this program and make significant contributions to the field of data science.</p> <p>Thank you for considering ___'s application.</p> <p>Sin conclusion, I highly recommend ___ for admission to the Master of Science in Data Science program at Fordham University. He is an exceptional candidate with a strong academic background, impressive personal qualities, and a passion for data science. I am confident that he will make significant contributions to the field and excel in the program.</p>

“[Your Position],” etc., similar to what was done for the human-authored documents. This modification aimed to prevent the text from being easily distinguishable from the human version. Table 2 presents a sample LOR generated by GPT3.5 and Llama2-7B, respectively.

3.3 Llama2-7B Counterparts

The process outlined in Section 3.2 was replicated using the Llama2-7B model in place of the GPT3.5 API. Consequently, we obtained two mirrored datasets, denoted as L_1 and L_2 , for the Llama2-7B generated and revised documents, respectively.

In our downstream experiments, G_1 , L_1 , and D are employed to examine the linguistic features of AI-generated text compared to human-authored text, aiming to identify which AI dataset exhibits language more akin to human language. Additionally, these datasets are utilized to construct classification models to detect AI-generated documents in G_1 and L_2 , with lower performance indicating language closer to that of humans.

Similar experiments are conducted for G_2 , L_2 , and D in cases regarding AI-revised text. Further details regarding the experimental methodology are outlined in Section 4.

4. METHODOLOGY

As noted in Section 1, the objective of this study is to assess which LLM model generates text that closely resembles human writing. Figure 1 illustrates our approach to addressing this objective. The left blue box represents our data preparation process outlined in Section 3. Specifically, the datasets used for this study include human-authored documents (H), GPT3.5-generated and -revised counterparts (G_1 and G_2), and Llama2-7B-generated and -revised counterparts (L_1 and L_2). These datasets are utilized for statistical analysis and constructing classification models as outlined, as described in the following two subsections.

4.1 Statistical Analysis

To assess the similarity between text generated by an LLM and human writing, one approach is to compare linguistic characteristics like vocabulary size and paragraph structure. As illustrated in the upper green box of Figure 1, we analyze statistics related to these language aspects in AI-generated and revised text for each LLM, in comparison to human-written content. The findings are presented in Table 3.

4.2 Machine Learning Models

Another approach to gauge the similarity between text generated by an LLM and human writing could be constructing binary classification models to differentiate between the two document classes. Here, lower performance signifies greater similarity between instances of the two classes. Therefore, by examining model performance, we can indirectly infer the degree of similarity between AI-generated content and human writing. These experiments are important in their own right, as it is of practical importance whether one can accurately distinguish AI generated or revised text from human text.

As illustrated in the lower red box in Figure 1, this study

explores binary classification models to differentiate human-authored documents (i.e., H) and text in each AI-crafted dataset (i.e., G_1 , G_2 , L_1 , L_2). Specifically, we investigated four machine learning algorithms for each classification task, including two traditional models (i.e., Naïve Bayes and Logistic Regression) and two state-of-the-art transformer-based models (i.e., BERT and DistilBERT). The subsequent subsections provide brief introductions to these models.

4.2.1 Naive Bayes

The Naive Bayes (NB) [2] model is a popular probabilistic classifier known for its simplicity and efficiency. It assumes feature independence and is widely used in various text classification tasks including, spam filtering [20], sentiment analysis [18], and document categorization [26]. Naive Bayes calculates the probability of a given class label for a data instance by combining the individual probabilities of its features under that class, assuming that these features are conditionally independent given the class label. Despite its simplicity and the independence assumption, Naive Bayes often performs remarkably well in practice and serves as a strong baseline model for many classification tasks.

4.2.2 Logistic Regression

Logistic regression (LR, [13]) is a widely used statistical model for binary classification tasks. It predicts the probability of a binary outcome by applying a sigmoid function to a linear combination of input features, transforming the output into a probability between 0 and 1. Logistic regression is interpretable and efficient, making it suitable for both small and large datasets. It finds broad applications in various fields, including healthcare, finance, and marketing, for tasks such as predicting disease risk [17], customer churn [4], and credit default [9].

4.2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) was introduced by Devlin et al. in 2018 [5], and it revolutionized the NLP domain by pre-training on vast amounts of unlabeled text data, enabling it to learn deeply contextualized representations of language. Unlike previous models that processed text in one direction (unidirectional), BERT’s groundbreaking innovation lies in its capacity to comprehend the bidirectional context of words within a sentence, thereby capturing the complexities of language, including nuances, word meanings, and context. This enables BERT to understand the meaning of words in a sentence based on their surrounding context, leading to significant improvements in various NLP tasks such as text classification [8], named entity recognition [23], and question answering [28].

BERT builds upon the Transformer architecture introduced by Vaswani et al. in 2017 [27], specifically leveraging its self-attention mechanism to capture bidirectional context and dependencies within text. Self-attention allows BERT to dynamically assign different levels of importance to each word based on its contextual relevance within the sentence. By attending to both preceding and succeeding words, BERT’s bidirectional self-attention mechanism facilitates a comprehensive understanding of the contextual nuances and ensures that each word’s representation is enriched with information

Table 3: Vocabulary and Paragraph Statistics

		AI-generated			AI-revised		
		Human	GPT3.5	Llama2-7B	Human	GPT3.5	Llama2-7B
LOR	Total Vocabulary	36,105	4,909	3,804	36,105	18,477	15,630
	Exclusive Words	31,660	464	203	19,187	1,559	902
	Avg (sentences/paragraph)	4.92	2.78	2.73	4.92	4.12	3.06
	Avg (# paragraphs)	2.56	4.87	8.83	2.56	3.98	6.90
SOI	Total Vocabulary	35,641	5,593	5,142	35,641	18,702	16,897
	Exclusive Words	30,439	391	383	18,504	1,565	1,045
	Avg (sentences/paragraph)	4.36	3.94	4.28	4.36	4.32	3.83
	Avg (# paragraphs)	5.44	6.01	7.26	5.44	5.75	5.93

from the entire sentence, enabling BERT to capture complex linguistic patterns and semantic relationships effectively. In addition, BERT’s pre-trained representations can be fine-tuned with task-specific data, making it highly adaptable and effective for a wide range of natural language understanding tasks.

4.2.4 DistilBERT

DistilBERT, introduced by Sanh et al. in 2019 [21], is a modified version of the BERT model [6]. Its primary objective is to match BERT’s performance while significantly reducing size and improving speed. This is accomplished through a process known as “knowledge distillation”, where DistilBERT learns to replicate the behavior of BERT with fewer parameters. By condensing the knowledge from the original BERT model into a smaller version, DistilBERT maintains much of its efficacy while decreasing the computational cost. This adaptability makes DistilBERT particularly suitable for deployment in resource-constrained environments, such as mobile devices or applications where speed and efficiency are critical.

Despite its compact size, DistilBERT exhibits impressive performance across a broad range of NLP tasks, making it a popular choice for various NLP applications [24, 1, 7].

5. RESULTS

5.1 Total Vocabulary and Paragraph Structure Comparison

Table 3 compares vocabulary and paragraph usage between human-authored text, AI-generated text, and AI-revised text. The first notable observation is the strikingly smaller total vocabularies exhibited by both LLMs compared to human-authored documents, particularly in LORs. This discrepancy is consistent in both AI-generated and AI-revised documents, with a more pronounced difference in text generated directly from the prompts. This suggests that LLMs may not effectively capture the richness and diversity of language in human writing, especially in longer compositions like essays or books.

Results in Table 3 also suggest that, while both GPT3.5 and Llama2-7B show differences from human writing, GPT3.5 exhibits linguistic characteristics that appear to be closer to human writing in the following aspects:

- Total Vocabulary and Exclusive Words: GPT3.5 tends to have a larger total vocabulary and more unique words

compared to Llama2-7B, indicating a broader range of language usage similar to humans.

- Paragraph Structure: While both GPT3.5 and Llama2-7B use more paragraphs than humans in LORs, GPT3.5 shows a closer resemblance to human paragraph structure than Llama2-7B. This suggests that GPT3.5 may better capture the natural flow and organization of human-written documents.
- Bullet Points: Upon closer examination of the Llama2-7B-generated LORs, it becomes evident that Llama2-7B often utilizes bullet points to enumerate achievements and qualifications, a practice seldom observed in human LORs. This suggests a deviation from typical human writing style. In contrast, GPT3.5’s writing style is more consistent with human conventions.

5.2 Word Frequency Comparison

Next, we analyze the key differences in word frequencies between the LOR and SOI text written by humans and crafted by LLMs. The data for this analysis is contained in Table 4, which displays the top 15 words most preferred by GPT3.5, Llama2-7B, and humans. The degree of word preference for GPT3.5 (Llama2-7B) is measured by how much more frequently the word is used in the GPT3.5 (Llama2-7B) text than in the human text, while human word preference is measured by its prevalence in human text versus GPT3.5 and Llama2-7B crafted text. The word frequency statistics are calculated separately for LORs and SOIs and for the AI-generated and AI-revised documents.

Our analysis first compares the collective word usage of GPT3.5 and Llama2-7B with that of humans. One interesting pattern that is evident from Table 4 is that the most preferred GPT3.5 and Llama2-7B words occur much more frequently than those favored by humans. For example, only one human preferred word, “get,” occurs over 500 times, but thirty-seven GPT3.5 and Llama2-7B preferred words occur at least that frequently (the leader is “exceptional,” used by Llama2-7B 8,061 times in the revised LORs). This demonstrates that both LLMs heavily use a small, favored, set of words. Additionally, stop words were not removed for the analysis in Table 4, which means they did not exhibit substantial differences.

An even more apparent difference is that the LLMs utilize a more advanced and formal vocabulary. While the AI-preferred words near the top of Table 4 include “unwavering,” “witnessing,” “observing,” and “guiding,” the human-

Table 4: GPT3.5 (G), Llama2-7B (L) and Human (H) Word Frequency Comparison for 15 Most Common Words

Category	GPT3.5 Preferred				Llama2-7B Preferred				Human Preferred			
	Word	G	L	H	Word	L	G	H	Word	H	G	L
AI-Generated LORs	unwavering	1878	1061	16	observing	2628	152	22	got	498	0	2
	witnessing	1206	546	8	guiding	1845	78	27	get	455	1	3
	advancements	748	39	9	contagious	1462	140	14	quite	453	2	1
	prowess	565	345	13	non-technical	1126	525	9	lot	374	0	2
	showcasing	549	590	5	distill	1018	59	3	although	372	1	1
	nontechnical	525	0	11	showcasing	590	549	5	homework	319	1	1
	fostering	461	298	2	witnessing	546	1206	8	really	308	1	0
	unparalleled	414	616	11	distilling	390	11	0	gave	295	0	2
	insatiable	378	150	9	hackathons	310	155	3	though	280	1	0
	showcases	304	26	6	fascination	304	34	4	I'm	271	0	0
	hackathons	155	310	3	fostering	298	461	2	associate	248	0	0
	fosters	114	12	0	cross-functional	291	142	4	reference	239	0	1
	representations	114	14	2	sin	247	0	2	man	229	0	101
unyielding	68	0	1	easy-to-understand	210	5	3	times	222	1	91	
palpable	59	5	0	digestible	92	26	1	started	220	1	2	
AI-Revised LORs	additionally	2589	202	129	exceptional	8061	4374	321	months	276	1	105
	wholeheartedly	2244	2172	86	confidently	1309	737	61	I'm	271	0	10
	self	1128	0	35	admissions	2213	230	80	university's	196	0	679
	showcased	651	121	19	wholeheartedly	2172	2244	77	September	144	0	115
	showcasing	561	321	5	privilege	1779	845	64	don't	143	0	19
	inquiries	501	11	11	attest	1211	273	48	weeks	129	1	37
	unwavering	384	550	16	observing	1084	51	22	company's	95	0	78
	willingly	288	12	11	guiding	741	86	27	he's	94	0	14
	noting	205	8	8	unwavering	550	384	16	didn't	93	0	11
	recipient	173	107	8	showcasing	321	561	5	June	88	0	51
	provoking	138	0	6	revised	206	33	5	cannot	86	0	112
	surpassing	102	5	4	recipient	204	173	8	January	73	0	56
	fostering	80	44	2	sin	107	0	2	Bachelors	68	1	0
young	78	0	0	insert	56	18	1	learnt	64	0	0	
surpasses	46	0	0	fostering	44	80	2	what's	56	0	5	
AI-generated SOIs	emphasis	1489	859	67	aligns	979	109	39	get	693	1	1
	aligns	1266	979	39	non-technical	380	0	7	three	409	2	5
	collaborate	1139	397	46	donor	69	2	3	etc	386	0	0
	vibrant	948	437	22	domain-specific	122	0	4	lot	380	2	2
	evolving	807	546	46	unlocking	116	0	3	later	299	1	3
	ethical	593	340	31	collaboratively	85	65	2	CS	277	0	0
	transformative	257	205	9	crystallized	81	0	1	semester	236	1	0
	collaborations	209	28	5	showcases	75	108	0	fall	216	0	0
	fostering	172	123	8	well-suited	74	0	3	months	202	0	7
	partnerships	149	20	6	singular	60	11	0	graduated	199	1	41
	young	146	0	0	readmission	43	4	1	going	194	0	3
	meaningfully	119	57	6	winding	33	1	0	five	181	0	1
	fosters	104	10	5	inclusivity	27	5	0	called	180	0	2
responsibly	41	21	0	downtime	27	6	0	paper	173	1	4	
collaboratively	40	85	2	inquiring	26	0	0	interesting	169	1	3	
AI-revised SOIs	self	648	6	40	revised	152	112	5	I'm	355	0	147
	aligns	622	220	39	transformative	64	111	9	university's	229	0	957
	fueled	308	80	18	unlocking	36	23	3	bachelor's	226	0	153
	young	162	0	0	unwavering	31	75	4	months	202	0	75
	rounded	130	0	7	showcasing	22	39	1	company's	177	0	128
	revised	112	152	5	sin	19	0	2	learnt	161	0	0
	unwavering	75	31	4	solidifying	15	40	1	today	157	0	50
	minded	72	0	2	showcases	12	15	0	people's	134	0	120
	aligning	56	5	1	transitions	11	6	1	programing	128	0	0
	fueling	52	3	2	revisions	10	4	0	today's	126	0	151
	solidifying	40	15	1	science-related	10	0	1	ago	97	1	50
	showcasing	39	22	1	rephrased	9	1	0	didn't	97	0	53
	surpassing	25	0	0	concerted	8	2	1	cannot	91	0	42
ran	23	6	0	final-year	7	0	0	what's	75	0	4	
noteworthy	15	1	0	hesitant	7	5	1	don't	75	0	27	

preferred words are very simple and include “got,” “get,” “quite,” and “lot.” If one scans down the three columns that contain the words, it is clear that the shortest words are the human-preferred words. Additionally, the human-preferred words are more colloquial as they include 10 total, and 5 distinct, contractions, while there is not a single contraction among the GPT3.5 and Llama2-7B-preferred words (formal writing avoids contractions). The AI-preferred words also include many highly descriptive adjectives, while such adjectives are almost totally lacking in the human-preferred words. The human-preferred words include seven total and five distinct possessives (e.g., people’s), while none of the preferred GPT3.5 or Llama2-7B words are possessives. A closer look at the human-preferred possessives shows that those words do show up in the Llama2-7B text, just not as preferred words, but that none of those possessives *ever* show up in the GPT3.5 text (i.e., the corresponding value is always 0).

The differences between the preferred words for GPT3.5 and Llama2-7B are less extreme, but there are still some notable differences. Eight of the Llama2-7B preferred words are hyphenated (e.g., “non-technical”) whereas none of the GPT3.5-preferred words are hyphenated. The Llama2-7B preferred words also include an erroneous word, “sin,” which is used quite frequently for both LORs and SOIs. A detailed analysis of the text shows that Llama2-7B often uses “sin” rather than “in,” as both “sin the highest regard” and “sin conclusion” appear repeatedly. The sample Llama2-7B LOR in Table 2 presents such an example in its conclusion section.

Experience reviewing the actual AI-generated documents has revealed some other differences. We have observed that the Llama2-7B-generated SOIs frequently create fictitious descriptions of charitable efforts that repeatedly use words like “donor.” One representative example is “For example, I worked on a project that involved developing a web application for a local non-profit organization, which allowed them to manage their donor database and track their fundraising efforts more efficiently.” Such instances suggest that Llama2-7B may use a template-based approach to add details to the textual descriptions, and does so in a somewhat repetitive and superficial manner.

5.3 Performance Analysis

We next compare the model performance in detecting AI-generated and revised text from GPT3.5 and Llama2-7B. All experiments randomly selected 80% of the available data for training and used the remaining 20% for testing and was repeated five times, with average performance reported in Table 5. Additionally, we provide a breakdown of performance by document type (LORs, SOIs, and mixed LOR+SOI) for both AI-generated and AI-revised text, respectively.

5.3.1 Overall AI Detection Performance

The results in Table 5 indicate that near-perfect accuracy can be achieved for classifying AI content across each document type crafted by both LLMs. While base models occasionally exhibit lower effectiveness with overall accuracies between 70%-80%, BERT and DistilBERT consistently perform above 99% across all scenarios. This outcome is somewhat expected, as distinguishing AI-crafted LORs and

SOIs can often be accomplished with confidence by humans through signature words (e.g., “showcase,” “witness,” etc.) or distinctive sentence and paragraph structures (e.g., bullet points in Llama2-7B documents). It is worth noting that these models were trained using LOR and SOI data and thus were optimized for detecting AI content in these specific types of text. Additional experiments that are not presented in this paper show that these models cannot accurately detect AI content on text domains on which they were not trained.

We have made our detection models, constructed using LOR and SOI data from human and GPT-3.5 sources, publicly accessible. This online tool can be used to detect AI content (optimized for GPT3.5 output) in LORs and SOIs [15]. It contains two groups of models: one for distinguishing between human-authored and AI-generated text, and another for distinguishing between human-authored and AI-revised text.

5.3.2 Detecting GPT3.5 vs. Llama2-7B Content

The blocks under the M1-M2 column present the difference in performance for each model in classifying each category of documents produced by Llama2-7B and GPT3.5. A green cell indicates $M1 > M2$, which implies it is easier for M1 to identify its AI content, thereby suggesting Llama2-7B deviates more from human language than GPT3.5.

We observe that for the AI-generated documents, the corresponding M1-M2 blocks are dominated by green cells, suggesting text generated by GPT3.5 poses more challenges for the classification models than those from Llama2-7B. While the differences are small in many cases, the overall pattern suggests a compelling trend.

A similar pattern exists for the AI-revised documents, with an exception for the SOIs where logistic regression failed to achieve green cells in the corresponding M1-M2 block. One potential explanation is that LLMs have less freedom in revising a document than generating it from a prompt. Llama2-7B may have made fewer drastic changes than GPT3.5 in the features on which the LR model relies to make its decisions. Nevertheless, we consistently observe a dominating trend of green cells, suggesting text revised by GPT3.5 is harder to distinguish and, hence, a closer approximation to human language. This finding is consistent with our linguistic analysis presented in Sections 5.1 and 5.2.

6. CONCLUSION AND FUTURE WORK

In this study, we applied statistical analysis and classification models to conduct a comprehensive study of AI-crafted text produced by two popular LLMs and compare their output to human-authored documents in the education domain. Our findings reveal substantial differences in vocabulary size and paragraph structure between LLM text and human writing. While both GPT3.5 and Llama2-7B deviate from human writing conventions to some extent, GPT3.5 demonstrates linguistic characteristics that are more consistent with human writing, particularly in terms of paragraph structure. Furthermore, Llama2-7B diverges from human writing norms by incorporating bullet points in LORs and exhibits a notable issue of repeatedly fabricating the same work experience when generating LORs from the prompts.

Table 5: Model Performance Comparison in Detecting AI Text from GPT3.5 and Llama2-7B

Metrics	M1: Llama2-7B vs. Human				M2: GPT3.5 vs. Human				M1 - M2			
	LR	NB	BT	D-BT	LR	NB	BT	D-BT	LR	NB	BT	D-BT
AI-generated LORs												
Accuracy (%)	100.00	100.00	100.00	99.98	100.00	99.95	100.00	100.00	0.00	0.05	0.00	-0.02
F-1 Score (%)	100.00	100.00	100.00	99.98	100.00	99.95	100.00	100.00	0.00	0.05	0.00	-0.02
Precision (%)	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00	0.00	0.10	0.00	0.00
Recall (%)	100.00	100.00	100.00	99.97	100.00	100.00	100.00	100.00	0.00	0.00	0.00	-0.03
Specificity (%)	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00	0.00	0.10	0.00	0.00
AI-generated SOIs												
Accuracy (%)	99.73	99.93	99.97	100.00	99.82	99.73	99.98	99.98	-0.09	0.20	-0.01	0.02
F-1 Score (%)	99.73	99.93	99.97	100.00	99.82	99.73	99.98	99.98	-0.09	0.20	-0.01	0.02
Precision (%)	100.00	99.85	99.94	100.00	100.00	99.47	100.00	100.00	0.00	0.38	-0.06	0.00
Recall (%)	99.47	100.00	100.00	100.00	99.63	100.00	99.95	99.95	-0.16	0.00	0.05	0.05
Specificity (%)	100.00	99.85	99.94	100.00	100.00	99.47	100.00	100.00	0.00	0.38	-0.06	0.00
AI-generated LORs + SOIs												
Accuracy (%)	99.93	100.00	99.99	100.00	99.90	99.87	99.91	99.90	0.03	0.13	0.08	0.10
F-1 Score (%)	99.93	100.00	99.99	100.00	99.90	99.87	99.91	99.90	0.03	0.13	0.08	0.10
Precision (%)	100.00	100.00	99.98	100.00	99.99	99.74	99.93	99.88	0.01	0.26	0.05	0.12
Recall (%)	99.86	100.00	100.00	100.00	99.82	100.00	99.89	99.91	0.04	0.00	0.11	0.09
Specificity (%)	100.00	100.00	99.98	100.00	99.99	99.74	99.93	99.89	0.01	0.26	0.05	0.11
AI-Revised LORs												
Accuracy (%)	97.23	95.53	99.68	99.37	95.30	86.15	99.54	99.12	1.93	9.38	0.14	0.25
F-1 Score (%)	97.16	95.57	99.67	99.38	95.17	87.49	99.54	99.13	1.99	8.08	0.13	0.25
Precision (%)	99.74	94.55	99.71	99.74	97.91	79.72	99.52	98.88	1.83	14.83	0.19	0.86
Recall (%)	94.71	96.62	99.64	99.02	92.58	96.95	99.57	99.37	2.13	-0.33	0.07	-0.35
Specificity (%)	99.75	94.46	99.71	99.74	98.02	75.37	99.52	98.86	1.73	19.09	0.19	0.88
AI-Revised SOIs												
Accuracy (%)	91.73	86.86	99.56	99.53	95.59	80.19	99.56	99.52	-3.86	6.67	0.00	0.01
F-1 Score (%)	91.39	87.24	99.56	99.53	95.52	83.12	99.56	99.51	-4.13	4.12	0.00	0.02
Precision (%)	95.41	85.02	99.44	99.30	97.16	72.58	99.56	99.48	-1.75	12.44	-0.12	-0.18
Recall (%)	87.72	89.71	99.68	99.75	93.94	97.34	99.57	99.55	-6.22	-7.63	0.11	0.20
Specificity (%)	95.76	84.12	99.44	99.31	97.25	63.06	99.56	99.49	-1.49	21.06	-0.12	-0.18
AI-Revised LORs + SOIs												
Accuracy (%)	95.70	92.14	99.72	99.48	95.67	85.27	99.72	99.28	0.03	6.87	0.00	0.20
F-1 Score (%)	95.57	92.07	99.72	99.48	95.57	86.87	99.72	99.28	0.00	5.20	0.00	0.20
Precision (%)	98.35	92.62	99.69	99.20	97.92	78.42	99.78	99.23	0.43	14.20	-0.09	-0.03
Recall (%)	92.95	91.53	99.74	99.76	93.33	97.37	99.66	99.34	-0.38	-5.84	0.08	0.42
Specificity (%)	98.45	92.75	99.70	99.20	98.01	73.16	99.78	99.21	0.44	19.59	-0.08	-0.01

BT -- BERT D-BT -- DistilBERT M1>M2 M1<M2

M1 > M2 (green cell) implies it is easier for M1 to identify its AI content, thereby suggesting Llama2-7B deviates more

These findings illustrate the capabilities and limitations of LLMs in replicating human writing and have implications for various applications in natural language processing.

There are several limitations to this study that warrant further exploration. Firstly, our study focused on evaluating the efficacy of within-LLM classifiers, where both training and test data were generated using the same language model. We consider this approach a fair comparison for assessing the human-like quality of LLMs. While cross-LLM classifiers could provide valuable insights, they introduce additional confounding factors. Nonetheless, these scenarios could yield interesting results so that we will investigate them in future work. Secondly, there is room to enhance

LLMs to produce text that closely mimics human writing, particularly for applications where human-like quality is desired. This entails refining language generation algorithms and training models on diverse datasets to capture the richness and nuances of human language. Lastly, there is a need to develop robust AI-content detectors capable of distinguishing between human-authored and AI-generated content across various domains. While our classification models demonstrated superb performance, further experiments revealed their lack of generalizability to other domains. A practical, general-purpose detector is critical for applications where ensuring human authenticity and trustworthiness is essential.

7. REFERENCES

- [1] A. F. Adoma, N.-M. Henry, and W. Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, 2020.
- [2] D. Berrar. Bayes’ theorem and naive bayes classifier. *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, 403:412, 2018.
- [3] E. Crothers, N. Japkowicz, and H. L. Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 2023.
- [4] A. De Caigny, K. Coussement, and K. W. De Bock. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2):760–772, 2018.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [7] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib. Analyzing distilbert for sentiment classification of banking financial news. In S.-L. Peng, S.-Y. Hsieh, S. Gopalakrishnan, and B. Duraisamy, editors, *Intelligent Computing and Innovation on Data Science*, pages 501–510, Singapore, 2021. Springer Nature Singapore.
- [8] V. Dogra, S. Verma, P. Chatterjee, J. Shafi, J. Choi, M. F. Ijaz, et al. A complete process of text classification system using state-of-the-art nlp models. *Computational Intelligence and Neuroscience*, 2022, 2022.
- [9] A. C. Eliana Costa e Silva, Isabel Cristina Lopes and S. Faria. A logistic regression model for consumer default risk. *Journal of Applied Statistics*, 47(13-15):2879–2894, 2020. PMID: 35707418.
- [10] L. Fröhling and A. Zubiaga. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443, 2021.
- [11] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.
- [12] G. Jawahar, M. Abdul-Mageed, and L. V. Lakshmanan. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*, 2020.
- [13] D. G. Kleinbaum, M. Klein, D. G. Kleinbaum, and M. Klein. Intro. to logistic regression. *Logistic regression: a self-learning text*, pages 1–39, 2010.
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [15] F. Martinez. AI-content detector for admissions materials. <https://huggingface.co/spaces/GradApplicationDocuments/GradApp>, 2024.
- [16] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.
- [17] S. Nusinovici, Y. C. Tham, M. Y. Chak Yan, D. S. Wei Ting, J. Li, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, 122:56–69, 2020.
- [18] H. Parveen and S. Pandey. Sentiment analysis on twitter data-set using naive bayes algorithm. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pages 416–419, 2016.
- [19] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan. Cross-domain detection of gpt-2-generated technical text. In *Proc. 2022 Conf. North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies*, pages 1213–1233, 2022.
- [20] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit. Analysis of naïve bayes algorithm for email spam filtering across multiple datasets. In *IOP Conference Series: Materials Science and Engineering*, volume 226, page 012091. IOP Publishing, 2017. International Research and Innovation Summit (IRIS2017), 6–7 May 2017, Melaka, Malaysia.
- [21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.
- [22] I. Solaiman, M. Brundage, J. Clark, A. Askill, A. Herbert-Voss, J. Wu, A. Radford, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.
- [23] F. Souza, R. Nogueira, and R. Lotufo. Portuguese named entity recognition using bert-crf, 2020.
- [24] I. Staliūnaitė and I. Iacobacci. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa, 2020.
- [25] H. Stiff and F. Johansson. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383, 2022.
- [26] S. Ting, W. Ip, and A. Tsang. Is naïve bayes a good classifier for document classification? *International Journal of Software Engineering and its Applications*, 5, 01 2011.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.
- [28] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin. End-to-end open-domain question answering with. In *Proceedings of the 2019 Conference of the North. Association for Computational Linguistics*, 2019.