Article

Iterative Bias Evaluation Approach for Sentiment Aware Text Summarization

Phuoc Nguyen [1,†,‡], Atqiya Munawara Mahi [2,‡]*Chris Argenta [2,‡]*

[1] Affiliation 1; e-mail@e-mail.com
[2] atqiyamunawaramahi@student.uml.edu
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

Abstract: Bias is a pervasive challenge in intelligence summaries, whether they are generated by humans or Large Language Models (LLMs). LLM-based abstractive summarizers can inadvertently amplify existing biases when summarizing, making it even more important for intelligence analysts to identify and mitigate bias in their reporting. One way to detect bias in summaries is to ensure that the summaries accurately reflect the content of the original articles. Aligning the summaries with their source articles and using tools like Textblob and VADER can help analysts identify and correct potential biases. More effective bias analysis can help intelligence analysts produce more objective summaries. This research explores methods that intelligence analysts can use to curb bias in summaries, whether created by humans or AI. We present examples of the risk of LLM-based automatic abstractive summarizers inadvertently magnifying bias. We demonstrate detecting bias by comparing article summaries and their respective articles for coherence and using TextBlob and VADER to evaluate sentiment differences. We summarize some best practices for bias analysis to assist intelligence analysts in generating more balanced summaries.

Keywords: bias, intelligence summaries, LLMs, Textblob, VADER.

2. Introduction

The quality and accuracy of intelligence summaries have become increasingly critical for informed decision-making across many sectors. Summaries are frequently the foundation for decision-making in policy, business, and security. The summarization process creates a significant opportunity for the introduction of biases that potentially change the intended interpretation of the original content. Since a summery is often used as a replacement for reading the original content, these biases may go undetected and misattributed. The challenge of detecting and avoiding such bias, both conscious and unconscious, poses a significant hurdle in achieving accurate and objective intelligence summeriesoo.

Biases of this sort can distort the essence of information, leading to skewed perceptions and potentially flawed decisions. This is particularly concerning when the source of bias is not readily apparent or is deeply ingrained within an information processing system. The popularity of generative AI technologies has recently increased for task like automatic abstractive summarization using large language models (LLMs). While LLMs have revolutionized the field of text summarization by generating concise and coherent summaries, we have observed that they can amplify existing biases based on the data they were trained, leading to a concerning propagation of these biases in their outputs and the skewing of reports based on these biases..

This paper suggests a strategy for detecting and mitigating such bias in intelligence summaries to ensure their objectivity and validity. Specifically, we present the complexities of bias in LLM-generated summaries, exploring how these models embed and/or amplify

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

biases subtly embedded and inadvertently. It also highlights the potential consequences of unchecked biases, emphasizing practical methods for detecting and mitigating bias. These include the comparison of summaries with their original articles to identify consistency and potential skew, and sentiment analysis tools like Textblob and VADER to detect subtle differences in potential interpretations. These tools can help in quantifying the sentiment and subjectivity in text, providing a more objective measure to assess bias.

Furthermore, the paper provides insights into how intelligence analysts can curb the propensity for bias. It explores strategies for both human-generated summaries and those created by LLMs, acknowledging that each has unique challenges and requires tailored approaches. By shedding light on these issues and offering practical solutions, this paper aims to contribute to the ongoing discourse on bias in intelligence summaries and the broader field of artificial intelligence. (add bullet points)

3. Background and Literature Review

Large language models have experienced remarkable progress and popularity since their inception. From early models that introduced pre-training and fine-tuning concepts to the recent breakthroughs in transformer architectures, these models have transformed the landscape of natural language understanding and generation. The field of large language models gained traction with models like ELMo[1] and ULMFiT[2]. These early models introduced the concept of pre-training on a vast corpus of text, followed by fine-tuning for specific tasks. While they laid the foundation for subsequent advancements, these models were limited in size and performance. (will add the referances later)

Introduced by Google in 2018, BERT[3] revolutionized natural language processing by introducing a bidirectional training approach. Unlike previous models that relied on left-to-right or right-to-left context, BERT considered both directions simultaneously during training.

OpenAI's GPT series is the currently leading approach in the field of large language models. The original GPT model, released in 2018, leveraged transformer architectures to achieve significant improvements in language understanding and generation. By training on massive amounts of text data, GPT models exhibited the ability to generate coherent and contextually relevant text. In 2019, OpenAI unveiled GPT-2[4], a groundbreaking model that pushed the boundaries of size and performance. With an astonishing 1.5 billion parameters, GPT-2 showcased unprecedented text generation capabilities. The release of GPT-3[5] in mid-2020 marked a milestone in the development of large language models. Boasting a staggering 175 billion parameters, GPT-3 became the largest language model at the time. Its language understanding and generation abilities were exceptional, often producing impressively coherent and contextually relevant responses.

Since GPT-3, the field of large language models has witnessed notable advancements. Researchers have explored techniques such as scaling laws, model distillation, and improved training strategies to further enhance model performance. Focus has also been placed on addressing limitations, such as biases in generated text and improving sample quality.

Traditional evaluation metrics exist for summarization assess the quality and effectiveness of automatic text summarization systems. These metrics help compare machine-generated summaries to human-written references and provide a quantitative measure of their performance. Among them, ROUGE[6] is a popular set of metrics used to evaluate the quality of a summary. It measures the overlap of n-grams (sequences of n words) between the generated summary and the references. Originally designed for machine translation evaluation, BLEU[7] has also been adapted for summarization evaluation. It calculates the n-gram precision between the generated summary and the reference summaries. BLEU is widely used but is less effective for evaluating short summaries. Precision and recall are commonly used evaluate summarization results. Precision is measured as accuracy of the generated summary by calculating the ratio of the correctly included information to the total information in the summary. Recall measures the completeness of the summary by calculating the ratio of the correctly included information to the total information in the reference summary. The F1 score another metric which is the harmonic mean of precision and recall. It provides a balanced measure of both precision and recall and is often used when there is an uneven class distribution between summaries and references. Originally developed for language modeling, perplexity is used as an evaluation metric for abstractive summarization models. It measures how well the model predicts the reference summaries given the source text. Lower perplexity values indicate better performance. While automated metrics are useful, human evaluation remains an essential aspect of summarization evaluation. Human judges can assess the overall quality, coherence, and informativeness of the generated summaries in a more nuanced and context-aware manner.

4. Methodology

4.1. Dataset Description

For this work, we used a random subset of 30 articles from the CNN/Daily Mail dataset. The CNN/Daily Mail dataset is a widely used benchmark dataset in the field of natural language processing (NLP) and machine learning. It was created by researchers at the University of Oxford and is named after the two news sources it primarily draws from: CNN (Cable News Network) and the Daily Mail. The dataset consists of news articles paired with human-generated summaries. It was originally introduced for the task of document summarization, where the goal is to generate a concise summary of a given news article. The articles in the dataset are diverse in topic, covering a wide range of news events and stories. Each example in the dataset consists of three parts: an article, a summary, and some additional metadata. The article is text of typically several paragraphs in length, and the summary is a shorter version that captures the key points and main ideas of the article. The metadata includes information such as the article's headline, the publication date, and other details. The dataset is particularly valuable for training and evaluating models that focus on abstractive summarization, where the generated summary is not limited to extracting sentences or phrases directly from the article. Instead, the models are expected to understand the content and generate human-like summaries that capture the essential information.

4.2. Model Description

We evaluate and compare 7 large language models including BERT, FALCON, GROOVY, ORCA, WIZARD, GPT 3.5, GPT 4.

GPT3.5: GPT-3.5, or Generative Pre-trained Transformer 3.5, is a subset of GPT-3 Models developed by OpenAI in 2022. OpenAI released updated versions of GPT-3 and Codex in its API on March 15, 2022, with additional features like edit and insert capabilities, labeled as "text-davinci-002" and "code-davinci-002."

GPT4: GPT-4 is an expansive multimodal model that can process both image and text inputs, producing text outputs. Although it may not match human capabilities in all real-world situations, it showcases human-level performance on numerous professional and academic benchmarks.

BERT (Bidirectional Encoder Representations from Transformers): BERT is a pre-trained language model developed by Google that uses transformer-based architectures. It is bidirectional, meaning it considers both left and right context of a word, resulting in better understanding of word meanings and context. BERT has been widely adopted for various natural language processing tasks due to its effective transfer learning capabilities.

FALCON: Falcon AI is a powerful, open-source Generative Language Model with 40 billion parameters, trained on 1 trillion tokens of RefinedWeb data. Its transparency and optimized architecture for inference make it stand out. Users can fine-tune Falcon for commercial use, and it outperforms state-of-the-art models on the OpenLLM Leaderboard. Falcon also offers Instruct versions for easy chat application creation. The extensive training on AWS Cloud with 384 GPUs, along with custom-made, high-quality data from RefinedWeb, contributes to Falcon's exceptional performance.

ORCA: Microsoft Research has introduced a novel AI model named Orca, which adopts an imitation-based learning approach from large language models. The research paper indicates that Orca aims to address the limitations of smaller models by emulating the reasoning processes of substantial foundation models like GPT-4. Models such as Orca have the advantage of task-specific optimization and can be trained using large language models like GPT-4. Due to its compact size, Orca demands fewer computing resources for its operation. This feature empowers researchers to optimize their models based on their needs and run them independently, reducing the reliance on large data centers.

Groovy: Also known as GPT4All groovy, it is a current leading commercially licensable model on GPT-J and trained by Nordic AI on the latest curated GPT4All dataset.

WIZARD: Researchers successfully trained large language models (LLMs) using AI-evolved instructions, outperforming human-created ones. The resulting WizardLM model showed promise in enhancing LLM capabilities, achieving over 90% of ChatGPT's capacity in 17 out of 29 skills.

The methodology employed in this study comprises various steps, which include the collection of data, pre-processing, summary generation by multiple large language models, summary evaluation using several metrics, and sentiment analysis. The following sections detail each step of the process.

4.3. Data Collection and preprocessing:

The first step in our methodology was to source our dataset. For this study, we chose a collection of articles from CNN and the Daily Mail. These sources provided a diverse range of topics and writing styles that enabled an extensive evaluation of the language models' summarization capabilities.

The articles were pre-processed during collection to fit the input format for the language models, Äîthis involved cleaning the text, such as removing HTML tags and other non-textual elements. We addressed all encoding problems during this phase.

4.4. Summary Generation:

The next phase involved utilizing each large language model: BERT, Falcon, Groovy, Orcar, Wizard, GPT-3.5 Turbo, and GPT-4, to generate summaries from the articles. Each model produced 30 summaries, resulting in a total of 210 summaries for each article. The prompts used requested the model generate the summaries without specifying any length or other constraints to observe the inherent abstracting capability of each model.

4.5. Summary Evaluation:

The generated summaries were then evaluated based on multiple metrics to assess their coherence with the original text. The metrics used in this study included Compression Ratio, ROUGE-1, Latent Semantic Analysis (LSA), Term Frequency-Inverse Document Frequency (TF-IDF), and Bilingual Evaluation Understudy (BLEU).

We performed the following evaluation analyses:

The Compression Ratio was used to compare the length of the original text and its corresponding summary.

ROUGE-1 was used to calculate the overlap of unigrams between the generated summaries and the original texts.

LSA was used to analyze the conceptual similarity between the original articles and the generated summaries.

TF-IDF was used to identify the importance of a word in a document compared to the corpus.

BLEU score was utilized to measure the similarity between the generated summaries and reference articles.

4.6. Sentiment Analysis:

After evaluating the summaries for coherence with the original text, we performed sentiment analysis using TextBlob and VADER (Valence Aware Dictionary and sEntiment Reasoner). The summaries were classified into one of three categories: positive, negative, or neutral, according to the polarity assigned by these tools. This multi-step process allowed us to deeply investigate the capabilities of the various large language models in text summarization tasks and how the sentiments they convey relate to the original text.

4.7. Bias Evaluation:

Drawing upon the top three models selected via performance metrics, we conducted a comparative analysis to ascertain whether traditional machine learning methods (TextBlob), a lexicon-oriented approach (VADER), or an integrated combination of these methodologies would yield the most reliable sentiment analysis. In this experiment, the GPT3.5 model was leveraged to generate three sets of summaries, each containing 30 samples that varied in sentiment: positive, negative, and fear. Subsequently, these summaries were evaluated using TextBlob (assessing Polarity and Subjectivity) and VADER to determine which sentiment analysis approach provided the highest accuracy.

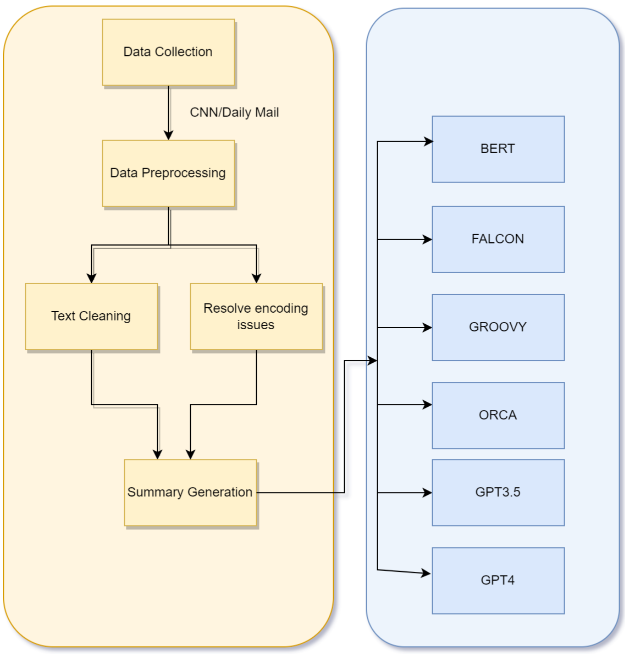

Figure 1. Data collection and pre-processing.

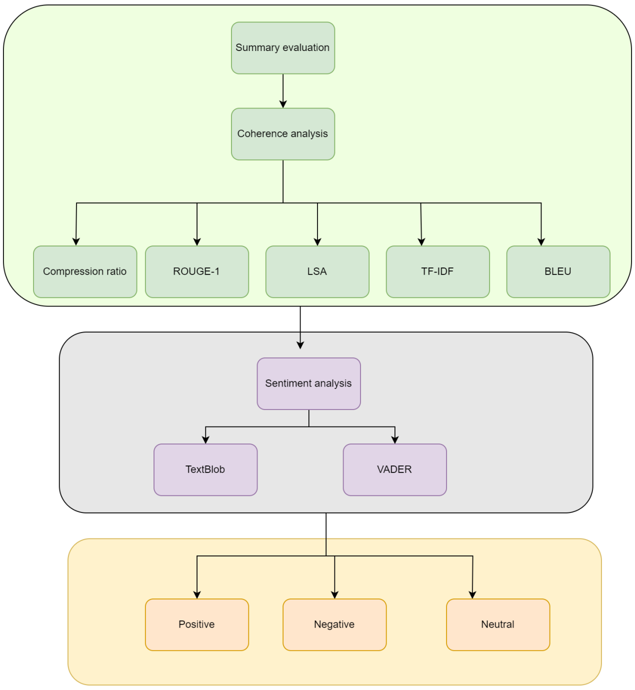

Figure 2. Sentiment evaluation framework.

Table 1. Assessing the coherence of summaries through conventional metrics.

Coherence	Definition	Evaluation metric
Compression Ratio	Refers to the measure of reduction in word count achieved when condensing an original text into a summary. It is the ratio between the summary's word count and the original text's word count, indicating the level of compression applied to the content.	The compression ratio quantifies text condensation by comparing the word count of a summary to the original text. A higher ratio indicates a greater level of compression applied to the source sentence.
ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation)	The ROUGE-1 and (their harmonic means) F-1 score measures the overlap and similarity between a generated summary and a reference summary at the unigram level, providing a single value that indicates the quality of the match.	Higher ROUGE-1 F-1 scores indicate a higher level of agreement and similarity between the generated summary and the reference summary regarding shared unigrams. Conversely, lower scores indicate a lower level of agreement and similarity.
Latent Semantic Analysis (LSA)	LSA similarity is a measure that quantifies the similarity between two pieces of text based on their underlying latent semantic meaning.	Lower LSA similarity scores indicate a lower level of similarity and may imply that the summary does not align well with the underlying semantic content of the input text.
TF-IDF (Term Frequency-Inverse Document Frequency)	TF-IDF assigns higher weights to terms that are frequent in a document but rare in the overall document collection, helping to identify key terms that are representative of the document's content.	Terms with higher TF-IDF scores are considered more significant or characteristic of the document's content.
BLEU (Bilingual Evaluation	A metric used in natural language processing to evaluate the quality of machine-generated translations by comparing	It ranges from 0 to 1, where 1 means the machine-generated translation perfectly matches

VADER (Valence Aware Dictionary and sEntiment Reasoner)	VADER utilizes a lexicon-based approach, where sentiment scores are assigned to individual words based on their semantic orientation. VADER also considers the context of the text, including punctuation, capitalization, and degree modifiers, to provide more accurate sentiment analysis results.	A sentiment score of -1 signifies a highly negative sentiment, +1 indicates a highly positive sentiment, and 0 represents a neutral sentiment. The score reflects the overall sentiment or emotional polarity of the text.

5. Results

We scored the summaries generated using BERT, FALCON, GROOVY, ORCA, WIZARD, GPT3.5, GPT4 for thirty random articles from CNN/Daily Mail dataset with the traditional metrices i.e Compression Ratio, Rouge, LSA, TF-IDF, BLEU, Polarity, Subjectivity, VADER. The results are in Table 2

Table 2. Metrics results.

Model	Compression Ratio	ROUGE	LSA	tf-idf score	BLEU score	Polarity	Subjectivity	Vader score
BERT	0.171	0.013	0.047	0.372	0.251	0.559	0.529	0.111
FALCON	0.252	0.060	-0.017	0.298	0.397	0.670	0.593	-0.084
GROOVY	0.226	0.0178	0.011	0.333	0.341	0.557	0.498	-0.187
ORCA	0.259	0.058	0.0349	0.450	0.384	0.673	0.616	-0.075
WIZARD	0.226	0.018	0.011	0.333	0.341	0.557	0.498	-0.186
GPT3.5	0.391	0.099	0.019	0.349	0.459	0.750	0.684	-0.177
GPT4	0.514	0.153	0.013	0.359	0.501	0.773	0.716	-0.047

Later to evaluate bias, we performed sentiment analysis with the GPT3.5 summaries modified based on 3 sentiments: positive, negative, and fear and forther evaluated with polarity, subjectivity and VADER score. The results are in Table 3

Table 3. Bias evaluation.

Model	Polarity	Subjectivity	Vader score
Positive	0.301	0.188	0.601
Negative	0.307	0.192	-0.687
Fear	0.290	0.148	-0.729

6. Discussion

The GPT-4 model performed best across much of our analyses. A close examination of several traditional metrics for summarization efficacy informs this conclusion. For instance, GPT-4's Compression Ratio of 0.514 leads the pack, reflecting its significant capacity to distill vital information effectively without losing crucial details. Similarly, its ROUGE-1 score, which measures unigram overlap between generated and reference summaries, stands at 0.153, again the highest among the models compared. This score testifies to GPT-4's excellent ability to match reference summaries, which is critical to producing high-quality summaries. While GPT-4 does not score highest in Latent Semantic Analysis (LSA), it maintains consistency, unlike models such as FALCON and GROOVY, which score negatively, suggesting difficulties in preserving the semantic meaning from the original text in their summaries. Regarding Term Frequency-Inverse Document Frequency (TF-IDF),

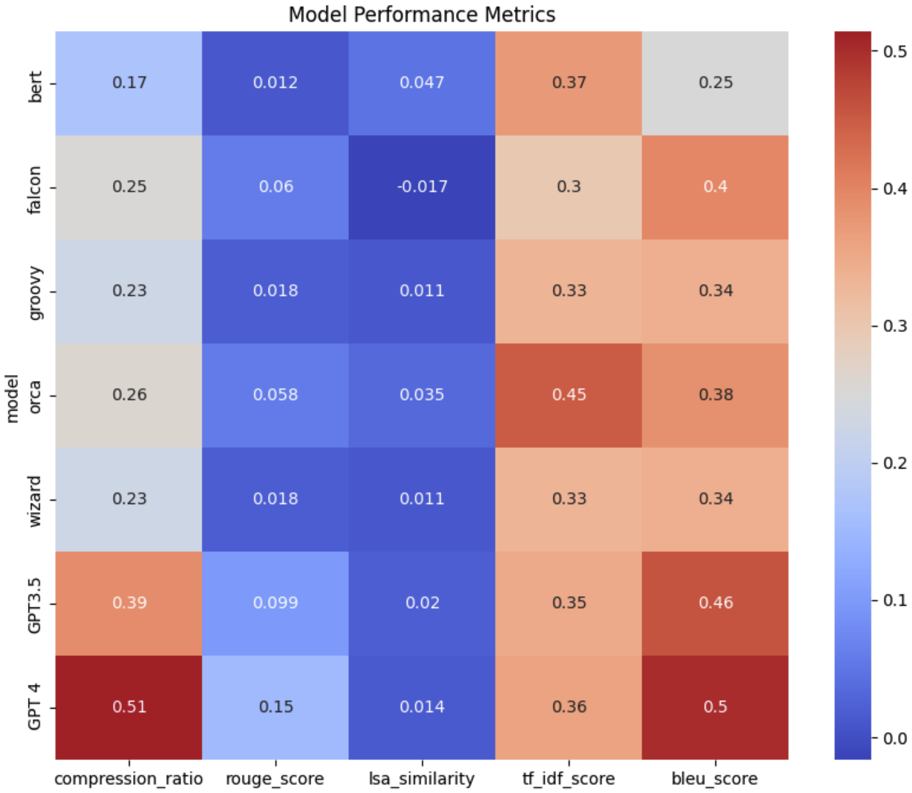

Figure 3. Metrics results.

Figure 4. Bias evaluation.

GPT-4 posts a solid score of 0.359, indicating its competence in identifying and retaining key terms that encapsulate the document's context and meaning. Another impressive aspect is the BLEU score, which gauges the accuracy of machine-generated translations.GPT-4 outperforms all other models with a score of 0.501, underlining the model's proficiency in generating translations that closely align with human-produced versions. Given the results derived from these metrics, it is safe to conclude that GPT-4 exhibits the highest efficacy in generating coherent summaries among the models analyzed.

GPT 3.5 Turbo secured the second-best position according to the performance metrics analyzed. Notably, GPT3.5 performs effectively in terms of text compression, achieving a Compression Ratio of 0.391, thereby establishing itself as a proficient model in retaining the core essence of information while achieving brevity. Additionally, its ROUGE-1 score of 0.099 signifies a substantial degree of agreement with the reference summary at the unigram level, marking it as second-highest in this aspect, indicative of its capability to generate summaries closely mirroring the reference.

Though not the front-runner in Latent Semantic Analysis (LSA), GPT3.5 nonetheless demonstrates relative stability, outperforming several other models and thereby illustrating its competence in preserving semantic correlation between the source text and the generated summary. In terms of the TF-IDF metric, GPT3.5 yields a score of 0.349, which, while not the peak score, still highlights its adeptness at identifying and incorporating crucial terms into its summaries. Furthermore, GPT3.5 achieves a BLEU score of 0.459, placing it second in terms of producing summaries that align well with the orginal article. Cumulatively, these results clearly reflect GPT3.5's commendable performance in generating coherent summaries, substantiating its ranking as the second-best model.

ORCA emerges as the third-best alternative for generating coherent summaries from the given text. An exploration of the data reveals the reasons for this ranking. Despite having the fourth-highest Compression Ratio of 0.259, ORCA demonstrates an adequate capability for text compression, a fundamental aspect of summary generation. In terms of ROUGE-1 score, which measures the overlap of unigrams between the generated and reference summaries, ORCA's score of 0.058 ranks third among the compared models, indicating a respectable degree of similarity with the reference summary.

The model's Latent Semantic Analysis (LSA) score, which gauges the semantic similarity between the original text and the produced summary, stands at 0.0349, suggesting a moderate level of preservation of semantic meaning in the generated summaries. Notably, ORCA outstrips all other models in the Term Frequency-Inverse Document Frequency (TF-IDF) measure, posting a score of 0.450. This signals a robust capability for identifying and retaining key terms that capture the essence of the document's content.

Lastly, ORCA's BLEU score, a metric that evaluates the closeness of machine-generated translations to original article, is 0.384, thereby ranking it third. This score signifies a reasonable degree of alignment between ORCA's generated summaries and the original text. Given these observations, it can be inferred that ORCA offers a solid, third-best option for creating coherent summaries.

Sentiment Analysis: Table 2 provided insight into the sentiment of the evaluated models, encompassing traditional approaches like BERT, FALCON, GROOVY, ORCA, and WIZARD, as well as the more recent ChatGPT3.5 and GPT4. Based on TextBlob's Polarity and Subjectivity metrics, traditional models generally produced positive summaries, with Polarity scores from 0.559 (BERT) to 0.673 (ORCA). Subjectivity scores ranged from 0.498 (GROOVY and WIZARD) to 0.616 (ORCA), indicating more subjective summaries. However, Vader scores, while negative, were close to zero, suggesting a slight discrepancy with TextBlob's generally positive Polarity scores. ChatGPT's evaluations showed a more positive sentiment, with Polarity scores for GPT3.5 and GPT4 at 0.750 and 0.773, respectively. Subjectivity scores were also higher, but Vader scores echoed the trend in traditional models with near-neutral results. When comparing traditional metrics with GPT4's assessments, GPT4 exhibited a stronger positive sentiment and greater subjectivity. However, the contradiction between the positive sentiment from TextBlob and near-neutral

Vader scores across all models, including GPT4, may be attributed to different sentiment quantification methods. Overall, the analysis suggests that, when using Vader scores, all models, particularly GPT4, Falcon, and Orca, generate summaries with the most accurate sentiment.

Bias Evaluation: Analyzing the results from table 3, we can observe how GPT3.5 performed when generating summaries with different sentiments and the corresponding evaluation of these summaries by TextBlob and VADER metrics. The Polarity scores from TextBlob indicate a positive sentiment in all three cases (positive, negative, and fear), which is unexpected. While the GPT3.5 positive sentiment summaries are correctly labeled as positive (0.301), the negative and fear summaries are also scored as positive with 0.307 and 0.290, respectively, contradicting our expectation that these should give a negative result.

The Subjectivity scores from TextBlob are pretty low across all categories, indicating that the summaries are more objective than subjective. The values are close, ranging from 0.148 (fear) to 0.192 (negative), providing little distinction between the different sentiment categories. When looking at the VADER scores, a different picture emerges. VADER successfully identifies the sentiment of GPT3.5 summaries in line with our expectations. The GPT3.5 positive summaries have a high positive VADER score of 0.601. The VADER scores are negative, as anticipated for the GPT3.5 negative and fear summaries, with -0.687 and -0.729, respectively. These scores accurately reflect the intended sentiment of the summaries. Based on the analysis of these results, we can conclude that in this context, VADER outperforms TextBlob in accurately assessing the sentiment of the summaries produced by GPT3.5. The lexicon-based approach of VADER, which also considers the context of the text, has proven to be more effective in distinguishing between positive, negative, and fear-based sentiment. Therefore, we recommend using VADER for sentiment analysis of text generated by the GPT3.5 model.

7. Conclusion

This study evaluates a set of LLM models and their propensity to introduce biases during text summarization, highlighting their strengths and weaknesses. GPT-4 performed best in generating coherent summaries, with GPT 3.5 Turbo and ORCA closely behind. A discrepancy was observed between the generally positive TextBlob Polarity scores and the near-neutral Vader scores, possibly due to different sentiment quantification methods. GPT-4 showed a tendency towards positive, somewhat subjective sentiment. Interestingly, VADER effectively gauged sentiment in GPT 3.5-generated summaries, surpassing TextBlob in context-driven sentiment analysis. These insights can improve automated text summarization, content analysis, and sentiment analysis, potentially benefiting news summarization, content filtering, and social media sentiment analysis. Future research should focus on reconciling sentiment analysis discrepancies, examining model performance in different languages and text genres, and addressing model biases in pursuit of ethical AI systems. These findings can help inform the intelligence communities about best practices to reduce bias in summaries created by humans or LLMs, improving analysis quality and rigor. Ensuring unbiased summaries is critical to support their role in decision-making processes. Despite their ability to generate relatively accurate summaries, models like GPT-4 and GPT 3.5 Turbo can introduce some bias. VADER, a lexicon approach, considering its effectiveness in context-based sentiment analysis, can help identify and mitigate bias.

References

1. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018.
2. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification, 2018, [arXiv:cs.CL/1801.06146].
3. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].

4. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
5. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
6. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 2004; pp. 74–81.
7. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.