[1]Praloy Biswas
[2]Subhrendu Guha Neogi
[3]A. Daniel
[4] A. Mitra

Evaluating Generative Artificial Intelligence on Multilingual Sentiment Analysis

Journal of Electrical Systems

Abstract: - This study shows that natural language processing (NLP), large language models (LLM), and generative Artificial Intelligence are becoming the most important. LLM is considered essential for many NLP tasks, including question-answering, classification, interpretation, and writing. New LLMs must be able to analyze and create documents in different languages as they know how to train training materials in different languages such as ChatGPT, BLOOMZ, and others. Considering how frequently LLMs are used, it is critical to assess the effectiveness in multi-lingual environments. In a zero-shot context, the present generative models are still useful in producing text in Indian languages. On the other hand, generative models routinely outperform human quality-based evaluation when it comes to English language generation and Indian languages. LLMs are not meant to be used in a zero-shot manner in downstream applications due to poor generating performance.

Keywords: NLP, LLMs, ChatGPT, BLOOMZ.

I. INTRODUCTION

The researchers have shown great interest in Large Models (LLMs) like ChatGPT and GT-4 because of the significant advancements in their capabilities, which include reasoning, fluency of generation, and context maintenance throughout discussions. Numerous users have reported testing these systems in languages other than English, with different degrees of success. Additionally, recent demonstrations of these models (Warren, 2023) have been presented in several languages, although ones with a lot of resources.

Multilingual sentiment analysis is challenging for machine learning due to linguistic and cultural differences across languages. Generative AI models like large language models can help overcome this by leveraging large datasets in multiple languages for training. Key evaluation metrics include sentiment classification accuracy, contextual semantic understanding, and calibration across languages. Models should be evaluated on representative datasets in each target language. Major benchmarks include multilingual Amazon Reviews, Multilingual Sentiment Twitter, and others spanning languages like English, Chinese, Hindi etc. Limitations of current generative models include poorer performance on low-resource languages, difficulty capturing nuances, and bias issues. It might be beneficial to enhance model designs and training data. Future work is needed to improve model robustness, mitigate biases, adapt better to new languages and domains, and enhance understanding of linguistic/cultural nuances. More diverse multilingual datasets and testing are important.

The MMLU multiple choice questions benchmark was recently used to assess the GPT-4 model (OpenAI, 2023) by automatically translating it into 26 languages. It was discovered that a few Latin-script low-resource languages showed great promise. Their multilingual capabilities originate in these models' pre-training data, which includes hundreds of millions of non-English tokens even in big, primarily English corpora (Blevins and Zettlemoyer, 2022). Documentation of evaluation procedures, data, and results is important for transparency and identifying areas for improvement. Their multilingual capabilities originate in these models' pre-training data, which includes hundreds of millions of non-English tokens even in large-scale corpora with a preponderance of English tokens (Blevins and Zettlemoyer, 2022). It has been reported that the unlabeled pre-training data for GPT-3 contains 119 languages (Brown et al., 2020), with around 93% of the tokens being in English1. With 60% and 18% of their pre-training data being non-English, respectively.

The objectives are as follows:

[1] *Corresponding author: Researcher Scholar, Amity University Madhya Pradesh

[2] Professor, The Neotia University, West Bengal

[3] Associate Professor, Amity University Madhya Pradesh

4 Anirban Mitra Associate Professor, Computer Science and Engineering Department Amity University Kolkata, West Bengal

1. This study to examine the zero-sample performance of LLM in an Indian language and test different languages to verify English connections.
2. This study is about the existing open-source LLMs and their performance in text production for Indian languages when measured by the ROUGE criterion. Similar trends are also seen in the results for cross-lingual generation, or generation from Indian languages to English.
3. It is not recommended to use off-the-shelf LLMs in downstream applications in Indian languages. The optimized models can outperform the zero-shot LLMs. For improved performance, fine-tuning with task and language-specific data is crucial.
4. In both mono-lingual and cross-lingual contexts, information produced by LLMs can be more accurate, fluent, and relevant than content produced by humans.

It is important to note that the ROUGE metric evaluation and the manual evaluation of quality metrics yield incongruent results, underscoring the lack of good correlation between automated metrics and human assessments. Overall, rigorous evaluation of generative models on multilingual sentiment ensures fairness, accuracy, and reliability as these models are deployed in real applications. Generative AI has promising capabilities for multilingual sentiment analysis but still requires improvement to handle the complexity and diversity of languages and expressive content. Rigorous benchmarking on realistic data is key to advancing the state-of-the-art.

II. Related Work

Several initiatives aimed at a comprehensive assessment of LLMs' capabilities have been prompted by the increased interest in LLM evaluation. Although Srivastava et al. (2023) covers a wide range of BIG-bench activities, most of the non-English tasks are translation-oriented, which restricts the more general task-based inferences that can be made for such an evaluation. Similarly to, Liang et al. (2022) assessed 30 language models using 42 scenarios and 7 metrics, proposing a taxonomy of scenarios and metrics in the Holistic Evaluation of Language Models (HELM) to define the space of LLM evaluation. All the scenarios, nevertheless, are concentrated on datasets written in either standard or regional English. Lately, there has been a great deal of interest in assessing the many capacities of LLMs, as evidenced by extensive studies such as HELM (Liang et al., 2022) that assess these models on a broad range of capacities. Nevertheless, the majority of these studies use on English language data, and there is a dearth of comprehensive assessments of LLMs' multilingual competencies. The necessity of such an evaluation cannot be overstated, given the present rate at which new language technologies utilizing LLMs are being produced, as instances of disparities in the performance of prior generation models across languages have been well-documented (Blasi et al., 2022).

Targeted efforts have been noted lately to assess these LLMs' performance in the context of various tasks, languages, and modalities. Lai et al. (2023) thoroughly assess ChatGPT's capabilities in several languages for a variety of jobs. Comparably, Bang et al. (2023) focus on reasoning and hallucination while conducting a thorough investigation of ChatGPT in multilingual, multimodal, and multitask settings. Experiments assessing ChatGPT's Text-to-SQL performance are documented by Liu et al. (2023) to investigate the system's capacity to produce structured SQL text from given natural language text. For the binary classification work, 3,120 unique text documents used from 16 varied datasets and for the three-class task, 792 documents from 4 datasets. For the balanced dataset, there is an equal amount of text documents from each class. With the exception of the meta-analytic sample obtained by Hartmann et al. (2023), nineteen of the datasets used in our comparative study are publicly available. To address potential concerns about data contamination, the three LLMs on data produced after their last training session using an Amazon review dataset from 2022 is evaluated. All of the LLMs that were tested were "zero-shot" because they inferred sentiment directly without any explicit task-specific training. The outputs were nearly predictable because the model's temperature was adjusted to zero.

An intriguing use case for generative models is annotation and evaluation. Numerous studies have been documented investigating this idea. Wang et al. (2023b) investigate whether ChatGPT can be used to assess the quality of natural language. Guo et al. (2023) investigate ChatGPT in-depth to see how similar it is to human experts. In a similar vein, Tornberg et al. (2023) find that ChatGPT performs better when it comes to annotation than both experts and crowd workers. In these works, high-resource languages like English and Chinese are taken into consideration. This is investigated in the works of Zhu et al.(2023), Kuzman et al. (2023), and Chen et al. (2023), the use of generative models as opposed to human assessors and annotators. All other Indian languages, excluding Hindi, are classified as low or extremely-low-resource languages (Lai et al., 2023). The data set used to train LLMs contains a smaller

representation of these languages. LLMs do reasonably well in these languages in spite of this. Punjabi (pa) and Odia (or) languages sometimes do remarkably better than Hindi, a language with comparatively more resources. We embody a concept that is considered to be a socio-technological system and a generator of AI as socio-technics. Generative AI research includes a prior class of experiences, which gives us a historical starting point. Attention is directed fundamentally to the language models being used. The use of AR/VR might also be considered as a tool with education (Peres et al., 2023) or in a specific application. The significant positive effect of education on future employment and life quality (Kasneci et al., 2023; Gimpel et al., 2023). The individual whose languages are preserved has better future employment opportunities and a better life quality overall. Additionally, most of the research on sentiment analysis with ChatGPT has been done with datasets in English, with a small amount of work done in the local tongue. Given the speed at which generative models are developing, there is a great deal of interest in comprehending and assessing these models' performance. One way to examine the behavior of these models is to experiment in different scenarios, as many of them (like Bard and ChatGPT) have not revealed all their technical and data specifications. There has been a great deal of interest in assessing the many capacities of LLMs, as evidenced by extensive studies such as HELM (Liang et al., 2022) that assess these models on a broad range of capacities. Nevertheless, the majority of these studies use on English language data, and there is a dearth of comprehensive assessments of LLMs' multilingual competencies. The necessity of such an evaluation cannot be overstated, given the present rate at which new language technologies utilizing LLMs are being produced, as instances of disparities in the performance of prior generation models across languages have been well-documented (Blasi et al., 2022).

III. Empirical Framework and Conceptual Designs

A. Multilingual Generative AI's Mathematical Foundations

Different from discriminative modeling mathematically, generative modeling is the basis of generative artificial intelligence (Ng and Jordan, 2001) and is frequently employed in data-driven decision support. The mathematical underpinnings of generative AI combine ideas from statistics, probability theory, and optimization. We can now mathematically express the chance that x and y will occur jointly thanks to generative modeling. It learns the distribution of different classes and attributes, not the border. With LLMs like ChatGPT and Llama 2 being so easy to get, businesses and researchers can use them for natural language processing jobs like sentiment analysis in ways that have never been possible before. This makes it easier to get an LLM, but it also makes it harder to choose because there are so many of them, and we don't fully understand their pros and cons yet. This wide range of options can make it easy to use a single model for all tasks, without considering how well each model works with different mood analysis tasks. First, it's clear that there isn't a complete empirical framework that tells businesses and academics how to choose the right sentiment classification methods in this age of Generative AI. Second, there aren't any performance comparisons between different LLMs and well-known transfer learning models, such as SiEBERT, on a standard set of data. This makes study harder. Lastly, more research needs to be done on how the types of data and the way they are analysed affect how well LLMs can classify things. The framework guides the planning of the next experiments, which are made to fill in the gaps in current study on sentiment analysis.

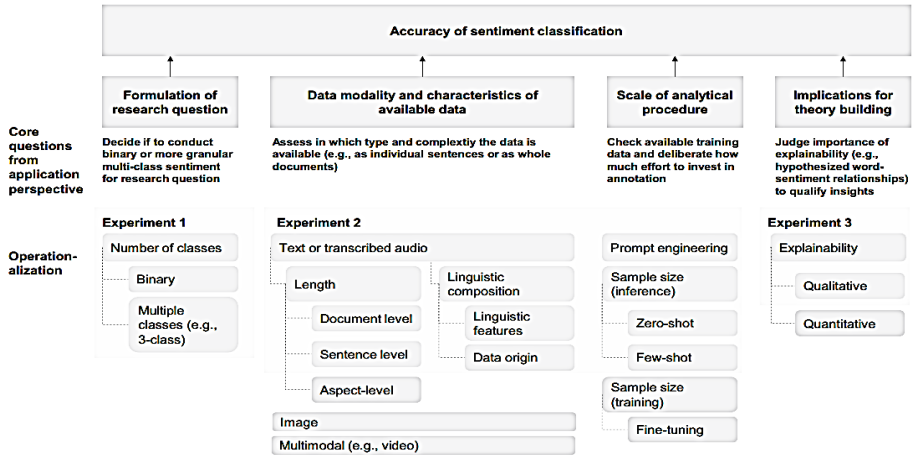

Fig. 1: Empirical framework for sentiment analysis in the age of Generative AI, adapted from Hartmann et al. (2023)

The visualization of models shows that, in addition to all the traits of both species' tails and ears, a generative model is also capable of predicting other properties of a class. This implies that it learns characteristics and how they connect to get a general idea of how those species seem. In Figure 1, we can see the revised framework that incorporates the Generative AI component, namely LLMs, within the preexisting sentiment analysis approach benchmark proposed by Hartmann et al. (2019 & 2023). The decision between simpler binary classification tasks and more complex multi-class classifications is shaped by the research topic and context, which in turn affects the accuracy of the classification. Consequently, these considerations are crucial when selecting a sentiment classification method. Based on their research aims, researchers need to decide how many classes sentiment analysis will use. From simple binary tasks, such predicting whether social media posts would be positive or negative or detecting firestorms, to more complicated multi-class tasks, like distinguishing emotions from email headlines, this decision-making can take many forms.

B. Models for Generative AI

These are the two most widely used generative AI models in our study. Generative Adversarial Networks (GANs) are the discriminator and generator networks are beyond reference. Supervised machine learning is used to model a customer against an existing base of other customers. The generative adversarial network is an amazing work of development in AI. A group of researchers, Goodfellow et.al. (2014) led by Jan Goodfellow at the University invented the first AI that produced original pieces of art. In Generative Adversarial Networks, GAN architecture shows a great deal of study of the real world applications have made GANs the most widely used of generative AI model. A group of deep learning language models known as GPT-3 were developed by the OpenAI group, an artificial intelligence research center situated in San Francisco. GPT-3 is the name of the generative pre-trained transformer model. GAN architectures consist of two sub-models in which the discriminator is a neural network that ascertains if a given sample originates from the domain or is a fictitious sample from a generator. A neural network is a generator that takes a random input vector—a set of mathematical variables—and creates fictitious input or fake samples out of it. The discriminator is essentially a binary classifier that produces probabilities, which are differentiable integers between 0 and 1. The closer the value approaches zero, the higher the likelihood of fake production. On the other hand, values that are closer to 1 suggest a higher likelihood of the forecast being accurate. CNNs (Convolutional Neural Networks) are frequently used to build both a generator and a discriminator, particularly when dealing with images. When a generator produces a fictitious sample that is sufficiently realistic that it can trick both people and discriminators, GANs will be successful. However, the game continues after that since it's time for the discriminator should be improved and updated. Transformer-based models are computer programs, such as Generative Pre-Trained (GPT) language models, that use data gathered from the Internet to produce textual material, such as press releases and whitepapers.

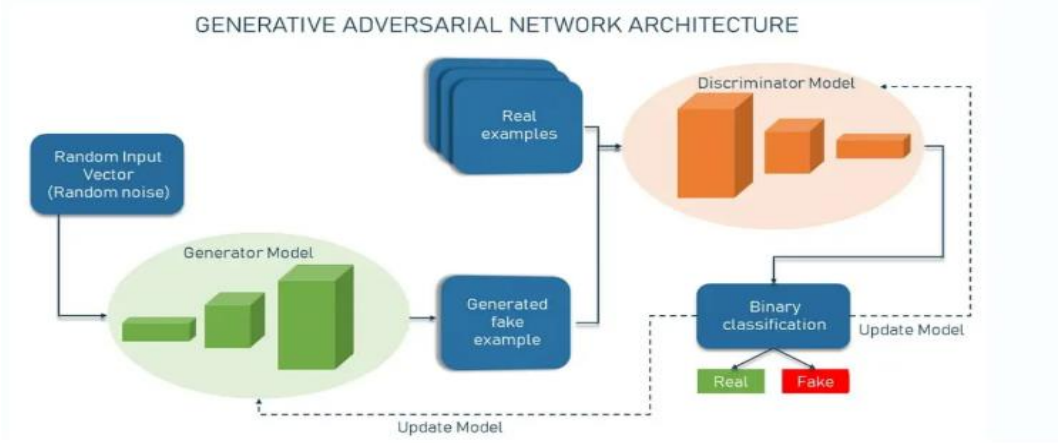

Fig. 2: The framework for generative adversarial network architecture

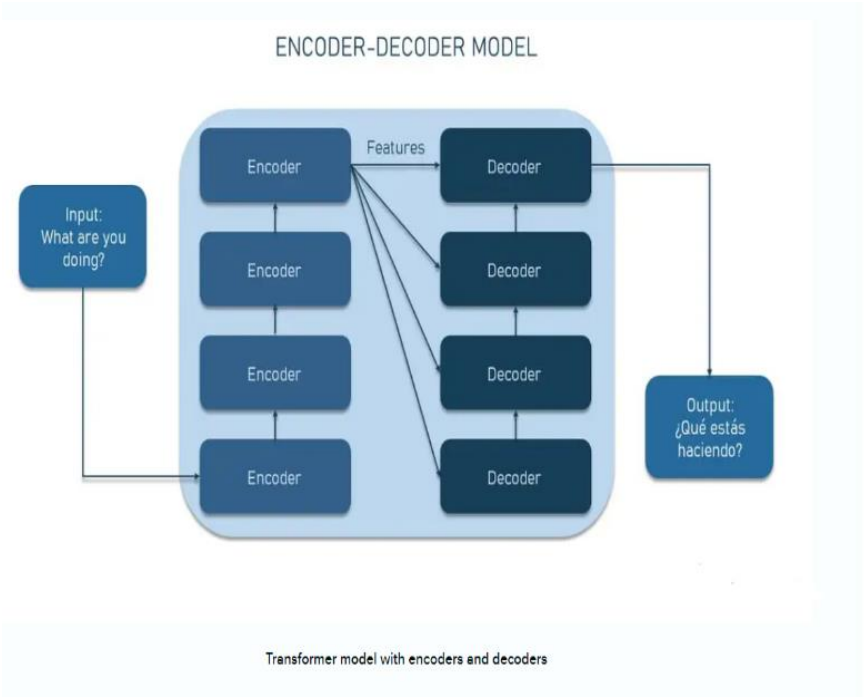

Fig. 3: The encoder-decoder transformer model

Transformers are strong deep neural networks, initially introduced in 2017 by Google research, that follow connections in sequential data, like the words in this phrase, to comprehend context and meaning. This explains why this technology is commonly used in NLP (Natural Language Processing) activities. GPT-3 and LaMDA are two of the most well-known transformer examples. The model can generate content that appears to have been produced by a human being. Discrimination and artificial neural network competition using a machine learning technique called generative adversarial network (GAN). In a zero-sum game involving two neural networks, the gains of each agent are equal to the losses of the other. Typically, a transformer has two sections. The input sequence is processed by the encoder. It takes a sequence and extracts all its features, turns them into vectors (like the semantics and sentence structure of a word), and sends them to the decoder. Working with the intended output sequence is the decoder. After the encoder receives its output, each decoder interprets the content and creates an output. The encoder and decoder of the transformer are made up of several encoder blocks stacked on top of one another. The output of one block becomes the input of another. The multilingual generative AI is a kind of machine intelligence that does multilingual and generative tasks. In learning architecture where the instances of the data are recursively created to form new ones. By utilizing multilingual AI algorithms, as well as translations of diplomatic documents, and patterns uncovered through AI; the information within the training data that indicates a correlation Gen artificial intelligence holds out great prospects; however, it would be premature to get rid of the human being. Deep neural networks have an edge for applications requiring the ability to perceive different levels of abstraction at the same time. For evaluating the effectiveness of data production, they can be routines created using data sources. Multiple structures to emulate unique statistical attributes for data types like sequential like human language or spatial-like encoding pictures (Janiesch et al., 2021; Kraus, et al., 2020).

A summary of major ideas and architectural schemas will be discussed widely used in the style of generative AI which might encompass anything from robotics to computer learning and natural language processing. Large language generative (LLMs) models for consistent text generation, using a transformer. For instance, the well-which consists of a family of LLMs identified as GPT which is abbreviated GPT-short for generative. The pre-trained transformer serves to be the foundation on which the text is formulated and casualties such as conversational bots. Large-scale generative AI models capable of precisely and completely detailing the human mental process.

In the Multi-lingual View of the systems, several components can be grouped or classified are connected and exchange information. For multi-lingual there are some key considerations when it comes to using generative AI systems. Among these are the underlying generative models that capture humans' ability to create new information. previously described. The translation app with another name of Human Generating AI model, besides the users who are familiar with the same language background from dealing with elements, modality, and which info is backing

their argument/ the elements, their modality, and the supporting data processing (for example, prompts). An illustration would be the adoption of the latest research findings, for example, Codex, as in deep learning models (Chen et el. 2021). GitHub Copilot demonstrates how simple it has become to create new codes despite the ease of the task. An ecosystem with intuitive design and rich features that will provide more useful and transparent information. For example, if there is a bot on the server, users can interact the with bot because the bot usually automates tasks to be able to make graphics using Ai-based image generation.

In the Application-Level View, the use cases of generative AI are the applications of generative AI systems which can be used for data management and processing, natural language processing, machine translation, and content generation. This solution of the network can give the impression of being an information system and such tasks as production processes controlling and human-machine systems whose main goal is making labor-saving production. For editors, this level of generative intelligence is boundless by the limitations that we have currently set. The optimization area can either be excluded or incorporated in the machine learning (Reisenbichler, et al. 2022), AI-based music (Garcia 2023) and deep fakes, (Metz) cover image generation are our current age's breakthrough innovations. AI has seen progress in the areas of facial and speech recognition, image captioning and automatic language translation (Chen et al. 2021).

IV. Experimental Results and Analysis

The focus of these studies is to evaluate generative models' abilities for data production and augmentation in multilingual sentiment analysis (we focus on Hindi). We compare three models, GPT-3.5, a Generative Pre-trained Transformer, and GPT-4 provided through OpenAI's ChatGPT; Bard AI by Google, with its PaLM 2 enabled. During this document, the terms "Bard" and "PaLM" will be used correspondently. For GPT-3.5 and GPT-4 we send prompts and get responses using ChatGPT API. For Pam 2, Bard AI accepts manual cues via an online interface from which respective responses are extracted. These models are also employed to generate new examples that have been systematically analyzed.

There are four primary experiments we conducted. In the Experiment 1, we have introduced different BERT-based Hindi language models using existing data and evaluate their performance. We used a variety of pre-trained models, including those trained on Twitter or social media profiles, from various Hindi corpora. The basic model will be the most effective model. In Experiment 2, GPT-3.5, GPT-4, and other generative models (PaLM 2) are evaluated by dividing test scores into "none," "poor," and "good." to evaluate the performance of each model based on test data. This attempt seeks to address RQs 1 and 2. In the Experiment 3, the aim was to ask the generative models to come up with m tweets on different attitudes ("positive", "negative", or "neutral"). The generated tweet samples will be subjected to manual assessment for their naturalness across several parameters. The objective is aimed at addressing RQ3. In the Experiment 4, two intended applications used for this data, which was generated in Exp. 3; first, it would be employed to refine the first training set; second, it will be refined through BERT-based models that were applied in Experiment One. Lastly, these same BERT-based models from experiment one are fine-tuned using pooled data sets acquired during this study. Test data will be used to evaluate how well both approaches perform. This experiment aims to answer RQ3. We employed the Hindi Twitter Dataset (HTD), which includes about 11,000 Hindi-English code-mixed tweets labeled for sentiment analysis, for the studies mentioned above. Hindi-English Tweets used Mixed Script. Approximately 93k tweets in Hindi, frequently with mixed codes, written in both Devanagari and Roman scripts.

For the Sentiment analysis dataset in Hindi and English, 15k Tweets with mixed Hindi-English code and sentiment annotations used. Our research focused on the first three types of tweets with tag of good, negative, and neutral, which accounted for 558, 1,632, and 500 tweets. During the research work, we used 2,690 tweets which randomly allocated HTD dataset into 25% for running tests on HTDtest and 75% for training purposes following class distribution. Additionally, we looked at the results of each suggested question by using thirty tweets from each class—that is three from HTDtrain which gives a total of ninety. We refer to these as ninety-nine tweets as HTDdev. Exp. 1 and Exp. 4 refine language models by use of HTDtrain. The predictions generated by generative models in Exp 2 are assessed using the set of data found in HTDtest and refined language models (Exp1 and Exp4). Human judgement is used in Experiment 3 and generative models' outputs are used to refine language models in Experiment 4. Using HTDdev, we evaluate different prompts for Exps.2 and Exps.3. However, due to financial constraints and time limitations, we made it balanced in classrooms since it is relatively small-sized as compared to other prompts some issues over varying shot counts could not be assessed because not all prompts were evaluated when there were many more shots involving students. However, on the entire test set HTDtest , optimal configurations for

number of shots and prompts will be evaluated. These prompts are based on or inspired by earlier studies that have been published (Alyafeai et al.,2023; Khondaker et al.,2023).

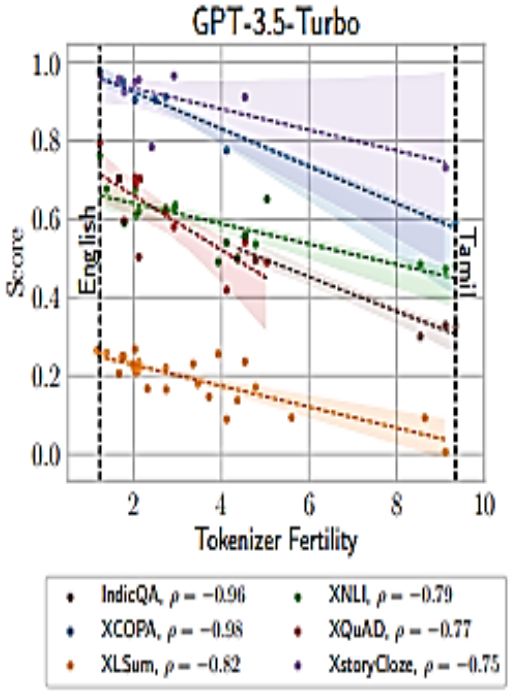

Fig. 4: The Turbo GPT model

In this section we will show and discuss the results of the above tests. We use the accuracy (Acc) metric and the micro-average of recall (R), precision (P), and F-1 score (F) values to evaluate the performance of the model. The main comparison metric we use to measure model performance is the F1 metric. For Experiment 1, Fine-tuning BERT Models (baseline) used. In these models all their pre-training done on Twitter or other social media data. All models were the subject of numerous trials with different hyperparameter settings. It should be emphasized, though, that each of these three datasets was included during the multitask fine-tuning phase for BLOOMZ, particularly for XQUAD and TyDiQA-GoldP, for which the evaluation-use validation data is probably also included in the fine-tuning data. The results show that the best performance is achieved by pre-training the model using the same data (Twitter data in this case) as the quality data only. In Experiment 2, the Analysis of Sentiment tested using Generative Models. In these tests, we tested Bard AI on HDDest along with ChatGPT (GPT 3.5 and GPT 4). Finding the best sign design for a generative model requires a step back compared to known control configurations. Hyperparameters for pre-trained languages used where at HTDtest, we conducted tests using seven designs in Hindi and English to classify tweets. We evaluate each clue pattern in HTDev and then use k shots (where k = {0, 1, 3, 5}) to select the most accurate clue. Each shot contains positive, negative, and neutral tweets. For K = 5, Bard AI's best recommendation is a total accuracy of 0.7 tweets. For example, have a look at these instances: a positive train tweet; a sentiment that is 1 (positive) neutral; a sentiment that is 0 (neutral) negative; a sentiment that is -1 (negative); a sentiment that is 1 (positive) neutral; a sentiment that is 0 (neutral) negative; a sentiment that is -1 (negative). Assisting others by predicting whether a Hindi tweet is Positive, Negative, or Neutral, you are a helpful assistance. Don't provide a disclaimer, explanation, or caution. Kindly reply to the test tweet with a tabular response that includes the tweet and its classification. The best prompt obtained accuracy scores of 0.81 and 0.91 for GPT-3.5 and GPT-4, accordingly, with k = 0. The message of X/Twitter provided a neutral, negative, or positive response.

Table 1: Comparison of Models used in Experiments

Model	**Classification**				**Question Answering**			**Sequence Labelling**		**Summarization**
	XNLI	**PAWS-X**	**XCOPA**	**XStory Cloze**	**XQuAD**	**TyDiQA-GoldP**	**MLQA**	**UDPOS**	**PAN-X**	**XLSum**
Metrics	**Accuracy**				**F1/EM**			**F1**		**ROUGE-L**
Fine-tuned Baselines										
mBERT	65.4	81.9	56.1	X	64.5 / 49.4	59.7 / 43.9	61.4 / 44.2	71.9	62.2	X

mT5-Base	75.4	86.4	49.9	X	67.0 / 49.0	57.2 / 41.2	64.6 / 45.0	-	55.7	28.1
XLM-R Large	79.2	86.4	69.2	X	76.6 / 60.8	65.1 / 45.0	71.6 / 53.2	76.2	65.2	X
TuLRv6-XXL	88.8	93.2	82.2	X	86 / 72.9	84.6 / 73.8	81 / 63.9	83.0	84.7	X
Prompt-based Baselines										
BLOOMZ	54.2	82.2	60.4	76.2	70.7 / 58.8	75.2 / 63.2	-	-	-	-
Open AI Models										
Text-davinci-003	59.27	67.08	75.2	74.7	40.5 / 28.0	49.7 / 38.3	44.0 / 28.8	-	-	-
Text-davinci-003 (TT)	67.0	68.5	83.8	94.8	x	x	54.9 / 34.6	x	x	-
gpt-3.5-turbo	62.1	70.0	79.1	87.7	60.4 / 38.2	60.1 / 38.4	56.1 / 32.8	60.2	40.3	18.8
gpt-3.5-turbo (TT)	64.3	67.2	81.9	93.8	x	x	46.3 / 27.0	X	X	16.0
gpt-4-32k	75.4	73.0	89.7	96.5	68.3 / 46.6	71.5 / 50.9	67.2 / 43.3	66.6	55.5	19.7

After determining which prompt works best, we test each of the three generative models on HTDest by asking them to classify each scenario as good, negative, or neutral. We decided that if a machine rejects a tweet due to its unsuitable content, it is predicted to have a negative outcome. By comparing the output of our generative model with the labels of the test data, we calculate the performance metrics. The three generative models' performance metrics are displayed in Table 2.

Table 2: Confusion matrix for models

Data Models	Accuracy	Precision	Recall	F1-score
GPT-4	74.97	81.39	75.12	77.32
Bard AI	79.78	77.89	79.22	77.45
GPT-3.5	69.87	71.93	70.16	70.45
Best BERT	79.84	79.67	79.87	79.85

The findings demonstrate that, when using the wholly supervised BERT-based models in a few shot conditions, GPT-4 and Bard AI perform similarly, with the F-1 score serving as the benchmark performance metric. Specifically, with an F-1 score of 0.77, GPT-4 performs better than the other fine-tuned models and is extremely close to the second-best BERT model. When compared to fully supervised models, Bard AI does reasonably well for sentiment analysis categorization, coming in second place with a score of 0.76. It is noteworthy that, with an F1 score of 0.76, it obtains performance comparable to the refined MARBERTv2 model, outperforming one of the BERT-based models. But GPT-3.5 performs poorly, lagging models based on BERT.The distinct class distributions are the reason for the notable discrepancy in the models' performance on the test set HTDtest and the development set HTDdev. Specifically, Bard AI does very poorly in the neutral class, which accounts for 33% of tweets in HTDdev. While GPT-4 and the refined BERT model perform similarly in a zero-shot environment, each model's performance differs significantly. Table 3 shows the F1 score for each sentiment class for both models. The results show that classifying neutral tweets was the most difficult assignment, with both models (BERT and GPT-4) doing better for classifying negative tweets, then positive tweets. When it came to the classification of unfavorable tweets, the optimally configured BERT model outperformed GPT-4. While GPT-4 and the refined BERT model perform similarly in a zero-shot environment, each model's performance differs significantly. Table 3 shows the F1 score for each sentiment class for both models. The results show that classifying neutral tweets was the most difficult assignment, with both models (BERT and GPT-4) doing better when classifying negative tweets, then positive tweets. In terms of classifying negative tweets, the best-configured BERT model fared better than GPT-4. Positive tweets performed much better than neutral ones in this regard, with a 4-point improvement in the F1 score for both.

Table 3: Comparison of sentiments for models

Data Models	Class		
	Negative	Positive	Neutral
GPT-4	84.7	77.9	57.2
Bard AI	82.4	75.6	57.5
GPT-3.5	79.8	76	58
Best BERT Model	89.4	76.6	59.8

In the Experiment 3, Data Generation by Generative Models used for the study. In a zero-shot configuration, we gave each generative model instructions to produce “positive”, “negative”, and “neutral” tweets. In both Hindi and English, we ran tests with eleven distinct prompt designs, and then, after evaluating the final output according to three criteria, we determined which prompt was the best:

A. The tweets' organic tone.

B. The Hindi dialect is used in tweets.

C. Refraining from sending too brief tweets.

We generated multiple results for each recommendation, evaluate them on a small scale (running each strategy several times), and select the best recommendations to offer the most excellent prompt above. Each generative mannequin (i.e., GPT-3.5, GPT-4, and Bard AI)was in contrast to its corresponding distribution in the training dataset HTDtrain, which is 391 for positive, , and 351 for negative, and so on, to assemble tweets for every category (i.e., [“positive”, “negative”, “neutral”]).To determine the calibre of the generated tweets and the emotions related to them, we randomly pick 50 tweets from every classification for every producing mannequin (a whole of 150 tweets per model). After that, for each tweet that used to be generated, two annotators dealt with the following binary questions (Yes/No)

Question 1: (Making Sense): Does the ensuing tweet have proper grammar and make sense?

Question 2: (Fit for Twitter): Would you count on seeing this content material on Twitter?

Question 3 :(Matching label): Does the ensuing tweet reflect the sentiment that was once specified?

Table 4: Comparison of models for confusion matrix

Data Models	Accuracy	Precision	Recall	F1-score
SDTC	78.97	78.39	78.12	78.32
Bard AI	69.78	70.89	69.22	67.45
GPT-3.5	60.87	63.93	60.16	55.45
GPT-4	69.84	74.67	69.87	67.85
Bard AI+SDTC	79.5	79.2	78.9	77.9
GPT-3.5+SDTC	77.6	77.5	77.2	76.3
GPT-4+SDTC	79.8	79.6	79.2	79.2
All Data Models	76.8	77.9	76.8	76.5

In the Table 4, when each annotator agrees on a "Yes" response, the generated tweet is deemed legitimate for that question; if not, it is deemed invalid. The proportion of sure solutions for each question and classification for every one of the three generative fashions is shown in Table 4. The findings exhibit that Bard AI outperforms GPT-3.5 and GPT -4 by way of a small margin, relying on whether all assessment questions are taken into account (46%). For the equal Q1, Q2, and Q3, percentages bought are95%, 62%, and 81%. When it comes to linguistic accuracy(Q1) and sentiment matching (Q3) for each “positive” and “negative” tweet, all fashions operate admirably. On the other hand, producing tweets that are appropriate for Twitter sure difficult (Q2). In the analysis of the generated impartial tweets when we evaluated the labels supplied with the aid of annotators with the impartial tweets produced by way of generative models, we can examine how stressed the tweets are with different classes covered in Exp. 1. For Bard AI, annotators recognized 96% of the tweets as positive. 13% of GPT-4 responses are categorized as terrible and 87% as positive, Comparatively,68% of GPT-3.5 responses had been labeled as “positive” and 32% as “negative”. “positive” content material seems hard for the generative models to distinguish from different kinds of information, particularly impartial material. This is in addition validated in experiment 1, the place where we adjusted BERT models, since neutral tweets are the toughest for BERT-based models to interpret accurately, main to inaccurate fantastic or negative classification.

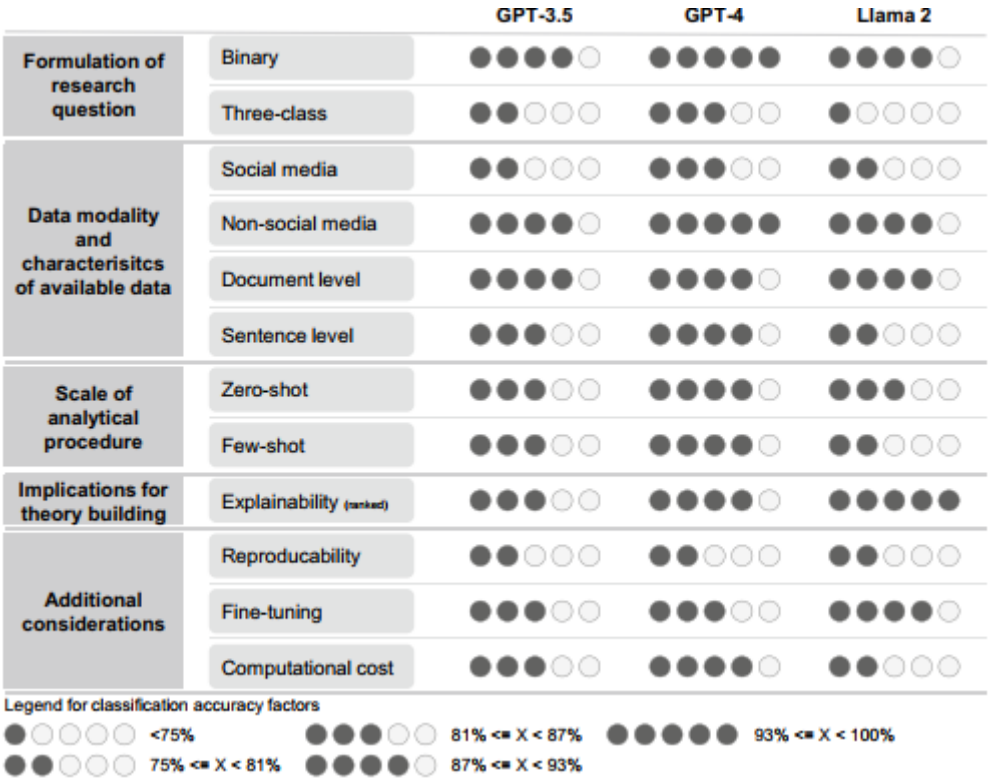

Fig. 5: Evaluation matrix for method selection summarizing the consolidated results

In Experiment 4 for optimizing the Ideal BERT Model with Generated Data, the same hyperparameters and the generated data together with HTDtrain were used with the seven iterative investigations for the best-performing BERT model. As with the other trials, the F1 measure served as the main comparative parameter and was the focus of our attention. According to the statistics, the model optimized using Bard AI data exhibited optimal performance throughout testing for every created dataset with a 0.67 F1 score. The mannequin with an F1 rating of 0.66 used to optimize the usage of information produced via GPT-4 got here in a shut second. The human assessment of the generated information indicates a high-quality correlation with the overall performance records. The cause for the models truly decreased overall performance is that the trying out statistics got here from an extraordinary population, and the generative models carried out a terrible job of classifying the generated data. The overall performance of the sophisticated mannequin remained unchanged when the generated statistics from every mannequin used to be blended with the authentic education set. It might also even have performed a phase in the model's declining efficiency. Performance suffers when the created statistics are mixed with unique data. The above-listed motives additionally account for this drop in performance. By using generative fashions to create new samples from one or more pictures of coaching data, overall performance can be better and an increased diploma of resemblance between the generated facts and the unique facts distribution can be guaranteed. By taking into consideration the possibility of data contamination and conducting a thorough examination of potential confounding variables including dataset origin, language features, and analytical process, our experimental design is based on a big and varied data sample. With the proliferation of state-of-the-art LLMs, the convention of developing or fine-tuning models for sentiment analysis on specific and proprietary datasets may become less relevant, according to marketing practitioners, who should be aware of the remarkable zero-shot sentiment classification performance achieved by all three tested LLMs. As a helpful reference for choosing a sentiment analysis approach, Figure 5 summarises our main findings. There are three important considerations to keep in mind while using LLMs for sentiment analysis: (1) Performance degrades as the number of classes increases; (2) The accuracy of classification is greatly affected by data features and analytical procedures; (3) Factors including computational costs, fine-tuning choices, and reproducibility also play a role in method selection.

V. Conclusion and Future Work

This research looks at LLMs' ability to speak more than one language by testing them with different models, activities, languages, and methods of prompting. To further understand the performance patterns, we examine essential attributes such as tokenizer quality and pretraining data volume. Languages written in the Latin script and those with fewer resources consistently fare worse, according to our research. At the same time as it recognizes the limitations of tactics like translate-test prompting, it stresses their efficacy. Our evaluation provides evidence that human review and automatic benchmarking should be prioritized in as many languages as possible. Finally, expanding on our study's focus on text as a data modality for sentiment analysis, future research should investigate the effectiveness of Generative AI in other data modalities. We hope that by taking this step, it will encourage additional study towards this goal. Unlike other proprietary models like PaLM, which includes training data in multiple languages, we were only able to compare the evaluation outcomes of GPT-3.5 and GPT-4 with BLOOMZ and state-of-the-art (SOTA) models. We did not evaluate all the available multilingual datasets, which is a limitation of our work. But, with the help of the research community, we plan to broaden our review in subsequent study editions. We evaluate all available multilingual datasets, but they don't cover all the languages that are under-resourced or have different typologies. For the time being, this is a major roadblock to expanding multilingual evaluation. The existing grading standards do not adequately account for African languages or Indigenous languages spoken in the Americas. Building a model to evaluate current LLMs for sentiment analysis was our main focus. In zero-shot and few-shot settings, it is critical to investigate alternatives to LLMs' restricted capabilities. Thus, it seems likely that LLMs have reached their bare minimal potential in sentiment analysis, according to our study's results. While our study does shed light on how LLMs perform differently when it comes to sentiment analysis (both at the document and phrase levels), more research into aspect-based sentiment analysis is required. When it comes to aspect-based sentiment analysis, it would be beneficial for future research to examine the biases and evaluate the capabilities of state-of-the-art Language Models (LLMs). To replicate our experiment's conditions, this comparison should be carried out in a few-shot and zero-shot scenario. It follows that the model can be fine-tuned or given specialist training to produce even more accurate outcomes. Particularly in domains requiring specialist knowledge, like financial or medical advice, it is crucial to evaluate how the amount of the training data affects the performance of LLMs in sentiment analysis.

References

[1] Ackermann, Noëmi Aepli, Hamid Aghaei, and Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. American Chapter of the Association for Computational Linguistics: Human Language Technologies, Associates, Inc.

[2] Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Proceedings of EMNLP 2019, pages 3685–3690.

[3] Baldridge. 2019b. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language

[4] Bang, Yejin & Cahyawijaya, Samuel & Lee, Nayeon & Dai, Wenliang & Su, Dan & Wilie, Bryan & Lovenia, Holy & Ji, Ziwei & Yu, Tiehzheng & Chung, Willy & Do, Quyet & Yan, Xu & Fung, Pascale. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. 10.48550/arXiv.2302.04023.

[5] Berger J, Humphreys A, Ludwig S et al (2020) Uniting the Tribes: Using Text for Marketing Insight. J Mark 84(1):1–25. https://doi.org/10.1177/0022242919873106

[6] Berger J, Milkman KL (2012) What Makes Online Content Viral? J Mark Res 49(2):192–205. https://doi.org/10.1509/jmr.10.0353

[7] Berger J, Packard G, Boghrati R et al (2022) Marketing insights from text analysis. Mark Lett 33(3):365–377. https://doi.org/10.1007/s11002-022-09635-6

[8] Berger J, Sherman G, Ungar L (2020) TextAnalyzer. http://textanalyzer.org/about. Accessed 15 Jan 2024

[9] Bommasani R, Hudson DA, Liang P and others (2021) On the opportunities and risks of foundation models. arXiv:2108.07258https://doi.org/10.48550/arXiv.2108.07258

[10] Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Boyd RL, Ashokkumar A, Seraj S et al (2022) The development and psychometric properties of LIWC-22. University of Texas at Austin, Austin, TX, pp 1–47. https://www.liwc.app/static/docum

[11] Burger B, Kanbach DK, Kraus S, Breier M, Corvello V (2023) On theuse of AI-based tools like ChatGPT to support managementresearch. Europ J Innov Manag 26(7):233–241. https://doi.org/10.1108/EJIM-02-2023-0156

[12] Burger, B., Kanbach, D. K., Kraus, S., Breier, M., & Corvello, V. (2023). On the use of AI-based tools like ChatGPT to support management research. European Journal of Innovation Management, 26(7), 233–241. https://doi.org/10.1108/EJIM-02-2023-0156

[13] C. Gong, Y. Shen et al., "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," arXiv preprint arXiv:2303.10420,2023.

[14] Chen M, Tworek J, Jun H, Yuan Q, Pinto HPdO, Kaplan J, EdwardsH, Burda Y, Joseph N, Brockman G, et al (2021) Evaluating large language models trained on code. arXiv:2107.03374

[15] Chiang T (2023) ChatGPT is a blurry JPEG of the web. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurryjpeg-of-the-web, accessed 25 Aug 2023

[16] Chouhan A, Halgekar A, Rao A et al (2021) Sentiment Analysis of Twitch.tv Livestream Messages using Machine Learning Methods.In: 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, pp 1–5

[17] Chui M, Yee L, Hall B, Singla A, Sukharevsky A (2023) The state of AI in 2023: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year. Accessed 17 Aug 2023

[18] Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2023). ChatGPT and its impact on research supervision: Insights from Australian postgraduate research students. Australasian Journal of Educational Technology, 39(4), 74–88. https://doi.org/10.14742/ajet.8843

[19] Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Demszky D, Movshovitz-Attias D, Ko J et al. (2020) GoEmotions:A Dataset of Fine-Grained Emotions. arXiv. https://doi.org/10.48550/arXiv.2005.00547

[20] Ding N, Qin Y, Yang G et al (2023) Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell5(3):220–235. https://doi.org/10.1038/s42256-023-00626-4

[21] Dwivedi YK, Kshetri N, Hughes L et al (2023) Opinion Paper:“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage

[22] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.10264

[23] Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., Röglinger, M., Ruiner, C.Schoch, M., Schoop, M., Urbach, N., Vandirk, S. (2023). Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education: A Guide for Students and Lecturers. University of Hohenheim, March 20, 2023.

[24] Giray L (2023) Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng 51(12):2629–2633. https://doi.org/10.1007/s10439-023-03272-4

[25] Liang, Sewon, Jordan Hoffmann, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. "Thinking fast and slow with deep learning and tree search." Advances in Neural Information Processing Systems 35 (2022).

[26] Metz C (2023) Instant videos could represent the next leap in AI.technology. https://www.nytimes.com/2023/04/04/technology/runway-ai-videos.html, accessed 25 Aug 2023

[27] Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems 14 (2001). NLP, pages 120–130, Online. Association for Computational Linguistics of Machine Learning Research, pages 12697–12706.

[28] Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Reisenbichler M, Reutterer T, Schweidel DA, Dan D (2022) Frontiers: supporting content marketing with natural language generation. Market Sci 41(3):441–452

[29] Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

[30] Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. Information Systems Research, 34(2), 399–408.

[31] Suzgun, Mirac & Scales, Nathan & Schärli, Nathanael & Gehrmann, Sebastian & Tay, Yi & Chung, Hyung & Chowdhery, Aakanksha & Le, Quoc & Chi, Ed & Zhou, Denny & Wei, Jason. (2022). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them.

[32] Syafri Bahar, et al. 2020. Indonlu: Benchmark and T. Kuzman, N. Ljubesi˘ c, and I. Mozeti ´ c, “Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification,” arXiv preprint arXiv:2303.03953, 2023.

[33] Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of ChatGPT et al.: The prospects of large language models. Business & Information Systems Engineering, 65, 95–101. https://doi.org/10.1007/s12599-023-00795-x

[34] Tom Warren. 2023. Microsoft’s chatgpt event live blog, accessed at 9th April, 2024.

[35] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang et al., “A survey on evaluation of large language models,” arXiv preprint arXiv:2307.03109, 2023.