Received: 1 April 2024 | Revised: 5 June 2024 | Accepted: 31 July 2024

DOI: 10.1111/exsy.13699

ORIGINAL ARTICLE

Expert Systems WILEY

Comparative analysis of paraphrasing performance of ChatGPT, GPT-3, and T5 language models using a new ChatGPT generated dataset: ParaGPT

Meltem Kurt Pehlivanoğlu[1] | Robera Tadesse Gobosho[1] | Muhammad Abdan Syakura[1] | Vimal Shanmuganathan[2] | Luis de-la-Fuente-Valentín[3]

[1]Department of Computer Engineering, Kocaeli University, Kocaeli, Turkey

[2]Department of Artificial Intelligence and Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India

[3]School of Engineering and Technology, Universidad Internacional de La Rioja, Logroño, Spain

Correspondence

Meltem Kurt Pehlivanoğlu, Department of Computer Engineering, Kocaeli University, Kocaeli 41001, Turkey.
Email: meltem.kurt@kocaeli.edu.tr

Funding information

Kocaeli Üniversitesi

Abstract

Paraphrase generation is a fundamental natural language processing (NLP) task that refers to the process of generating a well-formed and coherent output sentence that exhibits both syntactic and/or lexical diversity from the input sentence, while simultaneously ensuring that the semantic similarity between the two sentences is preserved. However, the availability of high-quality paraphrase datasets has been limited, particularly for machine-generated sentences. In this paper, we present ParaGPT, a new paraphrase dataset of 81,000 machine-generated sentence pairs, including 27,000 reference sentences (ChatGPT-generated sentences), and 81,000 paraphrases obtained by using three different large language models (LLMs): ChatGPT, GPT-3, and T5. We used ChatGPT to generate 27,000 sentences that cover a diverse array of topics and sentence structures, thus providing diverse inputs for the models. In addition, we evaluated the quality of the generated paraphrases using various automatic evaluation metrics. Furthermore, we provide insights into the strengths and drawbacks of each LLM in generating paraphrases by conducting a comparative analysis of the paraphrasing performance of the three LLMs. According to our findings, ChatGPT's performance, as per the evaluation metrics provided, was deemed impressive and commendable, owing to its higher-than-average scores for semantic similarity, which implies a higher degree of similarity between the generated paraphrase and the reference sentence, and its relatively lower scores for syntactic diversity, indicating a greater diversity of syntactic structures in the generated paraphrase. ParaGPT is a valuable resource for researchers working on NLP tasks like paraphrasing, text simplification, and text generation. We make the ParaGPT dataset publicly

Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la-Fuente-Valentín contributed equally to this work.

accessible to researchers, and as far as we are aware, this is the first paraphrase dataset produced based on ChatGPT.

KEYWORDS
ChatGPT, generative artificial intelligence, large language models, machine learning

1 | INTRODUCTION

Paraphrasing, the task of expressing a given sentence differently while preserving its meaning, holds significant importance across a spectrum of applications, including text simplification, machine translation, and question-answering. NLP, referring to the capacity of computers to comprehend and produce human language (Alshater, 2022), has experienced rapid growth with an increasing demand for sophisticated language models capable of understanding, generating, and manipulating natural language text.

However, a significant challenge in NLP, as identified by Khurana et al. (2022), lies in creating models that can effectively paraphrase text. Effective paraphrasing requires not only rephrasing or rewriting text to preserve its meaning and fluency but also changing its syntactical or lexical form. Despite the paramount importance of comprehensive paraphrase datasets and rigorous evaluation methodologies for the advancement of various NLP tasks, current literature has highlighted shortcomings in both the quality of paraphrase datasets and the comparative analysis of the paraphrasing performance of LLMs. Language models like GPT-3, T5, and ChatGPT have shown remarkable performance on various NLP tasks, such as text generation, question answering, and language comprehension (Brown et al., 2020). However, there has been little research comparing their performance on the task of paraphrasing machine-generated (synthetic) sentences.

One major issue with existing paraphrase datasets is their limited scope and variability. Many datasets consist of sentence pairs that lack diversity in terms of linguistic complexity and contextual richness. They typically focus on a narrow range of linguistic variations, failing to cover the wide array of possible paraphrases (Wieting & Gimpel, 2018). Moreover, existing datasets (discussed in the Section 1.1) either do not include comprehensive evaluation results for each reference-paraphrase pair or the evaluation has often relied on a limited set of metrics, which may not fully capture the nuances of paraphrased text. For instance, traditional metrics like BLEU and ROUGE, while useful, may not adequately assess semantic similarity and fluency (Lin, 2004; Papineni et al., 2002).

The shortcomings in paraphrase datasets and comparative analysis significantly impact both research and applications. The limited diversity and complexity in existing datasets restrict the ability of models trained on these datasets to generalize beyond simplistic and repetitive paraphrasing tasks, resulting in models that perform well on standard benchmarks but fail in complex, real-world scenarios. Inadequate evaluation further hinders the objective assessment of model performance, making it challenging to identify strengths and weaknesses in existing approaches and impeding the development of improved models.

These limitations also pose challenges in real-world applications such as text simplification, machine translation, and question-answering systems. High-quality paraphrases are crucial in these applications, and poor paraphrase quality can lead to misunderstandings or misinterpretations. This is particularly detrimental in fields like education and translation services, where accurate conveyance of meaning is essential. Paraphrase quality directly impacts the performance of question-answering systems, as models that fail to generate accurate and fluent paraphrases may not understand or retrieve the correct information, leading to incorrect or irrelevant answers and diminishing user trust and system reliability.

Moreover, these shortcomings inhibit comparative studies due to the lack of rigorous comparative analysis and benchmarking challenges. The limited comparative analysis of paraphrasing performance across different LLMs means that researchers and practitioners lack detailed insights into the relative strengths and weaknesses of models like GPT-3, T5, and ChatGPT. This knowledge gap makes it difficult to select the most appropriate model for specific tasks, leading to suboptimal application and deployment in real-world scenarios. The absence of standardized benchmarks and comprehensive comparisons hinders the establishment of best practices and common standards in the field, slowing down the overall progress in NLP research as innovations and improvements are not uniformly evaluated or recognized.

This paper aims to fill this gap by conducting a comparative analysis of the paraphrasing performance of these language models using sentences generated by ChatGPT and assessing the performance of each model using the automatic evaluation metrics: BERTScore, BLEU, ROUGE, METEOR, Google-BLEU (GLEU) and T5-STSB.

The choice of synthetic data was made intentionally to meet the study's specific goals. Synthetic data allows for a controlled and reproducible setup, ensuring that all reference sentences are consistently generated.

In addition to analysing and comparing the paraphrasing performance of the language models, we introduce the ParaGPT dataset. This dataset comprises reference sentences generated by ChatGPT and their corresponding paraphrases generated by ChatGPT, GPT-3, and T5 language models. When we refer to ‘dataset quality’, we mean datasets that go beyond merely containing sentence pairs. Our definition of a high-quality dataset includes those that incorporate evaluation scores for each reference-paraphrase sentence pair, as determined by multiple evaluation metrics. Our dataset, ParaGPT, not only contains the sentence pairs but also the results of the six aforementioned evaluation metrics for each pair, making it a more comprehensive and high-quality resource.

By providing detailed performance comparisons and a high-quality dataset, this paper contributes to the advancement of NLP research and offers valuable insights for developing, training, and evaluating new paraphrase generation models. Our contributions include the creation of a synthetic paraphrase dataset with comprehensive evaluation metrics' results and the thorough comparative analysis of state-of-the-art LLMs in the context of paraphrasing. We offer open access to the ParaGPT dataset and the source code employed in our experiments: https://github.com/massyakur/ParaGPT.

1.1 | Related works

In this subsection, we handle the related works by taking three research questions into consideration: (1) whether there are ChatGPT-Generated paraphrase datasets, (2) the methodologies and techniques employed in paraphrase generation, and (3) the rationale behind incorporating the three prominent LLMs: ChatGPT, GPT-3, and T5 in our study. Hence, we divide this section into three subparts: paraphrase datasets, paraphrase generation methodologies, and selected LLMs.

1.1.1 | Paraphrase datasets

We review some of the most popular paraphrase datasets in the literature. The first and most popular paraphrase dataset is Microsoft Research Paraphrase Corpus (MRPC) (Dolan & Brockett, 2005). 5801 sentence pairs in this dataset were manually labelled as paraphrases or non-paraphrases. Its manually curated labels make it a reliable resource for training and evaluating paraphrase models, but its relatively small size limits its diversity.

Another popular paraphrase dataset is the QQP dataset or Quora Question Pairs (DataCanary, 2017), which includes over 400,000 question pairs labelled as either paraphrases or non-paraphrases. This dataset, derived from questions on Quora, serves as a rich source of real-world context, reflecting natural language usage and making it particularly suitable for tasks related to question-answering and paraphrasing. However, it is important to note that the QQP dataset is primarily confined to question structures, which can limit its applicability to a broader range of text types and applications.

The Paraphrase Database, known as PPDB (Ganitkevitch et al., 2013), is a massive paraphrase database containing over 220 million paraphrase pairs. It is automatically constructed by mining sentence alignments from bilingual parallel corpora. Its vast scale contributes to comprehensive coverage, encompassing a wide variety of domains and sentence structures. This diversity makes it a valuable resource for paraphrasing tasks across different domains. However, the dataset's generality is worth noting as it is not tailored to specific Large Language Models (LLMs). Furthermore, the dataset's sheer size, while advantageous in terms of diversity, can also lead to noise in the dataset due to its automatic construction.

In ParaNMT-50M (Wieting & Gimpel, 2018) almost 50 million English-English sentential paraphrase pairs were automatically generated as part of the dataset utilizing neural machine translation. The dataset offers a substantial volume of sentence pairs, providing an extensive resource for paraphrase generation and related NLP tasks, but the dataset's generation process is rooted in machine translation, which can introduce translation-specific biases.

ParaSCI (Dong et al., 2021) is a recent addition to the field of paraphrase datasets that specifically targets the scientific domain. It is divided into two parts: ParaSCI-ACL, which includes 33,981 paraphrase pairs sourced from scientific papers published in the ACL conference proceedings, and ParaSCI-arXiv, which includes 316,063 pairs sourced from papers on the arXiv. ParaSCI caters specifically to research in scientific text, but its domain-specificity limits its general applicability.

The Twitter URL Corpus (Lan et al., 2017) is a dataset of sentential paraphrase pairs gathered from Twitter by connecting tweets via public URLs. The dataset contains 51,524 sentence pairs, labelled both by human annotators and automatically, where only the paraphrase sentence pairs are used for paraphrase generation. Despite being automatically labelled, the corpus provides a useful resource for paraphrasing research. However, due to the automatic annotation, the dataset has noisy labels, which can impact the accuracy of any downstream NLP tasks. Nonetheless, the simplicity of the data collection method makes it an attractive option for researchers seeking large-scale paraphrase datasets.

WikiAnswer dataset (Fader et al., 2013) is composed of about 18 million question pairs that are paraphrased and aligned word-by-word, providing synonym relationships. However, this dataset is limited to questions, which narrows down the scope of the paraphrases.

MSCOCO (Lin et al., 2014) is a large-scale object detection dataset consisting of over 120K images with human-annotated captions, with each image having five captions from different annotators. Originally intended for object detection, the dataset has been found to be valuable for paraphrase-related tasks due to the descriptions of the picture's dominant subject or action. The dataset contains approximately 500K pairs of paraphrases, making it a valuable resource for paraphrase generation.

TABLE 1 Highlights of main existing paraphrase datasets and our ParaGPT.

Name	Reference	Genre	Size (pairs)	Len	Char Len
PPDB	Ganitkevitch et al. (2013)	Phrase, words	220,000,000	2.85	16.25
WikiAnswer	Fader et al. (2013)	Question	18,000,000	11.43	54.33
MRPC	Dolan and Brockett (2005)	News	5801	22.48	119.62
TUC	Lan et al. (2017)	Twitter	56,787	15.55	85.10
ParaNMT-50M	Wieting and Gimpel (2018)	Novels, laws	51,409,585	12.94	59.18
MSCOCO	Lin et al. (2014)	Description	493,186	10.48	51.56
Quora	DataCanary (2017)	Question	404,289	11.14	52.89
ParaSCI-ACL	Dong et al. (2021)	Scientific Papers	59,402	19.10	113.76
ParaSCI-arXiv	Dong et al. (2021)	Scientific Papers	479,526	18.84	114.46
ParaGPT	This paper	ChatGPT-Generated Sentences[a]	81,000	18.53	117.03

Note: *Len* means the average number of words in a sentence, while *Char Len* represents the average number of characters in a sentence.
[a]Our study employed Chat GPT Jan 9, Jan 30, Feb 9, and Feb 13 versions from 2023.

In Table 1, we summarized all these paraphrase datasets to compare to our new dataset ParaGPT. It is clear ParaGPT is the first paraphrase dataset that includes ChatGPT-Generated sentences.

1.1.2 | Paraphrase generation methodologies

In recent years, because of their ability to comprehend complex semantic and syntactic characteristics of human language, pre-trained LLMs have gained popularity for the purpose of paraphrasing. Witteveen and Andrews (2019) proposes a novel approach to generating paraphrases for various text subjects and lengths using GPT-2 (Radford et al., 2019), including paragraph-level paraphrasing, without the need to break down the text into smaller chunks.

Another study (Hegde & Patil, 2020) shows that the generation capability of GPT-2 was leveraged to produce paraphrases in an unsupervised manner and evaluate its performance against other supervised and unsupervised methods. The experiments demonstrated that the model generated high-quality and diverse paraphrases, and the use of paraphrases as data augmentation improved the classification performance in downstream tasks.

A novel approach for generating high-quality paraphrases called the latent bag of words (BOW) model was proposed in Fu et al. (2019). Quora and MSCOCO datasets were used in the experiments, and the generated paraphrases were evaluated using BLEU and ROUGE metrics.

A unified, lightweight model is put forth in Palivela (2021), demonstrating its ability to classify whether provided pairs of sentences are paraphrases of one another and to generate multiple paraphrases from a single input sentence. This is accomplished through the utilization of a fine-tuned T5 model, as described in Raffel et al. (2020). The system performs at the cutting edge on both tasks and is evaluated using popular evaluation metrics including accuracy, precision for paraphrase recognition, and also like the metrics used in our project for paraphrase generation.

In order to paraphrase text in minority classes, Patil et al. (2022) proposes an oversampling technique for imbalanced text datasets that combines the WordNet corpus and also the T5 model. The balanced augmented dataset that results boosts the effectiveness of text classification systems. A robotic process automation tool is integrated with the model to facilitate automation. For evaluation, common classifiers like the Logistic Regression algorithm are applied. The proposed approach addresses the problem of insufficient data for minority classes in imbalanced datasets.

An approach known as Quality Controlled Paraphrase Generation (QCPG) (Bandel et al., 2022) has direct control over quality dimensions. The proposed method uses a pre-trained T5-base model and an Electra base (Clark et al., 2020) model to predict quality values. The paper also presents a method to identify optimal quality control points for generating high-quality paraphrases with greater diversity than uncontrolled methods. The results show that QCPG generates high-quality paraphrases with higher diversity than the uncontrolled baseline.

In a similar study (Wahle et al., 2022), large autoregressive transformers such as GPT-3 and T5 were used to generate machine-paraphrased text for a dataset to test against automatically generated paraphrasing. To train and evaluate the model, the dataset was created from various student theses, Wikipedia, and arXiv.

To address our first research question, we conducted a literature review to determine whether there is a ChatGPT-generated paraphrase dataset. Our findings show that there is limited research on ChatGPT-generated paraphrase datasets (recall Table 1).

As given above, there has been significant research on language models and their ability to generate high-quality paraphrased sentences. One of the most popular language models is ChatGPT, which has shown promising success in generating text with zero-shot or few-shot prompting.

However, there is a lack of research on ChatGPT's ability to generate paraphrased sentences. We conducted a comparative analysis of ChatGPT's ability to generate high-quality paraphrased sentences compared to other LLMs. Our findings show that there is still a research gap. This is the first paper in which we have addressed this problem and answered it.

1.1.3 | Selected LLMs

ChatGPT, or Conditional Generative Pre-trained Transformer, is an AI technology developed by OpenAI. It utilizes supervised and reinforcement learning techniques to autonomously generate natural language conversations. Trained on millions of diverse conversations, ChatGPT captures linguistic patterns and conversational nuances, making it proficient in generating natural language text. Its unique amalgamation of pre-trained deep learning models with a programmability layer enables robust conversation generation (Bahrini et al., 2023). ChatGPT's inclusion in this study is justified by its extensive research lineage and state-of-the-art capabilities.

Generative Pre-trained Transformer 3 (GPT-3) is a deep learning-based autoregressive language model known for producing natural-sounding text. With 175 billion parameters, it is one of the most sophisticated language models, displaying high performance across various NLP tasks such as sentence completion, question answering, and language translation (Heiß et al., 2022). GPT-3 excels in few-shot learning, adapting to different tasks with minimal instruction (Brown et al., 2020). Despite challenges in datasets where few-shot learning struggles or in situations where methodological issues are associated with training on large web corpora (Brown et al., 2020), its proficiency in generating human-like text, including news articles (Goyal et al., 2023), underscores its transformative impact on NLP research and makes it a compelling choice for paraphrasing tasks.

Text-to-Text Transfer Transformer (T5) plays a prominent role in transfer learning in NLP. It utilizes pre-training on data-rich tasks before fine-tuning on downstream tasks (Raffel et al., 2023), a transformative technique in NLP. T5 introduces a unified framework that converts various text-based language problems into a text-to-text format, facilitating a wide range of NLP tasks. Achieving state-of-the-art results on benchmarks covering tasks like summarization, question answering, and text classification, the T5 research team's commitment to facilitating future research is evident through the release of datasets, pre-trained models, and code to the research community (Raffel et al., 2023). In our study, a fine-tuned T5-based model is employed to assess its paraphrasing capabilities alongside other models, providing valuable insights into how this unified text-to-text approach performs in the context of paraphrasing machine-generated sentences.

1.2 | Paper organization

Section 1 provides a brief introduction to the related works and compared language models. Section 2 explains our new dataset ParaGPT and the process of paraphrase generation and describes the automatic evaluation metrics and the relevance of each metric to our study's objectives. Section 3 presents and analyses the experimental results, and Section 4 discusses the findings and their implications. Section 5 summarizes and outlines possible future work.

2 | METHODOLOGY

2.1 | ParaGPT dataset

We present ParaGPT, a novel paraphrase dataset, we constructed using a variety of AI models operating in different domains. Specifically, we used ChatGPT to generate a corpus of 27,000 sentences. To ensure the completeness of the dataset, we selected sentences that cover a wide range of topics (genres) and sentence structures, thus providing diverse inputs for our models.

ParaGPT achieves a notable level of syntactic diversity by design. The dataset was carefully curated to encompass a broad spectrum of syntactic structures in the paraphrases it contains. Syntactic diversity, referring to the variety in the arrangement of words, phrases, and clauses in sentences, plays a crucial role in natural language understanding. The grammatical function and meaning of a sentence are dependent on this structural organization, often referred to as syntax or syntactic structure. In traditional grammar, the four basic types of sentence structures are the simple sentence, the compound sentence, the complex sentence, and the compound-complex sentence (Verma & Srinivasan, 2019). ParaGPT embraces these structural variations by incorporating diverse reference sentences generated by ChatGPT, covering an extensive array of topics (genres) and sentence constructions. This diversity in reference sentences serves as the foundation for obtaining paraphrases from the LLMs: ChatGPT, GPT-3, and T5. To achieve this syntactic diversity, different prompts were used to generate sentences with varying structures. For instance, prompts like '*Generate sentences of kind simple*', '*Generate compound – complex sentences*', '*Generate sentences with a complex conditional clause*', and so forth, were employed. The paraphrases generated by these models are inherently influenced by the reference sentences' varying syntactic structures. As a result, ParaGPT offers not only semantic diversity but also a rich tapestry of sentence constructions, ranging from simple to complex, short to lengthy, and declarative to interrogative.

While generating 27,000 reference sentences, we have selected topics (genres) from a variety of fields (genres) given in Table 2 to ensure comprehensive coverage of different fields (genres). This diversity was intentionally sought to rigorously evaluate the paraphrasing abilities of the models across varied terminology and jargon inherent to different topics (genres). The selection process was randomized but designed to ensure that every major domain was adequately represented, providing a robust testbed for assessing the adaptability and accuracy of the models in handling diverse subject matter. Section 3 of our paper provides detailed results and analysis, demonstrating how the models' paraphrasing capabilities were influenced by the diversity of topics (genres). This analysis helps to highlight the strengths and weaknesses of each model, offering valuable insights into their performance across different domains and the potential need for tailored approaches depending on the subject matter.

Furthermore, an essential consideration in our dataset construction was the inherent content moderation capabilities of the language models used, particularly ChatGPT. Given ChatGPT's built-in security filters and restrictions against generating unsafe or inappropriate content, our dataset inherently excludes hate speech, offensive language, or any toxic content. This design choice ensures that the dataset is not only comprehensive in its coverage of various topics (genres) but also adheres to ethical standards of content safety, making it a reliable resource for further research in NLP without the risk of propagating harmful biases.

We then applied the three LLMs to paraphrase all of these sentences and generated three paraphrases for each 27,000 reference sentence. One paraphrase was generated using ChatGPT, another using GPT-3, and the remaining one using the pre-trained T5 model (*prithivida/parrot_paraphraser_on_T5*) (Damodaran, 2021).

For the 27,000 reference sentences generated by ChatGPT, we extracted their corresponding 81,000 paraphrased sentences generated by each of these three models. A total of 108,000 sentences (27,000 reference sentences, 27,000 ChatGPT-generated paraphrase sentences,

TABLE 2 Topics of reference sentences ChatGPT generated given in ParaGPT.

No.	Topics	# of sentences	No.	Topics	# of sentences	No.	Topics	# of sentences
1	Medical	3998	15	Random	590	29	Exam	316
2	Law	2086	16	Scientific	571	30	Creative	312
3	Cybersecurity	1763	17	Novel	559	31	World History	268
4	Nonfiction	1386	18	Education	534	32	Science	258
5	Politics	1282	19	Book	499	33	Robotics	256
6	Question	1055	20	Hobby	465	34	Tips	239
7	News	1039	21	Weather and Env.	459	35	Trade	228
8	Tourism	1012	22	Economics	414	36	Evolution	200
9	Daily Life	993	23	Sports	403	37	Nutrition	196
10	Conversation	684	24	Philosophy	363	38	Activities	184
11	Food	653	25	Cooking	357	39	Business	153
12	Space	614	26	Culture	350	40	Social Media	143
13	Technology	614	27	Football	347	41	Random Long	130
14	Film	596	28	Product Review	334	42	Twitter	97

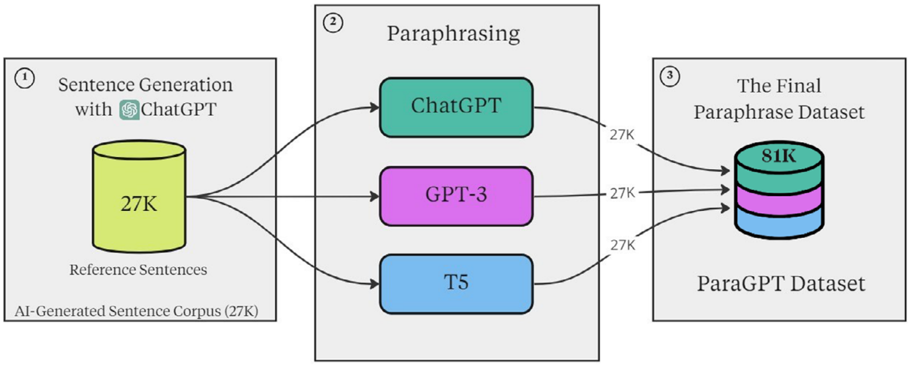
① Sentence Generation with ChatGPT
27K
Reference Sentences
AI-Generated Sentence Corpus (27K)
② Paraphrasing
ChatGPT
GPT-3
T5
27K
27K
27K
③ The Final Paraphrase Dataset
81K
ParaGPT Dataset

FIGURE 1 The general structure of ParaGPT dataset.

27,000 GPT-3 generated paraphrase sentences, and 27,000 T5-generated paraphrase sentences) were evaluated in this manner. That means ParaGPT contains 27,000 reference sentences, and 81,000 paraphrases of the reference sentences, adding up to a total of 108,000 machine-generated sentences.

In Figure 1, we give the general structure of ParaGPT dataset.

In addition to the extensive collection of reference sentences and their corresponding paraphrases, the ParaGPT dataset further enriches the repository of machine-generated sentences by incorporating a comprehensive set of automatic evaluation scores for each paraphrase. These scores are derived from a suite of widely recognized automatic evaluation metrics, including BERTScore, T5-STSB, BLEU, METEOR, ROUGE, and GLEU. This enables a robust assessment of the quality of each paraphrase, offering researchers a holistic perspective on the strengths and limitations of the machine-generated sentences. By including these evaluation scores, ParaGPT not only provides a wealth of paraphrase data but also valuable insights into the comparative analysis of the language models that produced them, enhancing its utility as a resource for NLP research and development.

The creation of ParaGPT involved a deliberate choice to use synthetic data as the foundation of our dataset. This selection was motivated by several key factors that are integral to the goals of our study.

2.1.1 | Controlled nature

Reference sentences are generated in a controlled and structured environment. In the context of paraphrasing, this control ensures that all reference sentences are consistently and systematically produced. The controlled nature of synthetic data minimizes variability, contributing to the reliability of our dataset.

2.1.2 | Resource advantages

Synthetic data offers resource advantages that are particularly valuable in constructing a large-scale dataset like ParaGPT. The efficiency of generating synthetic reference sentences allows for the creation of a comprehensive dataset with 27,000 reference sentences and their corresponding paraphrases. This resource efficiency facilitates the scalability and accessibility of ParaGPT.

2.1.3 | Diverse inputs

One of the strengths of synthetic data lies in its flexibility to generate diverse inputs across various domains and topics. With the ability to tailor the synthetic reference sentences to different areas of knowledge, ParaGPT encompasses a wide array of subject matter, ensuring that our dataset covers an extensive spectrum of content. This diversity enhances the dataset's applicability for a broad range of research and model evaluation tasks.

In summary, the choice of synthetic data for ParaGPT was a conscious decision aimed at achieving a controlled, resource-efficient, and diverse dataset, which aligns with the specific objectives of our study. This makes ParaGPT a novel resource for researchers interested in various tasks involving NLP, such as paraphrasing, text simplification, and text generation.

2.2 | Sentence generation using ChatGPT

Due to the enormous amount of data that GPT-2 and GPT-3 models have been exposed to during training, they have a very good capacity for generating sentences. So fine-tuning ChatGPT (Version of GPT language model) on sentence generation might not bring much improvement in performance. Therefore, the 27,000 sentences were generated without the need of fine-tuning the ChatGPT model.

2.3 | Paraphrasing using ChatGPT, GPT-3, and T5

In this section, we elaborate on our methodology for obtaining paraphrases for the reference sentences by leveraging the three LLMs: ChatGPT, GPT-3, and a pre-trained T5-based model known as '*prithivida/parrot_paraphraser_on_T5*', which has been fine-tuned specifically for text paraphrasing tasks.

Our process began with the manual paraphrasing of the 27,000 generated reference sentences. This operation was conducted within the ChatGPT environment, facilitated by a uniform prompt format specifying '*paraphrase* : [*sentences*]'. The outcome of this phase yielded a paraphrased sentence corresponding to each reference sentence. In Table 3, we give the corresponding paraphrased sentences of each model for one reference sentence.

Subsequently, we generated the second set of paraphrases using the GPT-3 API, specifically utilizing the Text-Davinci-003 model. Here, parameters were configured with a $temperature = 0.7$ *and* $top\_p = 1$.

For the T5-based model, accessed through *HuggingFace* − *prithivida*/*parrot_paraphraser_on_T5*, the paraphrasing prompt remained consistent. However, during the T5 paraphrasing process, it was noted that in some instances, the T5 model generated paraphrases identical to the reference sentences. In response, we applied a beam search technique to rephrase these identical sentences, thereby ensuring a diverse set of paraphrases.

As a result of these efforts, our approach yielded a total of three paraphrases for each of the 27,000 reference sentences. These paraphrases, each generated by one of the three distinct language models, constitute a core component of our ParaGPT dataset.

2.4 | Evaluation metrics and procedure

To evaluate the performance of language models on the task of paraphrase generation, the use of appropriate evaluation metrics that can effectively capture the quality of the generated paraphrases is crucial. The automatic evaluation metrics: BERTScore, T5-STSB, BLEU, METEOR, ROUGE, GLEU, and T5-CoLA are used to assess the quality of the generated paraphrases.

2.4.1 | BERTScore

BERTScore (Zhang et al., 2019) leverages the contextual embeddings from the pre-trained BERT model, calculating the cosine similarity between the embeddings of the reference and generated texts. This metric is particularly useful for assessing semantic equivalence in paraphrase tasks, providing a score that correlates well with human judgement of textual similarity.

TABLE 3 The corresponding paraphrased sentences of each model for one reference sentence.

Models	Sentences
Reference	The sound of my alarm clock ringing at 6 AM was a harsh reminder of another long day ahead, as I dragged myself out of bed and prepared for the morning rush.
ChatGPT	The blaring of my alarm at 6 AM served as a rude awakening to another long day, as I arose from bed, bleary-eyed and groggy, to face the morning rush.
GPT-3	The blaring of the alarm at 6 in the morning served as a jolting notice that the day was not going to be a short one, causing me to struggle out of bed and ready myself for the bustle of the morning.
T5	I heard my alarm clock ringing at 6:00 AM and dragged myself out of bed and prepared myself for the morning rush.
Reference	The plaintiff's claim of trademark infringement is denied by the court.
ChatGPT	The court rejects the plaintiff's claim of unauthorized use of their trademark.
GPT-3	The court rejected the plaintiff's allegation of trademark violation.
T5	The plaintiff's trademark infringement claim has been denied by the court.
Reference	How can companies use big data and analytics to improve their operations and make better business decisions?
ChatGPT	In what ways can big data and analytics be utilized by businesses to enhance their operations and make more informed decisions?
GPT-3	What strategies can businesses employ to capitalize on big data and analytics to optimize their performance and make better-informed decisions?
T5	How do I use big data in my business?
Reference	What are the potential consequences of a trade war between countries?
ChatGPT	What are the possible outcomes of a conflict over trade between nations?
GPT-3	What could be the potential impacts on countries if they enter into a trade war?
T5	What would a trade war between the US and China look like?

2.4.2 | T5-STSB

The T5-STSB metric, based on the T5 model adapted to the Semantic Textual Similarity Benchmark (STSB) (Cer et al., 2017), evaluates semantic equivalence by assigning a similarity score from 0 (no similarity) to 5 (complete equivalence). This nuanced scoring allows for a detailed assessment of how closely paraphrases capture the essence of the original sentences.

2.4.3 | BLEU

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), originally developed for assessing machine translation, computes the precision of n-grams between the candidate text and a reference text. To address situations where the generated sentence might be misleadingly short yet still achieve high precision, a Brevity Penalty is applied. This penalty reduces the BLEU score for outputs that are significantly shorter than the reference text, ensuring that overly concise paraphrases are penalized. The BLEU score is adjusted for brevity through this penalty: if a generated sentence is much shorter than the input, it receives a lower brevity score. A BLEU score approaching 1 signifies limited syntactic diversity, indicative of an inadequate paraphrase.

2.4.4 | METEOR

METEOR (Banerjee & Lavie, 2005) extends the evaluation beyond simple n-gram matching by including synonymy and stemming, offering a more nuanced measure of semantic similarity. This metric calculates unigram precision and recall, combined with a fragmentation penalty to assess the order and coherence of the text. METEOR's comprehensive approach allows it to better capture the quality of translations and paraphrases, often outperforming other metrics like BLEU in terms of correlation with human judgement.

2.4.5 | ROUGE

ROUGE measures the recall between the given reference and the generated text. It is frequently used in text summarization and machine translation tasks (Lin & Och, 2004). ROUGE-L compares how similar two sentences are based on the longest sequence of words they have in common. ROUGE-1 and ROUGE-2 measure the amount of bigrams and unigrams that are present in the generated text and the reference text, respectively.

2.4.6 | GLEU

GLEU (Palivela, 2021), a derivative of BLEU, refines the evaluation by focusing on n-gram overlaps that better align with human judgements. This metric assesses the fluency and order of n-grams in paraphrases relative to the reference, penalizing shorter average n-gram lengths in the generated text compared to the reference. GLEU's adaptations make it particularly suitable for assessing paraphrase quality in terms of structural and semantic accuracy.

2.4.7 | T5-CoLA

The T5-CoLA metric utilizes the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018) for assessing grammatical correctness. This binary classification task within the T5 framework evaluates whether sentences adhere to grammatical norms, making it crucial for analysing the syntactic integrity of generated paraphrases. By determining the acceptability of paraphrases, T5-CoLA contributes an additional layer of linguistic evaluation that complements semantic-focused metrics.

3 | EXPERIMENTAL RESULTS

As we mentioned in the previous section, we used various automatic evaluation metrics to assess the quality of the paraphrases in our ParaGPT dataset. These metrics include BERTScore with two different models (RoBERTa-large and DeBERTa-xlarge-mnli), which are known to be the best for semantic similarity and human correlation (BERTScore, 2023), as well as other standard metrics such as ROUGE, BLEU, and METEOR. We also used the T5-STSB metric, which is specifically designed for semantic textual similarity. Figure 2 shows the overall structure of our experimental setup.

3.1 | Performance comparison of paraphrasing models

In order to present the results of our evaluation, we have included tables below that display the scores of the three paraphrasing models for each sentence. These tables provide a comprehensive overview of the performance of ChatGPT, GPT-3, and T5 in generating high-quality paraphrases.

3.1.1 | BERTScore evaluation

In Table 4, we present the performance of the three LLMs on BERTScore with two different models, with and without rescaling: DeBERTa-xlarge-mnli and RoBERTa-large. According to the table, T5 outperforms the other two models on all four metrics, with scores ranging from 0.8471 to 0.9742. ChatGPT and GPT-3, on the other hand, show more mixed results. ChatGPT performs well on the BERTScore(R) (0.9686) but lags behind on the other metrics. GPT-3 has the lowest scores among the models on all metrics except for BERTScore(D), where it outperforms ChatGPT.

Overall, T5 appears to be the best-performing model across all four metrics, followed by ChatGPT and then GPT-3. It is important to note that BERTScore is just one measure of sentence-level similarity and may not capture all aspects of language generation quality. Nevertheless, the scores in this table can provide some insight into how well the three language models perform on this particular metric.

3.1.2 | Semantic similarity and diversity metrics

In addition, the semantic similarity was also evaluated using T5-STSB and METEOR metrics in Table 5. The T5-STSB metric and BERTScore are both used to evaluate the semantic similarity of two sentences. However, they use different approaches to compute the similarity score.

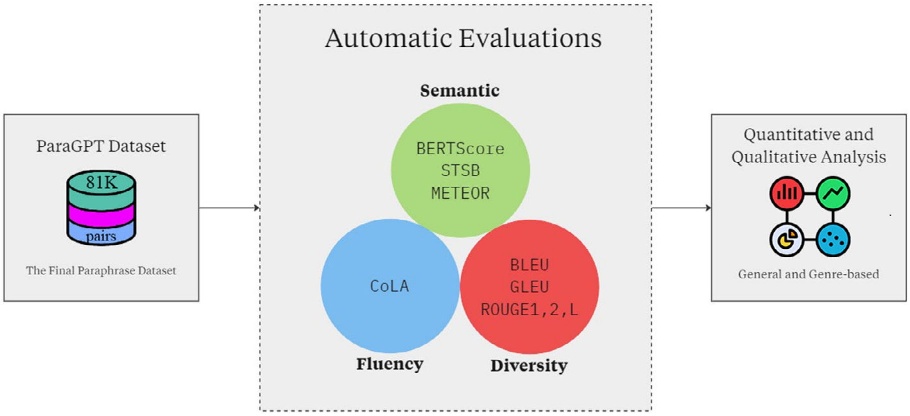

FIGURE 2 The overall structure of the experiments.

TABLE 4 Comparison of different models used in BERTScore.

	BERTScore(D)[a]		BERTScore(R)[b]	
Model	No rescale	Rescale	No rescale	Rescale
ChatGPT	0.9281	0.8517	0.9686	0.8140
GPT-3	0.8788	0.7501	0.9477	0.6901
T5	**0.9350**	**0.8660**	**0.9742**	**0.8471**

Note: Scores range from 0 to 1, and the bold values indicate the best results.
[a] DeBERTa-xlarge-mnli model.
[b] RoBERTa-large model (default).

TABLE 5 The automatic evaluation results.

Model name	T5-STSB[a]	METEOR[b]	GLEU[b]	ROUGE1[b]	ROUGE2[b]	ROUGEL[b]
ChatGPT	**4.8117**	0.7397	0.468	0.7353	0.5309	0.671
GPT-3	4.4122	0.5249	**0.2427**	**0.5213**	**0.2618**	**0.471**
T5	4.7069	**0.7958**	0.571	0.8269	0.6697	0.7769

Note: Lower GLEU and ROUGE scores indicate greater diversity. Bold values indicate the best results.
[a]Scores range from 0 to 5.
[b]Scores range from 0 to 1.

TABLE 6 The BLEU evaluation results.

Model name	BLEU	Brevity penalty
ChatGPT	0.3977	0.946
GPT-3	**0.1342**	**0.9645**
T5	0.5232	0.8993

Note: Bold values indicate the best results.

TABLE 7 T5-CoLA grammatically unacceptable paraphrases.

Name	Total
ChatGPT	29
GPT-3	49
T5	508

Note: The numbers in the 'Total' column represent the total number of unacceptable sentences (per 27,000 sentences) generated by each model on the T5-CoLA task. Bold values indicate the best results.

Therefore, it is possible for different models to perform differently on these metrics. In this case, Table 5 shows ChatGPT achieved the highest T5-STSB score, indicating that it is better at capturing semantic similarity between sentences compared to GPT-3 and T5. However, T5 had the highest BERTScore, which suggests that it excels in capturing the semantic similarity between sentences according to that metric. For all models, the scores for T5-STSB and BERTScore(R) were above 4 and 0.9, respectively, indicating that the generated paraphrases have a great semantic similarity.

In METEOR metric, T5 had the highest score, followed by ChatGPT, indicating that it is the best at capturing semantic similarity. Tables 5 and 6 also include metrics for the diversity of the generated paraphrases, such as BLEU, GLEU and ROUGE. Lower scores in these metrics indicate higher syntactic and lexical diversity. The GPT-3 model had the lowest scores in all of these metrics, indicating that it generated more diverse paraphrases compared to ChatGPT and T5.

3.1.3 | Brevity penalty and grammatical acceptability

Additionally, Table 6 reports the Brevity Penalty, which adjusts the BLEU score to penalize overly short paraphrases. This penalty ensures that paraphrases that are significantly shorter than the reference text receive a lower score, reflecting their reduced information content. In our results, the Brevity Penalty scores are given in the table, with GPT-3 obtaining the highest score. This indicates that paraphrases generated by GPT-3 are typically as long as the reference sentences, closely followed by ChatGPT. In contrast, T5 sometimes generates shorter paraphrases, which may lead to a loss of content or context.

We also evaluated the T5-CoLA metric to determine the grammatical acceptability of the generated paraphrases, which measures their fluency. Only a few paraphrases, as shown in Table 7, were considered to be unacceptable. However, the T5 model had the highest number of unacceptable paraphrases, with 508 out of 27,000. Note that for ChatGPT and GPT-3 models, the number of unacceptable paraphrases was much lower, only 29 and 49, respectively, out of a total of 27,000 paraphrases generated by each model.

Further analysis of the distribution of these grammatically unacceptable sentences across different genres is presented in Figure 3. This figure visually represents the percentage of unacceptable sentences for each model within various genres, providing a clear comparison of how grammatical acceptability varies not only by model but also by content type. For instance, the figure highlights that certain genres may pose more challenges for the models, potentially due to the specific linguistic or structural complexities associated with those genres. Such insights are invaluable for understanding model limitations and can guide future improvements in model training protocols to enhance grammatical accuracy across a broader range of topics.

In conclusion, we assessed the adequacy/semantic similarity, fluency, and diversity of the generated paraphrases. While each model has its strengths and weaknesses, T5 generally performed well in terms of semantic similarity. However, ChatGPT and GPT-3 produced more diverse paraphrases. The GPT-3 model generated paraphrases with the most syntactic diversity but had lower semantic similarity scores. All models had a high level of fluency, with only a small number of sentences being deemed unacceptable. These findings highlight the importance of using various evaluation metrics in order to obtain a comprehensive assessment of the quality of generated paraphrases.

There were instances where ChatGPT generated paraphrases that were very similar to the reference sentence and were rated as highly as T5's paraphrases. We also discovered that the input prompt used for each model, like ChatGPT and GPT-3, had a significant impact on the quality of the resulting paraphrases. In general, more specific prompts led to higher-quality paraphrases.

3.2 | Genre-based score distribution analysis

The histogram analysis of the evaluation scores across different genres, as given in Figures 4–8, revealed interesting insights into the performance of the three LLMs.

3.2.1 | Analysis of evaluation metrics across genres

Overall, the average evaluation scores of ChatGPT and GPT-3 remained relatively consistent across various genres, indicating a consistent performance across different domains. However, the T5 model exhibited a notable discrepancy in its performance, particularly evident in the significant drop in average scores for the ‘Question’ genre.

In the ‘Question’ genre, the T5 model exhibited noteworthy performance across evaluation metrics such as BLEU, GLEU, ROUGE1, ROUGE2, and ROUGEL, where lower scores indicate superior paraphrase quality. However, its performance was comparatively poorer on metrics like T5-STSB and METEOR, where higher scores signify better paraphrase quality.

We suspect that the reason why the T5 model performed exceptionally well on the ‘Question’ genre could be attributed to its fine-tuning on the ‘Quora question pairs’ dataset, among others. This dataset inclusion might have provided the T5 model with a competitive advantage in paraphrasing questions, thereby influencing its performance significantly.

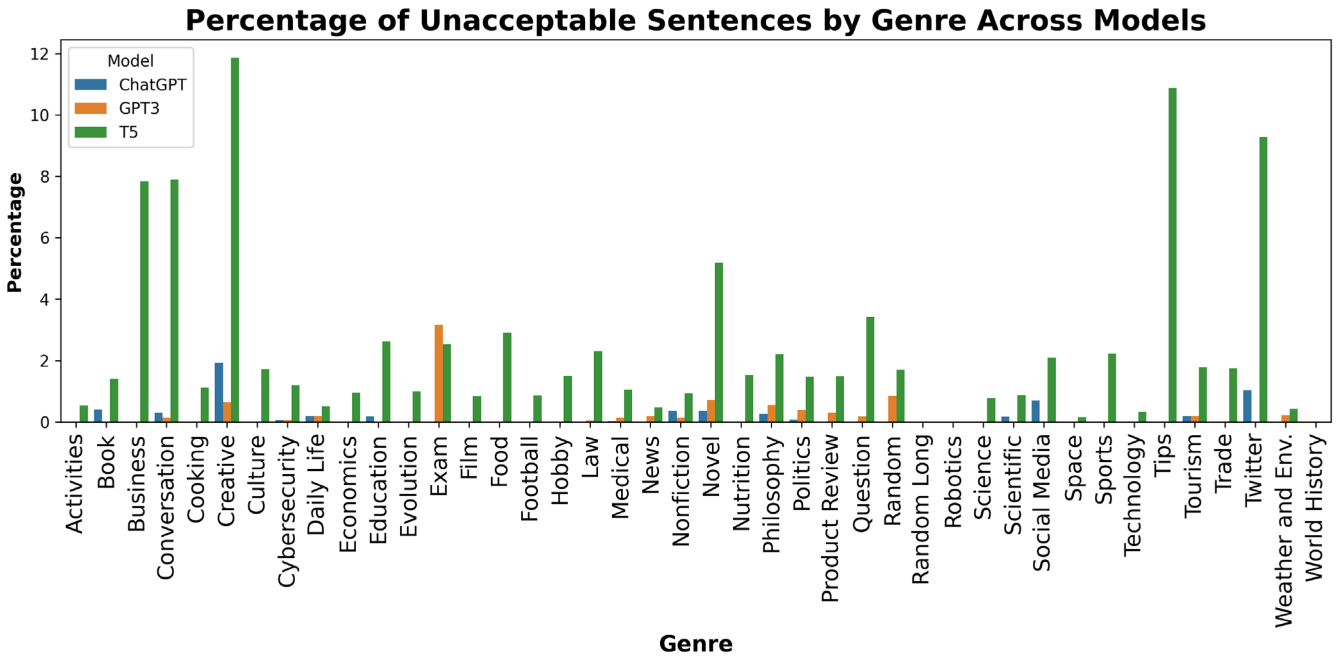

FIGURE 3 Percentage of grammatically unacceptable sentences by genre across models.

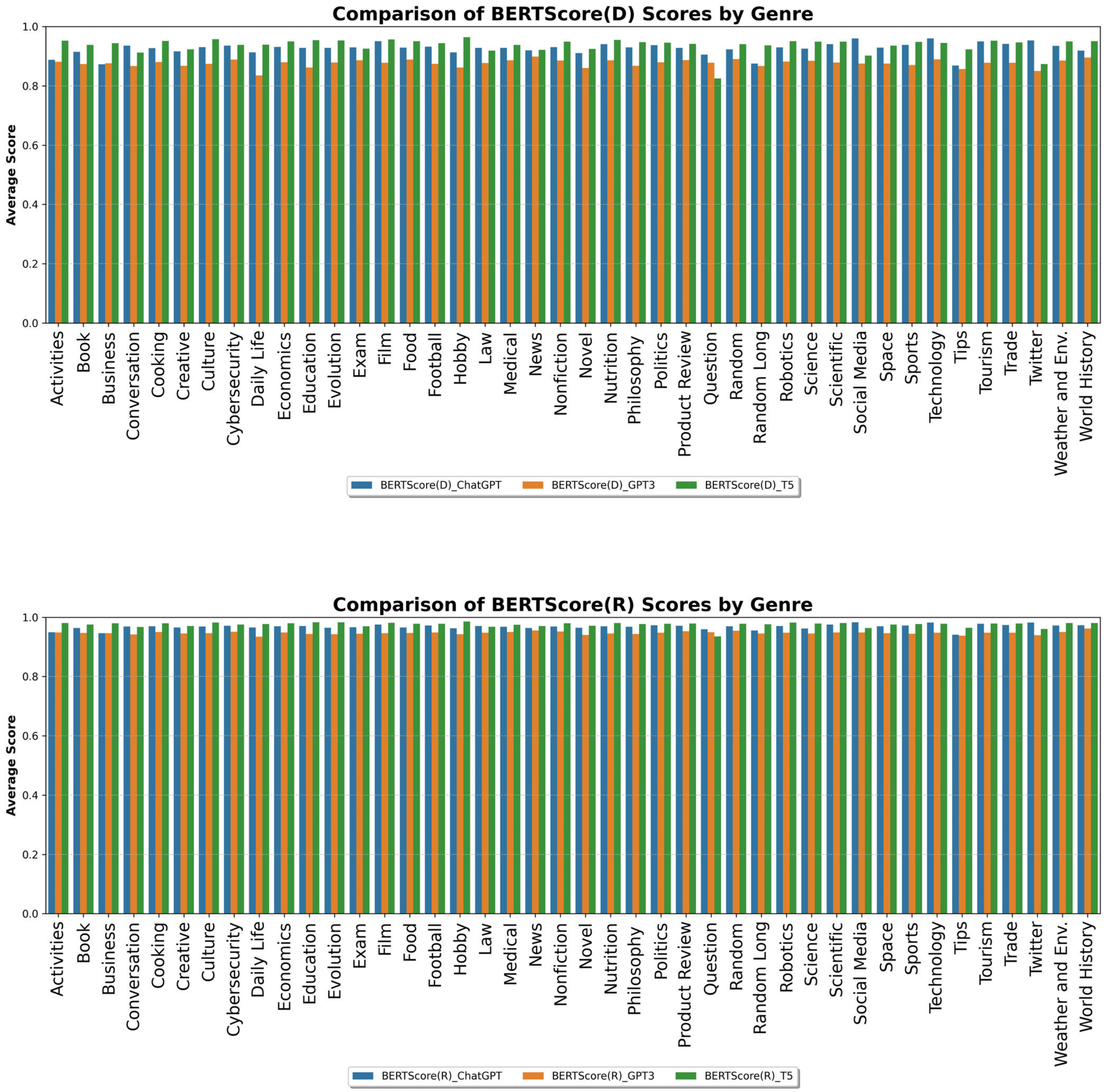

FIGURE 4 Comparison of average BERTScore(D) and BERTScore(R) scores across genres.

It is important to highlight that there was no significant genre-based discrepancy observed for BERTScore(D) and BERTScore(R) scores of T5. This nuanced analysis underscores the importance of considering multiple evaluation metrics to gain a comprehensive understanding of model performance across different text genres.

3.2.2 | Standard deviation and variability analysis

To provide a more comprehensive view of the variation in paraphrase quality across different genres, we also analysed the standard deviation of the evaluation metric results. The distribution (histogram) plot of standard deviation highlighted the variability in evaluation scores across genres for each model. Additionally, we presented the box plot of the distribution of scores across genres to better visualize the quartiles, interquartile range, and anomaly values, offering insights into the range of performance variation observed.

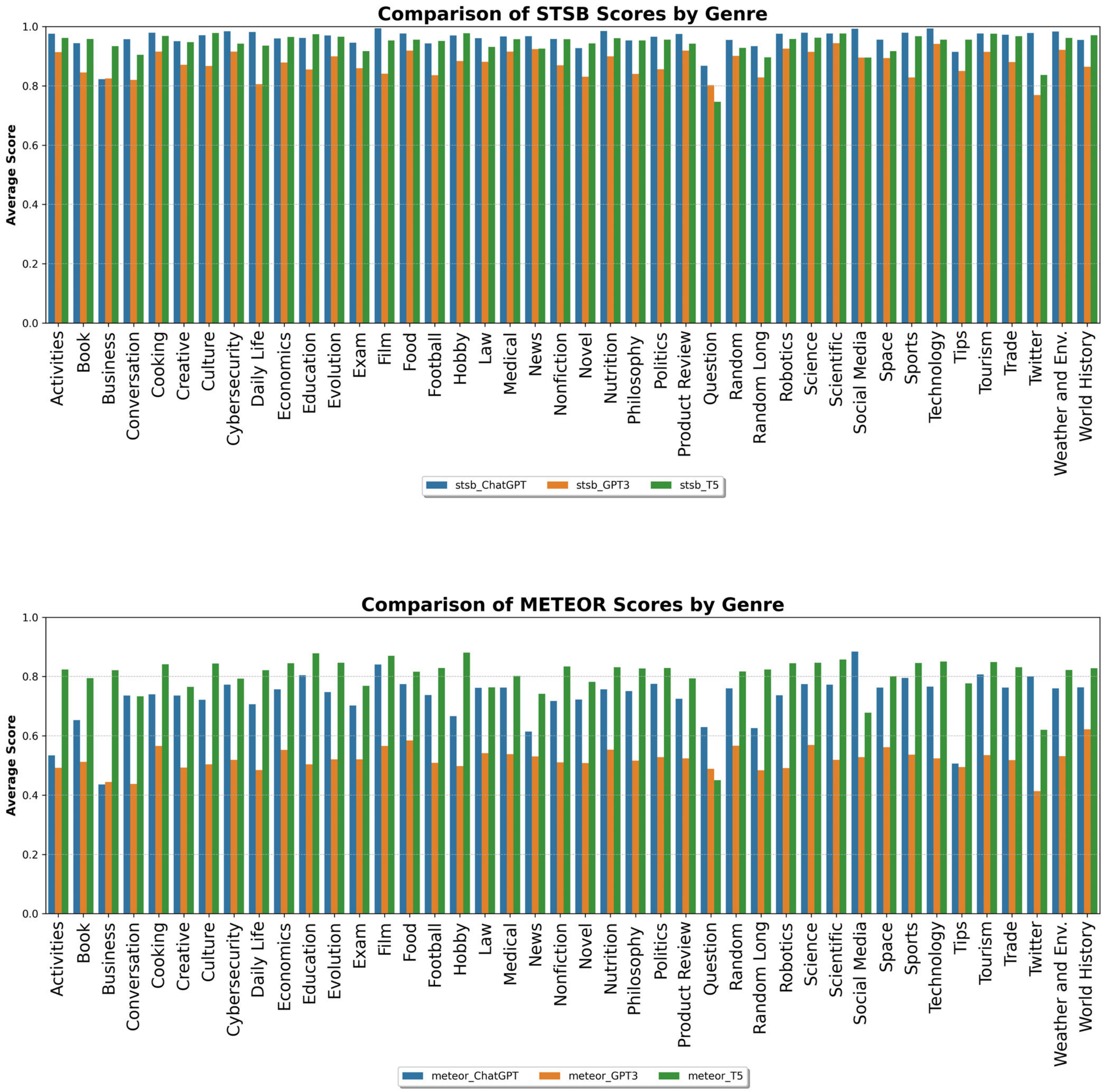

FIGURE 5 Comparison of STSB and METEOR scores across genres. STSB scores are normalized to a 0–1 scale.

As evidenced in Figure 9, an analysis of the standard deviation of scores across various genres reveals significant insights into the stability and reliability of language models. Notably, GPT3 emerges as the most stable model across the genres examined. This stability suggests that GPT3's training and architectural nuances may better equip it to handle diverse content with consistent quality, making it particularly robust in applications requiring uniform performance across varied genres (topics).

In contrast, ChatGPT and T5 display higher variability in their performance, with ChatGPT often registering the highest deviations. This observation could be attributed to operational practices in the paraphrasing tasks where sentences are processed in batches. Specifically, it was noted that ChatGPT's effectiveness in paraphrasing diminishes progressively throughout a batch. By the end of a batch, changes to sentences are minimal and less creative, indicating a potential diminution in performance quality with prolonged use in a single instance. This pattern was not observed with GPT3, suggesting a possible limitation in early ChatGPT's implementation or an inherent model fatigue that GPT3 manages to avoid.

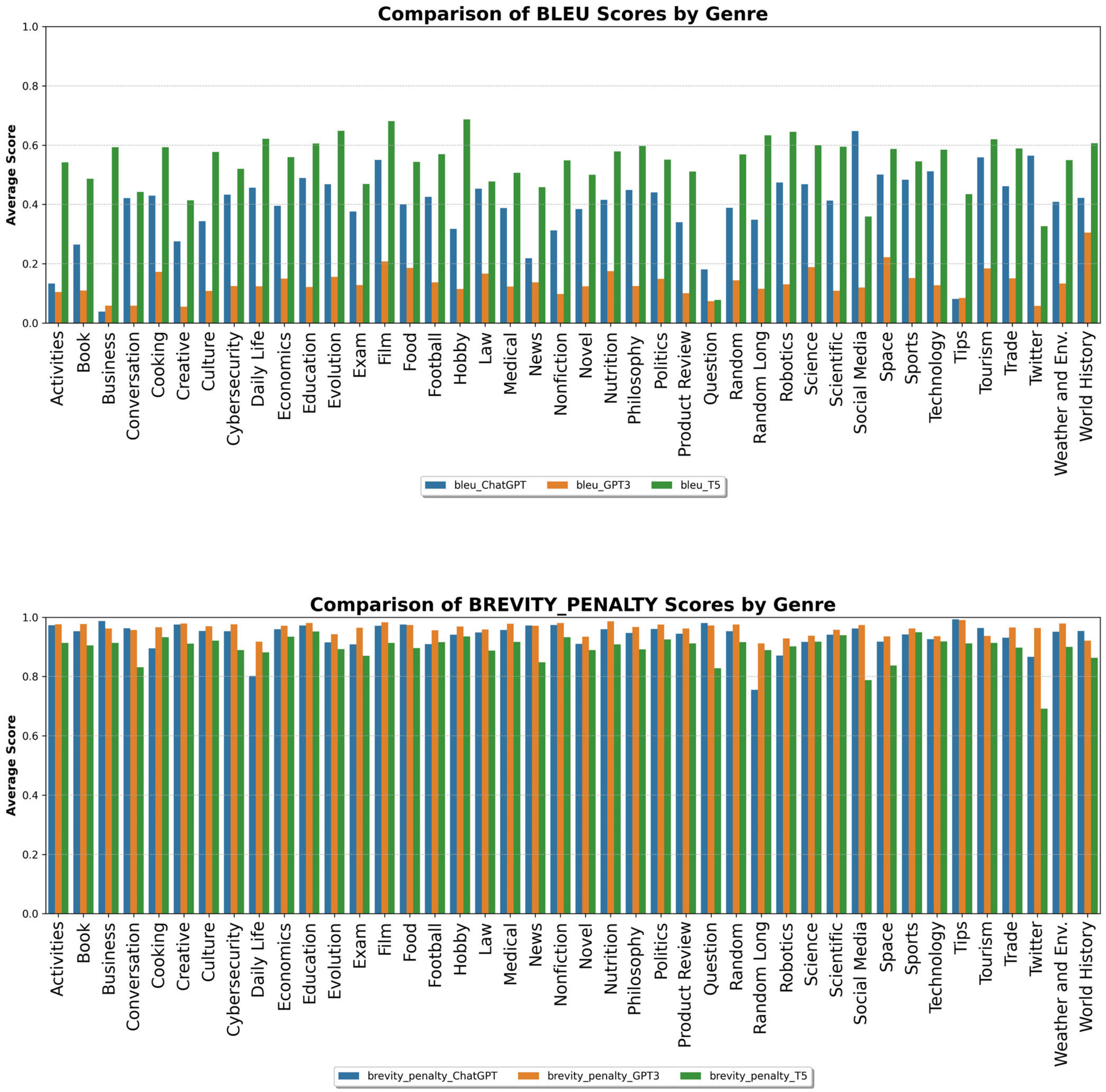

FIGURE 6 Comparison of average BLEU and BREVITY_PENALTY scores across genres.

T5, while also showing variability, does not exhibit as drastic a decline as ChatGPT, indicating a moderate level of stability that lies between ChatGPT and GPT3. The results for T5 reflect a balance, demonstrating neither the high variability of ChatGPT nor the robustness of GPT3, but a competent performance that might suit applications where moderate variability is acceptable.

The box plot illustrated in Figure 9 provides a comprehensive visual representation of the distribution of evaluation scores across various genres for each language model. It offers insights into the central tendency and variability of scores, crucial for understanding model performance across different text genres. The central line within each box denotes the median score, offering a measure of the typical score observed within each genre. The upper and lower edges of the box represent the interquartile range (IQR), encapsulating the middle 50% of the data and highlighting the variability in scores across genres. The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles, respectively, providing a visualization of the data spread. Additionally, any points beyond this range are identified as outliers, serving as indicators of potential anomalies in the data distribution.

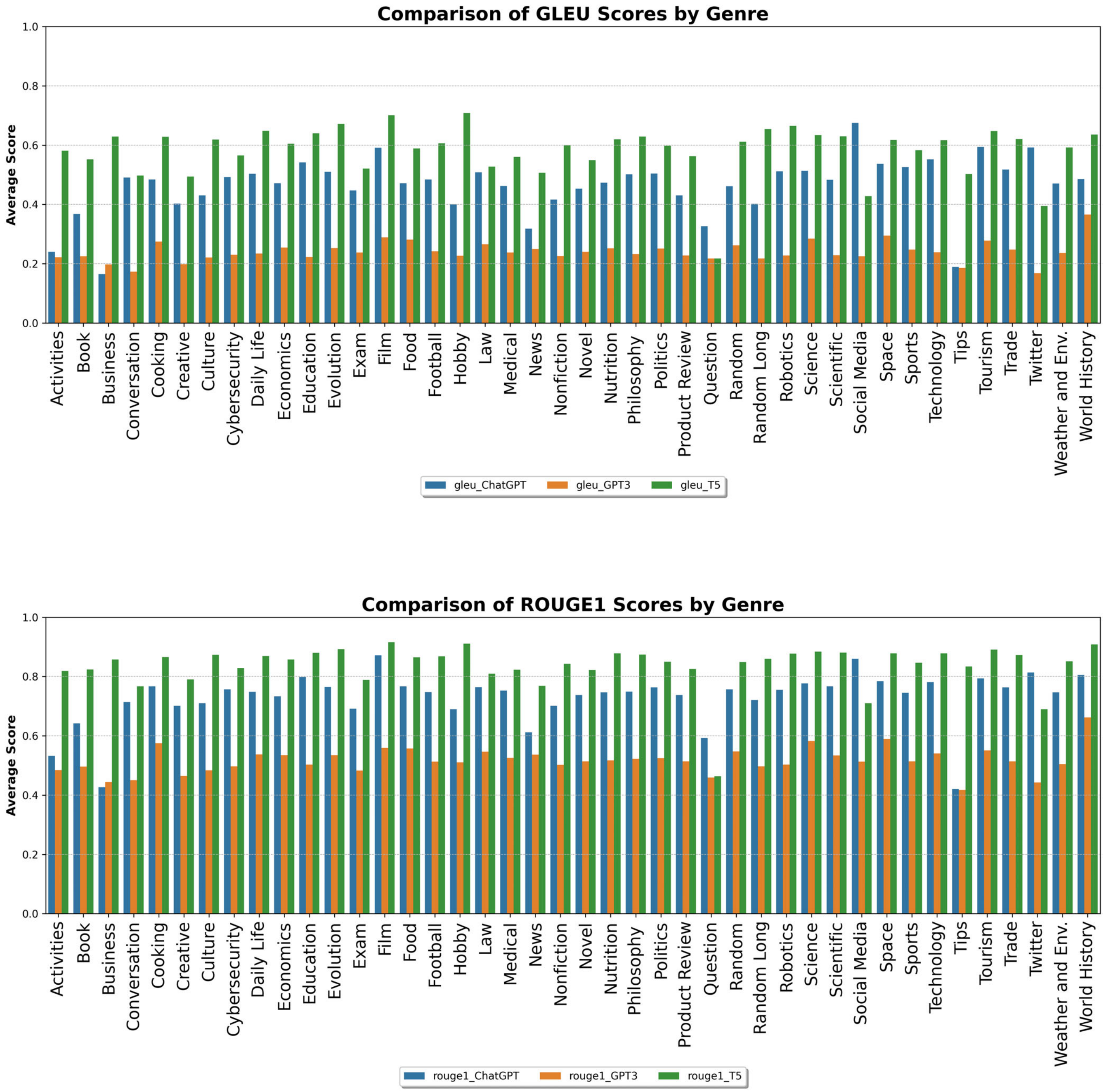

FIGURE 7 Comparison of average GLEU and ROUGE1 scores across genres.

In our analysis, we observed notable differences in the occurrence of anomalies across the three language models. ChatGPT exhibited the highest number of anomalies, totalling 43 across all evaluation metrics, indicating instances where the model's performance deviated significantly from the norm. T5 followed with 25 anomalies, while GPT-3 demonstrated the lowest number of anomalies, with only 19 instances recorded. This observation underscores the importance of identifying and addressing anomalies to ensure the reliability and robustness of model performance assessments across diverse text genres.

In terms of interquartile range (IQR), ChatGPT exhibits the widest spread of evaluation scores across genres, followed by T5, and then GPT-3, which has the narrowest range. This indicates that ChatGPT has the most variability in evaluation scores across different genres, followed by T5, and then GPT-3. In other words, ChatGPT's performance varies more widely across different genres compared to the other models.

Specifically, we observed significant variations in performance for ChatGPT across genres such as 'Business' and 'Tips'. For T5, notable variations were observed in genres like 'Question', 'Social media', and 'Twitter', with the 'Question' genre showing the most significant variation among all genres analysed for the three models.

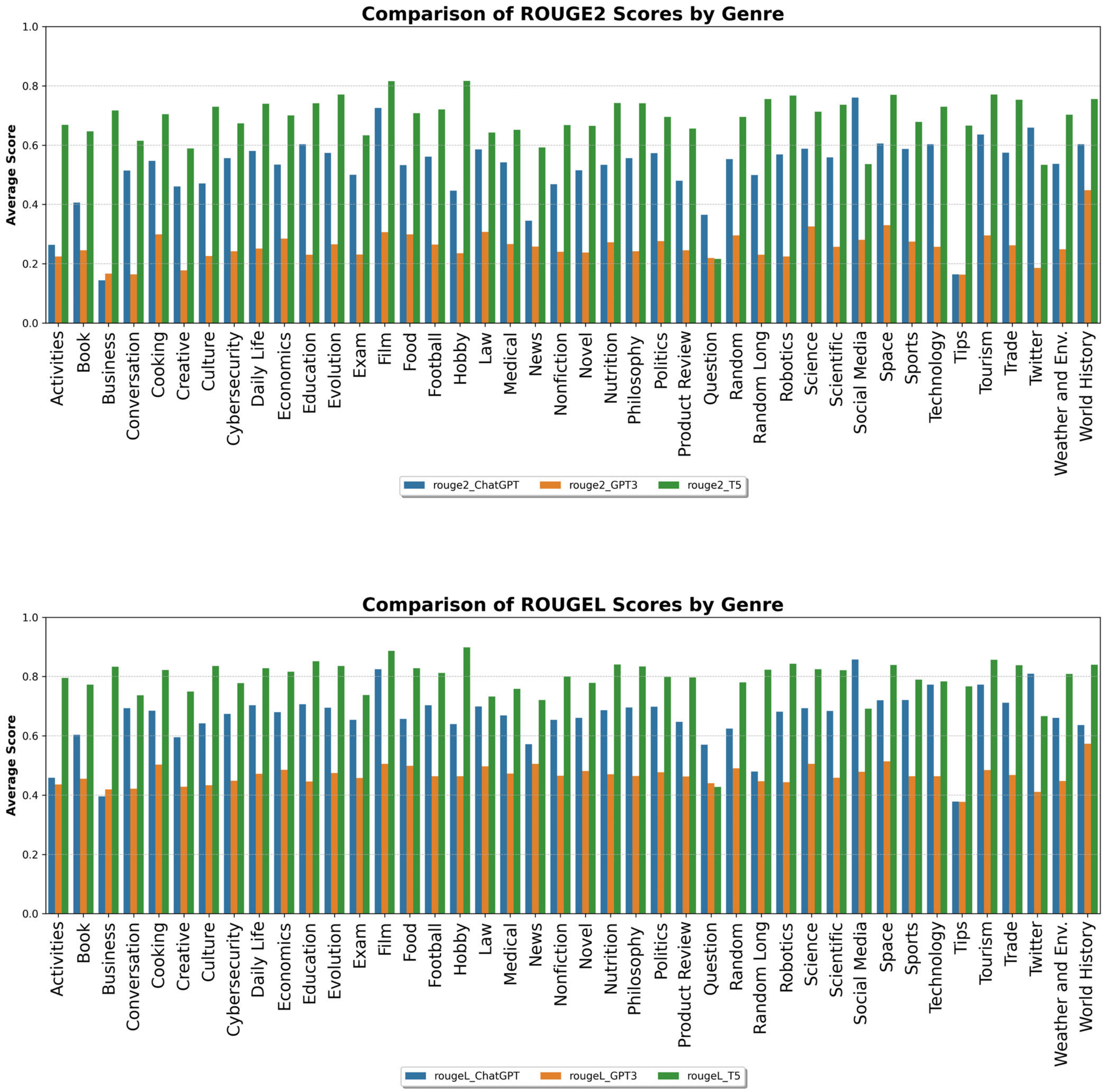

FIGURE 8 Comparison of average ROUGE2 and ROUGEL scores across genres.

4 | DISCUSSION

Our study represents a significant contribution to the field of NLP, providing a large-scale resource for training and evaluating paraphrase models. Our approach of combining multiple paraphrasing methods has resulted in a dataset that exhibits high fluency, diversity, and coverage of various types of paraphrases.

The utilization of APIs for T5 and GPT-3 models gives our method an edge, offering a distinct advantage in rapidly generating a large volume of paraphrases. However, one limitation we encountered is that the GPT-3 API is a paid service, and by the time our dataset was created (from January to February 2023), there was no official public API access to ChatGPT. Due to this unavailability, we had to resort to using the ChatGPT's web interface, requiring manual intervention to generate paraphrases using prompts multiple times. This manual approach significantly slowed down the process compared to using an API, impacting the efficiency of our method. Additionally, the high demand for ChatGPT's web interface can further exacerbate the delay in generating paraphrases. These challenges limit the accessibility of these models to a wider audience and can

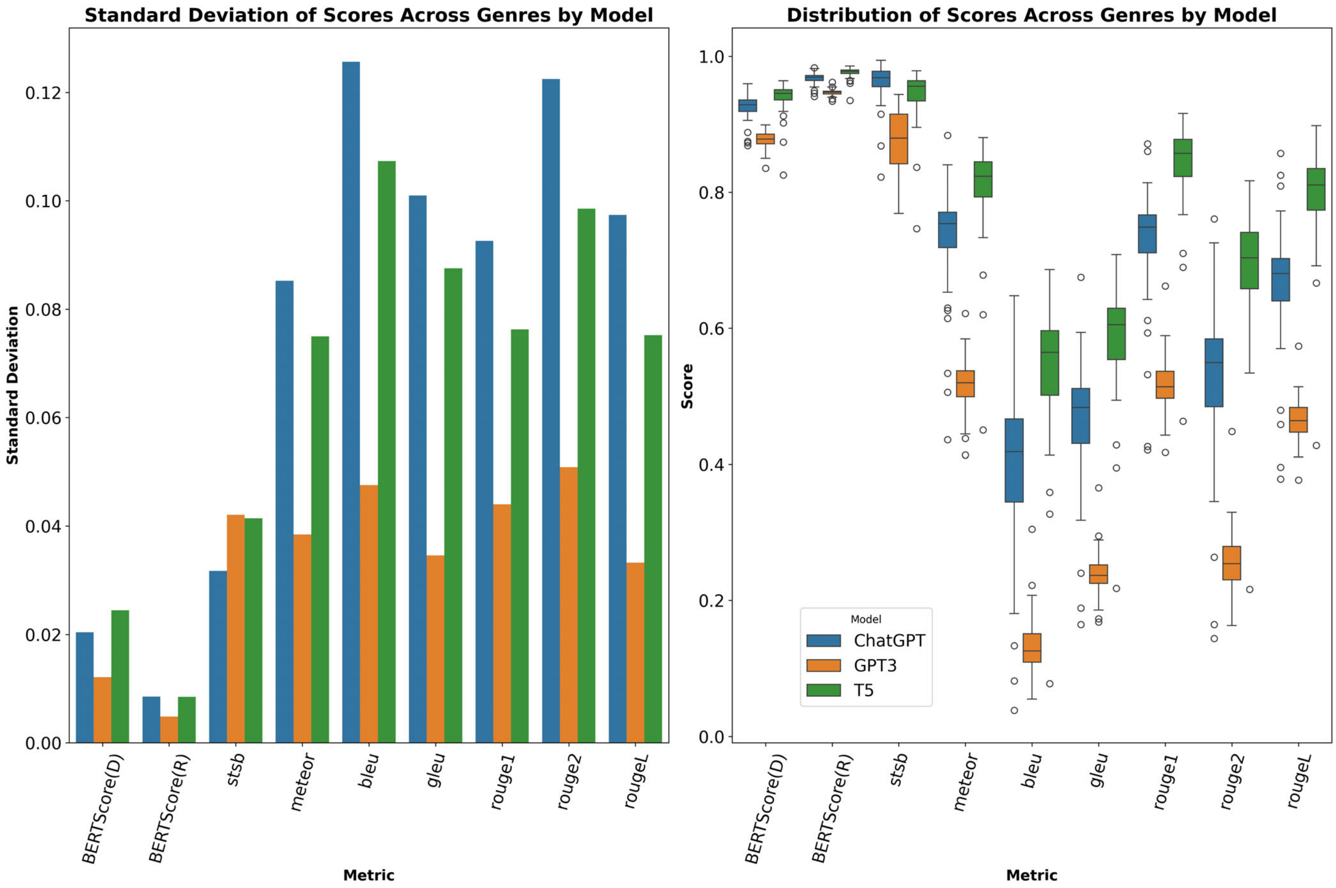

FIGURE 9 Standard deviation and box plot of scores across genre by model.

make it challenging to apply them to large-scale NLP tasks. Nonetheless, the results of our study demonstrate that, when used appropriately with better prompts and diverse domains, ChatGPT is a powerful tool for generating high-quality paraphrases. While our evaluation shows that ChatGPT can generate high-quality paraphrases in some cases, we found that the model's performance is inconsistent and can sometimes produce low-quality paraphrases that are very similar to (replica of) the original sentence. This could be due to the training data's limited scope, which may not cover all possible sentence structures and variations. To get the best results from ChatGPT, we recommend using carefully crafted prompts that are tailored to the specific task at hand. By providing more specific instructions, users can guide ChatGPT towards generating more accurate and diverse paraphrases and avoid some of the issues we observed with this model. As a result, future work could explore the use of other publicly available language models for paraphrase generation. Another potential direction for future research is to use this dataset to fine-tune language models, which could lead to better paraphrase generation in different domains.

The ParaGPT dataset is unique in that it contains paraphrases of the machine (ChatGPT) generated sentences from a diverse range of domains. This is important because many existing paraphrase datasets are limited in scope, which can hinder their applicability to real-world scenarios. The ParaGPT dataset contains paraphrases from domains such as finance, technology, and healthcare, making it a valuable resource for a variety of NLP tasks.

Another advantage of the ParaGPT dataset is that it can be used to train and evaluate models that can generate paraphrases in a wide range of contexts. The dataset includes paraphrases that involve lexical substitutions, syntactic transformations, and semantic rephrasing, providing a rich set of examples for models to learn from.

One limitation of our approach is that it relies on existing language models to generate paraphrases. While state-of-the-art models such as ChatGPT, T5, and GPT-3 have demonstrated impressive performance on a variety of natural language tasks, they may not be perfect in capturing all possible types of paraphrases. As such, it is important to continue exploring new methods for generating paraphrases and evaluating their effectiveness.

Overall, we believe that the ParaGPT dataset has the potential to facilitate progress in the development of high-quality paraphrase models, as well as other NLP tasks. We hope that the dataset will be a valuable resource for the research community and encourage further exploration of its applications in both academic and industrial settings.

4.1 | Automatic evaluation

We observed that some metrics, such as BLEU, tended to overestimate the quality of the paraphrases. This highlights the importance of using multiple evaluation metrics and not relying on a single metric to assess the quality of paraphrases.

The use of automatic evaluation metrics also facilitated the comparison of the performance of various paraphrasing methods. We observed that ChatGPT and T5 generally produced higher-quality paraphrases than the GPT-3 API. However, we also observed that the performance of each method differed depending on the specific domain of the paraphrase. This highlights the importance of testing and evaluating paraphrase models in multiple domains.

Our analysis of the paraphrasing performance of ChatGPT, GPT-3, and T5 language models, using synthetic reference sentences generated by ChatGPT, revealed valuable insights. As we delve into the discussion, we address concerns regarding the choice of synthetic data as reference sentences and offer justifications for its use.

4.2 | Justification of synthetic data

4.2.1 | Control and reproducibility

The selection of synthetic data for creating the ParaGPT dataset is grounded in the principles of control and reproducibility. Synthetic data offers a controlled environment where we can precisely generate reference sentences, ensuring consistency in the dataset. This controlled setting significantly reduces variability in reference data, thereby enhancing the reliability of our results. In essence, the use of synthetic data ensures that all reference sentences are consistently generated, contributing to the overall robustness of our study.

4.2.2 | Real-world paraphrasing versus model evaluation

It is essential to clarify the primary focus of our study, which is to evaluate the performance of language models (ChatGPT, GPT-3, and T5) in generating paraphrases. We do not claim that synthetic data is a perfect representation of real-world data. Rather, synthetic data serves as a valuable resource for assessing model capabilities in a controlled setting. By employing synthetic data, we create a standardized benchmark for evaluating paraphrasing performance, enabling us to measure the progress and capabilities of language models systematically. This approach enhances the transparency and reproducibility of our findings, supporting rigorous model evaluation.

4.2.3 | Comparison to real data

While our study leverages synthetic reference data, we recognize the value of exploring the gap between model performance on synthetic references and actual language usage. Future research endeavours may include the incorporation of real-world data, allowing for more extensive comparative analyses to gain a more comprehensive understanding of model capabilities.

In conclusion, we emphasize the contributions and limitations of our study. The use of synthetic reference data offers several advantages, including controlled and reproducible evaluations, resource efficiency, and diverse inputs. These benefits enhance the accessibility and scalability of the ParaGPT dataset. However, it is important to acknowledge that synthetic data also comes with constraints, such as potential deviations from real-world language usage. This nuanced understanding of the dataset's characteristics enables researchers to leverage its strengths while remaining mindful of its limitations.

5 | CONCLUSION

In this study, we introduced ParaGPT, a paraphrase dataset that consists of sentences generated by ChatGPT and their paraphrases produced by three state-of-the-art language models. We also provide evaluation scores for each reference-paraphrase pair using six different automatic evaluation metrics, making our dataset useful for various Natural Language Processing (NLP) applications.

Our findings underscore the potential of AI-generated paraphrases and affirm the high quality of the dataset, making it a valuable resource for diverse NLP tasks. However, it is essential to meticulously assess the quality of generated paraphrases and select the most appropriate model according to the specific task requirements.

Based on our results, we offer guidance on model selection:

- *For tasks requiring syntactic diversity*: GPT-3-generated paraphrases are recommended due to their capacity to introduce syntactic variety.
- *For tasks mandating close semantic and structural parity*: T5 emerges as an excellent choice, as it generates paraphrases closely resembling reference sentences in both wording and structure.
- *For tasks demanding both high semantic similarity and syntactic diversity*: ChatGPT excels by producing paraphrases that strike a balance between semantic similarity and syntactic/lexical diversity.

Notably, ChatGPT stands out as the model yielding the highest-quality paraphrases, as confirmed by our automatic evaluation metrics. However, enhancing the quality of results is dependent on using effective prompts.

As part of our future work, we plan to explore advanced prompts, alternative language models, and fine-tuning techniques to further elevate the quality of our paraphrases. This commitment to continuous improvement will ensure that ParaGPT remains a reliable resource for NLP research and development.

As a significant line of future research, we propose the following key areas:

5.1 | Expanding the comparative analysis scope

In future iterations of our research, we aim to broaden the scope of our comparative analysis by incorporating other language models such as Gemini, Llama, Claude, Grok, and more. This expansion will not only enhance the comprehensiveness of our study but also provide deeper insights into the diverse landscape of paraphrasing techniques and model performances.

5.2 | Exploring domain-specific paraphrasing

Future research can delve into the efficacy and challenges of generating paraphrases in specific domains, such as medical, legal, or technical texts, using Large Language Models (LLMs). Domain-specific paraphrasing poses unique challenges due to specialized vocabulary and strict syntactic and semantic constraints. Researchers can work on developing and evaluating paraphrase datasets that adhere to domain-specific constraints while maintaining high-quality paraphrase generation. This can involve fine-tuning LLMs on domain-specific corpora and considering the development of domain-adapted evaluation metrics tailored to the nuances of domain-specific language. Additionally, it is worth exploring the utility of generated paraphrases in domain-specific NLP applications, demonstrating the practical value of such datasets. This approach will extend the applicability of our research findings to more specialized and challenging paraphrase-generation scenarios, increasing the impact of our work.

5.3 | Mitigating bias in paraphrase datasets

Another significant area for future research is the identification and mitigation of bias in paraphrase datasets. This endeavour is crucial for ensuring that data used for training and evaluating paraphrase generation models are both representative and unbiased. Biases in paraphrase datasets can emerge from various sources, including the original texts and the methods employed to generate paraphrases, and may inadvertently lead to models that perpetuate these biases. Future research should focus on the development of methodologies and frameworks that systematically identify, quantify, and mitigate biases in paraphrase datasets. The goal is to ensure that these datasets provide a fair and unbiased representation of language, free from favouring or disadvantaging any particular group or perspective. Such scrutiny should extend to both the explicit content and implicit biases and assumptions encoded in paraphrases.

In conclusion, this research contributes not only to the field of NLP but also presents a valuable resource that can support and guide future studies across a range of domains and applications. By addressing these research questions and extending our work to specialized domains while ensuring fairness and inclusivity, we can foster advancements in paraphrase generation and the development of high-quality natural language generation models.

ACKNOWLEDGEMENTS

Within the scope of the TUBITAK ULAKBIM-Wiley Read & Publish agreement, Kocaeli University, a member institution of TUBITAK ULAKBIM, provides support for the open-access publication of this research.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in ParaGPT at https://github.com/massyakur/ParaGPT.

ORCID

Meltem Kurt Pehlivanoğlu https://orcid.org/0000-0002-7581-9390
Vimal Shanmuganathan https://orcid.org/0000-0002-1467-1206
Luis de-la-Fuente-Valentín https://orcid.org/0000-0001-9727-315X

REFERENCES

Alshater, M. (2022). Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *Available at SSRN*.

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). *ChatGPT: Applications, opportunities, and threats*.

Bandel, E., Aharonov, R., Shmueli-Scheuer, M., Shnayderman, I., Slonim, N., & Ein-Dor, L. (2022). Quality controlled paraphrase generation. *arXiv preprint arXiv:2203.10940*.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the Acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

BERTScore. (2023). *BERTScore default layer performance on WMT16*. https://docs.google.com/spreadsheets/d/1RKOVpselB98Nnh_EOC4A2BYn8_201tmPODpNWu4w7xI/edit#gid=0

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. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2001, https://aclanthology.org/S17-2001

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Damodaran, P. (2021). *Parrot: Paraphrase generation for NLU*.

DataCanary, L.J.M.R.N.D.t. (2017). *hilfialkaff: Quora question pairs*. Kaggle. https://kaggle.com/competitions/quora-question-pairs

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*. https://aclanthology.org/I05-5002

Dong, Q., Wan, X., & Cao, Y. (2021). Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume* (pp. 424–434).

Fader, A., Zettlemoyer, L., & Etzioni, O. (2013). Paraphrase-driven learning for open question answering. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (Long Papers) (Vol. 1, pp. 1608–1618). Association for Computational Linguistics. https://aclanthology.org/P13-1158

Fu, Y., Feng, Y., & Cunningham, J. P. (2019). Paraphrase generation with latent bag of words. In *Proceedings of Advances in Neural Information Processing Systems*, *32* (pp. 13645–13656).

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 758–764). Association for Computational Linguistics. https://aclanthology.org/N13-1092

Goyal, T., Li, J. J., & Durrett, G. (2023). *News summarization and evaluation in the era of GPT-3*.

Hegde, C., & Patil, S. (2020). Unsupervised paraphrase generation using pre-trained language models. *arXiv preprint arXiv:2006.05477*.

Heiß, S., Gierl, P., Rappl, V., Ardaya-Lieb, S., Möhring, C., Solisch, M., Solisch, T., Garschhammer, K., Kammerl, A., & Kolb, M. (2022). *Fms-berichte sommersemester 2022: Seminar zu aktuellen themen der elektro-und informationstechnik*.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744.

Lan, W., Qiu, S., He, H., & Xu, W. (2017). A continuously growing dataset of sentential paraphrases. In *Proceedings of The 2017 conference on empirical methods on natural language processing (EMNLP)* (pp. 1235–1245). Association for Computational Linguistics. http://aclweb.org/anthology/D17-1127

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out*.

Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-04)* (pp. 605–612). Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014. Proceedings, Part V* (Vol. 13, pp. 740–755). Springer.

Palivela, H. (2021). Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, *1*, 100025. https://doi.org/10.1016/j.jjimei.2021.100025

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Patil, A. P., Jere, S., Ram, R., & Srinarasi, S. (2022). T5w: A paraphrasing approach to oversampling for imbalanced text classification. In *2022 IEEE international conference on electronics, computing and communication technologies (CONECCT)* (pp. 1–6). IEEE. https://doi.org/10.1109/CONECCT55679.2022.9865812

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the limits of transfer learning with a unified text-to-text transformer*.

Verma, G., & Srinivasan, B. V. (2019). *A lexical, syntactic, and semantic perspective for understanding style in text*.

Wahle, J. P., Ruas, T., Kirstein, F., & Gipp, B. (2022). How large language models are transforming machine-paraphrased plagiarism. *arXiv preprint arXiv:2210.03568*.

Warstadt, A., Singh, A., & Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Wieting, J., & Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics* (Long Papers) (Vol. 1, pp. 451–462). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1042, https://aclanthology.org/P18-1042

Witteveen, S., & Andrews, M. (2019). Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

AUTHOR BIOGRAPHIES

Meltem Kurt Pehlivanoğlu received the B.Sc. and M.Sc. degrees in computer engineering from Trakya University, Edirne, Turkey, in 2010 and 2013, respectively, and the Ph.D. degree in computer engineering from Kocaeli University, Kocaeli, Turkey, in 2018. She held a postdoctoral fellowship from TÜBİTAK (Scientific and Technological Research Council of Türkiye). In 2019, she was a postdoctoral researcher at The Security of Advanced Systems Research Group, The University of Sheffield, for 8 months, and in 2022 for 4 months, totaling 1 year. She has been an Assistant Professor with the Department of Computer Engineering, Kocaeli University, since 2020. Her research interests include lightweight cryptography, secure novel intelligent systems, artificial intelligence, and machine learning.

Robera Tadesse Gobosho received his B.Sc. degree in Computer Engineering with high honors from Kocaeli University, Turkey, in 2024. Born and raised in Ethiopia, he completed his high school education there before pursuing his undergraduate studies in Türkiye. Robera has actively engaged in applying AI to embedded systems, participating in TÜBİTAK (Scientific and Technological Research Council of Türkiye) funded projects and Teknofest (Aerospace and Technology Festival, Turkey) competitions. He also gained international experience through an internship in Germany, focusing on embedded systems. In 2017, he won local and national science fair competitions in Ethiopia for robotics, demonstrating his early commitment to technological innovation. His research interests include machine learning and the application of AI in embedded systems.

Muhammad Abdan Syakura received a B.Sc. degree in Computer Engineering from Kocaeli University, Turkey, in 2024. Originally from Indonesia, he has focused his research on Natural Language Processing (NLP), Large Language Models (LLMs), Artificial Intelligence (AI), Machine Learning (ML), Cybersecurity, and Web Technologies. During his undergraduate studies, he published several research papers in international conferences, with work featured in IEEE. As a research intern at the National Research and Innovation Agency (BRIN) in Indonesia, he contributed to the Aeronautics and Space Research Organization's Rocket Technology Research Center – Guidance Laboratory. There, he focused on developing solutions for rocket soft landing guidance trajectories using optimal control, convex optimization, and Reinforcement Learning. His primary research focus is on advancing AI and ML technologies to address complex challenges.

Vimal Shanmuganathan is working as Professor, Department of Artificial Intelligence & Data science, Sri Eshwar College of Engineering, Tamilnadu, India. He has around Eighteen years of teaching experience, EMC certified Data science Associate and CCNA Certified professional too. He holds a Post. doctoral Fellow from Federal Institute of Science Education and Technology, Brazil, Ph.D in Information and Communication Engineering from Anna University Chennai and he received Masters Degree from Anna University Coimbatore. His areas of interest include Game Modelling, Artificial Intelligence, Cognitive radio networks, Network security, Machine Learning and Big data Analytics. He is a Senior member in IEEE and holds membership in various professional bodies. He has hosted 22 special issues in IEEE, Elsevier, Springer, Wiley and CMC tech science journal also hosted 3 International conference indexed by Scopus. He is a Senior Member in IEEE. He have been listed among Top 2% scientist in AI by Stanford University USA for the year 2022 & 2023

Luis De La Fuente Valentin received the Ph.D. degree in telematics engineering from the Universidad Carlos III de Madrid. He is currently a Postdoctoral Senior Researcher with the Universidad International de La Rioja, Spain, in the framework of the Vice-Rectorate for Knowledge Transfer and Technology. He has involved in the European projects INTUITEL and Hotel. His research interests include technology-enhanced learning strong background in learning standards and specifications, learning analytics, information visualization, student centered learning systems, recommendations, mobile learning, and gamification techniques.

How to cite this article: Kurt Pehlivanoğlu, M., Gobosho, R. T., Syakura, M. A., Shanmuganathan, V., & de-la-Fuente-Valentín, L. (2024). Comparative analysis of paraphrasing performance of ChatGPT, GPT-3, and T5 language models using a new ChatGPT generated dataset: ParaGPT. *Expert Systems*, *41*(11), e13699. https://doi.org/10.1111/exsy.13699