Summer 2024

Who Wrote the Scientific News? Improving the Discernibility of LLMs to Human-Written Scientific News

Dominik Soós
Old Dominion University, soos.domi@gmail.com

WHO WROTE THE SCIENTIFIC NEWS? IMPROVING THE DISCERNIBILITY OF LLMS TO HUMAN-WRITTEN SCIENTIFIC NEWS

by

Dominik Soós
B.S. May 2023, Old Dominion University
M.S. August 2024, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2024

Approved by:

Jian Wu (Director)

Vikas Ashok (Member)

Meng Jiang (Member)

ABSTRACT

WHO WROTE THE SCIENTIFIC NEWS? IMPROVING THE DISCERNIBILITY OF LLMS TO HUMAN-WRITTEN SCIENTIFIC NEWS

Dominik Soós
Old Dominion University, 2024
Director: Dr. Jian Wu

Large Language Models (LLMs) have rapidly advanced the field of Natural Language Processing and become powerful tools for generating and evaluating scientific text. Although LLMs have demonstrated promising as evaluators for certain text generation tasks, there is still a gap until they are used as reliable text evaluators for general purposes. In this thesis project, I attempted to fill this gap by examining the discernibility of LLMs from human-written and LLM-generated scientific news. This research demonstrated that although it was relatively straightforward for humans to discern scientific news written by humans from scientific news generated by GPT-3.5 using basic prompts, it is challenging for most state-of-the-art LLMs without instruction-tuning. To unlock the potential evaluation capability of LLMs on this task, we propose guided-few-shot (GFS), an instruction-tuning method that significantly improves the discernibility of LLMs to human-written and LLM-generated scientific news. To evaluate our method, we built a new dataset, SANews, containing about 362 triplets of scientific news text, LLM-generated news text, and the corresponding scientific paper abstract on which the news articles were based. This work is the first step for further understanding the feasibility of using LLMs as an automated scientific news quality evaluator.

I dedicate my thesis to my grandfather. I wish he could have been here to see this achievement and what's to come. His memory and influence continue to inspire me every day.

ACKNOWLEDGMENTS

First, I would like to thank my parents, who are my greatest supporters. My mother, whose passion and hard work continue to inspire me, and my father, who has shaped me into the person I am today. I would also like to appreciate Dr. Wu for his continued guidance throughout this process. This work would not have been possible without his support. I also want to acknowledge Dr. Jiang, who has been instrumental in guiding me with this project for over a year. Dr. Ashok, your assistance has greatly enhanced my understanding of NLP and language modeling. Thank you for your help.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

In recent years, the field of Generative Artificial Intelligence (GenAI), a branch of Natural Language Processing (NLP), has experienced rapid advancements, opening new opportunities for generating comprehensive textual content [24]. Among the various breakthroughs, large language models (LLMs) such as GPT3/4 [1], [8], [19], Claude 3 family [2], LLaMA 3 [39], and Mistral [18] have emerged as powerful tools for various NLP tasks, more specifically for producing informative text across domains [22], [36] and the automatic evaluation of such content [10]. LLMs have shown promising results regarding the scalability and evaluation of large amounts of news text. However, these models have also introduced potential biases toward LLM-generated content and this bias can potentially manifest itself by favoring certain styles and types of content that resemble the data they were trained on [6], [12]. Such bias is especially harmful in scientific communication, where accurate representation and evaluation are crucial. Scientific research papers often contain complex ideas and specialized terminology for a specific domain that can be intricate for nonexperts to comprehend. However, scientific news text plays an essential role in our society bridging the gap between scientific advancements and the general public. This thesis explores the potential of using LLMs as automated evaluators of news text and paving the way for mitigating their bias toward other LLM-generated content to ensure fair evaluations of news text.

1.2 PROBLEM DESCRIPTION

Fair evaluation of AI-generated content is of great importance for the advancement of GenAI research, and guarding the quality of such content disseminated on the web. The fact that neural retrievers are biased toward other LLM-generated content can be traced back to source bias [12]. Their study shows that neural retrieval models often rank texts generated by LLMs higher than similar human-written texts. This bias exists because LLMs generate content that is more semantically focused meaning that they have a higher semantic density therefore are easier for retrieval models to process [13]. On the other hand, LLMs are often trained on large datasets that may include LLM-generated content. These models then prefer other LLM-generated content. This phenomenon aligns with the principles of machine learning where a model's performance is optimized for data distributions similar to its training set. Models trained on data that closely resembles their test data perform better because they have learned the underlying patterns and features present in that specific distribution [17]. In the context of automated text evaluation, bias can result in preferential treatment toward LLM-generated content, which may not always reflect the most accurate or relevant information we want to give to the general public. In Figure 1 we can see the potential bias of the Evaluator LLM toward the news article written by Journalist GPT. We verify the claims of [12] that giving a direct single score may only produce suboptimal results. LLMs may use a set of underlying rubrics different from the rubrics built-in human common sense. It is not easy to know what these rubrics are and it may not be correct to interpret the rubrics in the response of an LLM as the true rubrics it uses. Many deep learning models make it difficult to pinpoint and address the exact source of bias [31]. Tracking down the source of bias in LLMs is out of the scope of this thesis. Instead, we are focused on mitigating the already existing bias toward

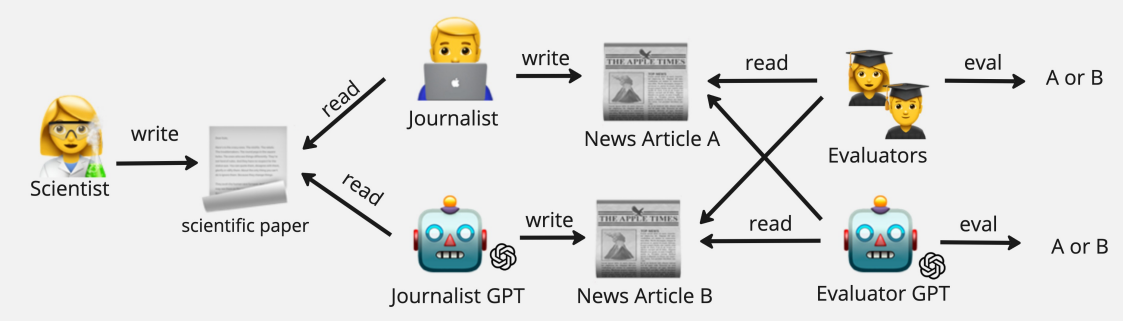

Figure 1. Process Flow of Evaluating between Human-Written news from LLM-generated ones

LLM-generated content by guiding these models using prompt engineering techniques. Mitigating the bias of LLMs is a larger topic and may need far more work than this thesis can cover. The proposed methods work well for discerning human-written and LLM-generated scientific news. Additionally, our task is to discern human-written and LLM-generated news, not to evaluate the quality. Future work may include the improvement of generating text as well as the quality evaluation.

1.2.1 Challenges in Mitigating Bias

Challenges of text evaluations, such as the limitations of the standard text summarization metrics and the known limitations of LLMs. Our task cannot be based on traditional token-based metrics. The direct scoring method was shown to be suboptimal for news summarization tasks [12]. We want to test it for ourselves using our data because scientific news text has not been explored extensively.

1.3 RESEARCH QUESTIONS

The main questions this thesis would like to address are the following;

- Can humans distinguish human-written text from LLM-generated scientific news?
- Can LLMs distinguish human-written text from LLM-generated scientific news?
- Do commercial and open-source LLMs have different capabilities for this task?
- How can we design prompts so LLMs can effectively discern human-written and LLM-generated scientific news?
- Can we use guided few-shot prompt engineering techniques to mitigate the bias of direct evaluation

In the future, we will explore LLMs to unlock their potential to be used as evaluators of *text quality*. Although the data we compared are scientific news, theoretically, the methods can be generalized to other news articles.

1.4 MAJOR CONTRIBUTIONS

The main goal of this work is to discover whether humans or LLMs are capable of distinguishing between human-written and LLM-generated content. In order to reach such goals, we propose methods to mitigate the bias toward other LLM-generated content. To develop these methods, first, we need to develop strategies to enhance the ability of LLMs to distinguish between human-written and AI-generated content. To ensure reliable evaluation methods, we propose methods through carefully crafted prompts to guide the model's output.

The contributions of this work are threefold:

- We developed a new dataset called SANews for examining the feasibility of LLMs as automated evaluators. It contains the metadata of scientific news articles with the triplet of annotated news articles, linked to research papers, and LLM-generated news articles written based on those papers.
- We propose guided few-shot, a novel approach of utilizing pairwise comparison with prompt engineering techniques and examples to improve the accuracy of LLMs.
- We verified the weaknesses of using direct scoring in the context of scientific news evaluations

This work provides a new baseline for pairwise comparison by providing multiple examples of the LLM to further prove that LLMs are few-shot learners [7].

CHAPTER 2

RELATED WORK

In recent years, the need for automated text evaluation has gained attention because of the thriving of the NLG. To evaluate machine-generated text, we usually take human evaluations as the gold standard. Usually, we want to measure the quality of the text using quantitative measures such as a score. In some cases, experts are needed for evaluating text containing domain knowledge while in other cases, the general human subjects are sufficient. The latter allows us to launch larger-scale evaluations using crowdsourcing. However, crowd-sourced human evaluation is constrained by funding availability and time scale and thus is not always feasible for many NLG tasks that need frequent evaluations on a large corpus of test samples, which calls for automatic evaluation metrics. Traditional evaluation metrics, such as ROGUE, can only be used for measuring lexical similarity, but not semantic similarity between the ground truth and the test samples. It also cannot be used for scoring the quality of a piece of text or directly comparing two pieces of text LLMs like GPT-4 offer a more powerful solution to provide fast and semantic text evaluation without any ground truth. It can also be used for pairwise comparison between two or multiple pieces of text.

2.1 NEWS ARTICLE GENERATION DATASETS

Previous work has not been done extensively in the scientific news domain. They mainly focus on methodologies for the generation and summarization task of general news articles [34], [40]. A recently published work introduced the **SciNews** dataset, which provides a comprehensive collection of scientific news and their corresponding research papers [29]. This dataset was designed to

help in the development of NLG models that can translate scientific content into accessible news reports by the general public. They proposed a new task namely Automated Scientific News Generation Task (SNG) tailored for this dataset. The dataset is mostly compiled from the ScienceX platform[1], ensuring high-quality content through rigorous editorial standards. Their data cleaning and quality control process includes automated BERT similarity checks and human evaluations to maintain consistency and quality [29]. Another study introduced by the Massachusetts Institute of Technology collected news articles written by freelance writers for the news article summarization task. [43]. They evaluated LLMs using the **CNN/DailyMail** and **XSUM** which are still one of the most frequently used datasets in the realm of news summarization domain [25], [26]. CNN/DailyMail dataset offers a large corpus of news articles paired with human-generated summaries. It is known for its extensive use for the news summarization task. XSUM, on the other hand, focuses on creating extremely concise summaries, providing a challenging benchmark for summarization models. **SumPubMed** is another dataset that is focused on summarizing scientific articles from the PubMed archive [16]. This dataset posed a unique challenge at the time, due to the complex domain-specific found in medical literature, making it a valuable resource for improving summarization models [16]. However, this dataset is not about scientific news articles. The **N24News** dataset stands out for its multimodal approach, featuring both text and images sourced from New York Times news articles [41]. This dataset includes a diverse range of categories and leverages images to enhance text classification tasks. The use of visual data allows for more sophisticated classification problems and provides a richer context for understanding the content, thereby improving the accuracy of classification models [41].

These datasets collectively contribute to advancing the field of automated news article summa-

[1]https://sciencex.com/

rization/generation. Each one offers a unique challenge and resources to improve models' robustness and accuracy in summarizing and generating scientific news content.

2.2 NEWS GENERATION EVALUATION METRICS

The potential of automatically evaluating the quality of human and AI-generated language has been explored extensively [10], [15]. Traditionally, they rely on the use of automated metrics such as ROGUE, BLEU, or BERTScore and human judgments. The motivation for using LLMs to evaluate news text stems from the limitations of existing evaluation methods.

2.2.1 Standard Evaluation Metrics

Several approaches have been proposed to use language models for NLG evaluation. For instance, BERTScore uses BERT embeddings to compare the generated text with reference text, showing a better correlation with human judgments than traditional metrics [42]. Similarly, BLEURT fine-tunes a BERT-based model on human evaluation scores to provide more nuanced assessments [32]. However, these models still face challenges in fully capturing the human alignment aspect due to their training limitations and the scope of data they were exposed to.

To evaluate the quality of the generated news articles, we assume that news articles written by news editors have the highest quality, and use them as the ground truth. The evaluation metrics include standard text summarization metrics.

- The ROUGE scores are recall-oriented scores that [20] measure the quality of a summary by counting the number of overlapping *n*-grams (ROUGE-1 and ROUGE-2) or the longest word sequences (ROUGE-L) between the GPT-generated news to be evaluated and the ground truth news created by news editors.

- The BLEU scores [27], originally proposed as a machine translation evaluation metric, calculate the precision of n-grams in the GPT-generated news by comparing them to the ground truth news created by news editors. The precision is then modified by a brevity penalty to account for generated news that is shorter than the ground truth news.
- METEOR [5] calculates precision and recall similarly to BLEU scores but, it produces a more comprehensive alignment with human judgment by also considering the meaning of words.
- BERTScore [42] computes the contextual similarity using the embeddings of each token in the GPT-generated news with each token in the ground truth.
- BLEURT [32] is using contextual embeddings from BERT to evaluate the quality of NLG by comparing it to human references.

While metrics like BLEU, ROUGE, and BERTScore provide quantifiable measures of similarity between generated text and reference text, they often fail to capture the nuanced aspects of language quality such as coherence, fluency, and contextual appropriateness [5], [20], [27]. Automated metrics, while efficient, do not align well with human judgment and can lead to misleading conclusions about the quality of generated text .Human evaluations, on the other hand, seem to be more accurate, they are time-consuming, subjective, and suffer from scalability issues.

2.2.2 LLMs as Evaluators

Traditional evaluation metrics, such as BLEU and ROUGE, have been criticized for their inability to fully capture the quality of the generated text. These metrics are largely based on n-gram overlap with reference texts and do not account for logical consistency, creativity, or contextual

appropriateness [30]. Scoring methods and pairwise comparison, on the other hand, align more closely with human judgment and can provide more reliable assessments of text quality. Baseline methods include prompting LLMs to give a direct score as a representation of the quality of the text given a reference text it was generated upon. Another approach involves using LLMs themselves to perform pairwise comparisons between the quality of two given texts. We explore both of these methods extensively.

Scoring Method

The capabilities of LLMs such as GPT-4 have been explored in G-Eval in evaluating natural language generation aligned with human judgment [21]. Its ability to generate and evaluate text with a high degree of fluency and coherence makes it a suitable candidate for this task. This method leverages GPT-4 to perform evaluations that have shown to be more closely aligned with human judgment than other LLMs or traditional metrics. This method allows GPT-4 to better resemble human-like reasoning in assessing text quality, resulting in more accurate and reliable evaluations. Purely comparing the quality of two texts may not be the most accurate evaluation technique [9]. They discovered that the use of Explicit Scoring outperforms others. Previous research has shown that incorporating model-based evaluations can improve alignment with human judgments, but these models were often limited in their capacity to understand and generate complex language structures [44]. The scoring method has shown to have its limitations and may only result in suboptimal evaluation performance [10].

Pairwise Comparison

Several studies have explored the use of pairwise comparisons for evaluating NLG systems. Pairwise comparison is a method commonly used in the evaluation of text generated by LLMs. This approach involves comparing two pieces of text side-by-side to determine which one is better according to some criteria, such as fluency, coherence, or relevance. Pairwise comparison helps to mitigate some of the limitations of traditional automated metrics by providing more nuanced and human-like evaluations [14]. One notable approach is the use of preference-based reinforcement learning, where human preferences between pairs of text are used to train models to generate higher-quality content that is more aligned with human content [11]. The pairwise comparison method has shown significant improvements in generating more human-aligned text and their evaluations.

Pairwise comparison aligns more closely with human judgment and can provide more reliable assessments of text quality. One approach involves using LLMs themselves to perform pairwise comparisons. For example, OpenAI's GPT models have been used to evaluate pairs of generated texts by scoring them based on various qualitative criteria [45]. This method leverages the language model's capability of understanding context and coherence to provide more sophisticated evaluations. Another work by OpenAI was used for learning to summarize with human feedback [37].

2.3 LIMITATIONS AND CHALLENGES

Despite its advantages, direct score methods and pairwise comparison face several challenges. One major issue is the scalability of obtaining human experiments, which can be costly and time-

consuming. Additionally, ensuring the consistency and reliability of human judgments can be difficult, as different evaluators may have varying opinions on what constitutes "better" text [38]. Another challenge is the potential for biases in the preference data used to train the models. If the training data is already biased toward AI-generated content, the model's evaluations may reflect those biases, leading to unfair evaluations. Addressing these challenges requires carefully crafted experiments and robust methodologies to ensure fair and accurate evaluations.

Previous research has shown that incorporating model-based evaluations can improve alignment with human judgments, but these models were often limited in their capacity to understand and generate complex language structures [44]. In their paper, [10] demonstrated that prompting the LLM to provide a single score value is suboptimal due to the lack of explanatory depth, which significantly improves the correlation between the model's ratings and human ratings. They evaluated the use of GPT-3.5 as the LLM, but they did not explore GPT-4 due to limited access. Open source LLMs such as LLaMA and Mistral underperformed GPT-4 on text evaluation and thus are not suitable for such tasks.

CHAPTER 3

DATA

One of the main contributions of this work is the dataset that was crawled, processed, and annotated over a long period.This dataset incorporates 23K scientific news articles' metadata. Due to the cost of using state-of-the-art LLMs and data selection criteria, only about 3% of the total crawled data were used in the experiments. Our process begins with the collection of the data, then pre-processing, and ends with the annotation of a training dataset comprising 506 news articles from ScienceAlert, each associated with one or more published scientific research papers.

3.1 WEB-CRAWLING

Our goal was to collect a corpus of news articles summarizing or reporting original scientific research studies, containing references to original scientific articles. To this end, we developed a custom web crawler to automatically download 23,674 HTML that contains the scientific news articles from ScienceAlert[1], which is an online publication platform focusing on scientific research and discoveries, with content created by human editors. The extracted data include key details such as the processed news article text, title, category, authors, publication dates, HTML content, and all URLs. The date of the articles range from 2014-2022.

3.2 PRE-PROCESSING

After collecting the metadata, the URLs in the HTML were processed to identify articles di-

[1]ScienceAlert: https://www.sciencealert.com/

rectly linked to scientific research papers from recognized academic domains. This step ensured our focus on articles summarizing or interpreting original research studies, rather than general scientific news. There may be more than one research paper linking to a ScienceAlert article.

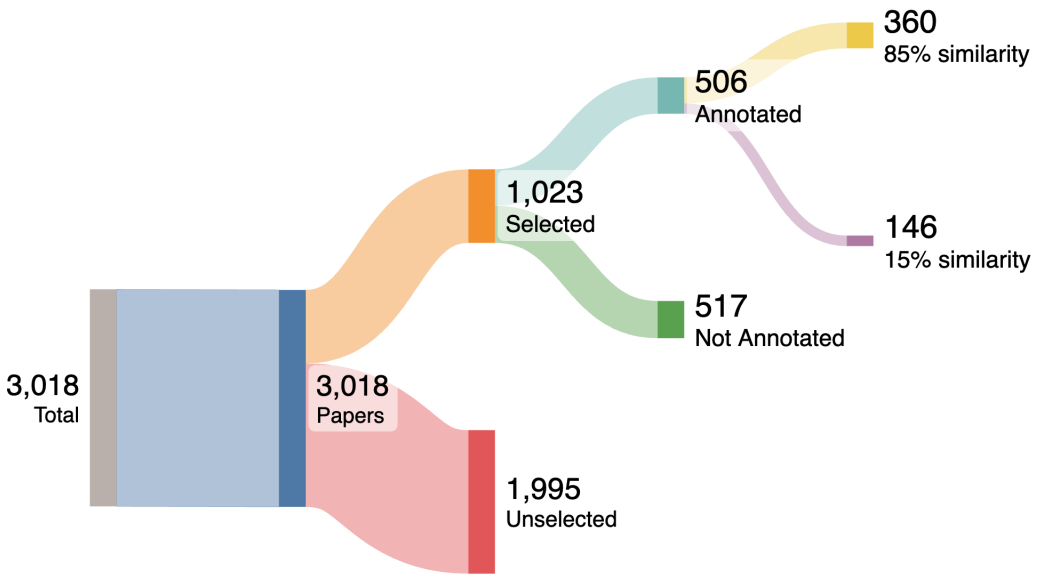

Figure 2. Sankey Diagram of the Breakdown of the Numbers

A large fraction of ScienceAlert articles do not contain a URL linking to a scientific publisher domain i.e. a research abstract. We found more than 3K URLs that link to research papers. The Sankey diagram in Figure 2 visually illustrates how the numbers changed after each filtering step. We tried to automatically retrieve the content of such links, but they are mostly protected by firewalls. Out of the entire set of scraped ScienceAlert articles that link to research papers (3,018), a distinct set was randomly selected based on the category distribution. Before processing, the ScienceAlert articles were split into equal numbers of articles per category. Since there may be

a ScienceAlert article that links to more than one Table 1 shows the distribution of the complete number of ScienceAlert articles that are linked to scientific research papers. A fraction of the news articles were labeled as "uncategorized" articles because they were among the articles published in 2014.

Table 1. URL and ScienceAlert Category Table.

Category	**URL**	**ScienceAlert**
Nature	59	49
Uncategorized	76	70
Tech	44	42
Physics	19	19
Health	65	59
Space	34	34
Humans	61	51
Environment	41	37
Society	1	1
Total	**400**	**362**

3.3 GROUND TRUTH ANNOTATION

From the processed URLs, we selected and annotated 600 news articles directly linked to scien-

tific research articles. There may be more than one research paper linking to a ScienceAlert article. These annotations encompassed essential information about the article, including its source, the research paper reference, and the domain of the research article. Through this effort, we have compiled a dataset that enables exploration of how scientific research is translated into news generation. This dataset, the first of its kind, will serve as a valuable resource for future research on article generation and related applications.

The annotation was done independently by two graduate students in the Computer Science Department. After the annotation, the final dataset was selected by calculating the similarity scores of two annotations using the longest common subsequence. Then a threshold of 85% was applied to the similarity scores. Figure 2 illustrates the change in numbers visually.

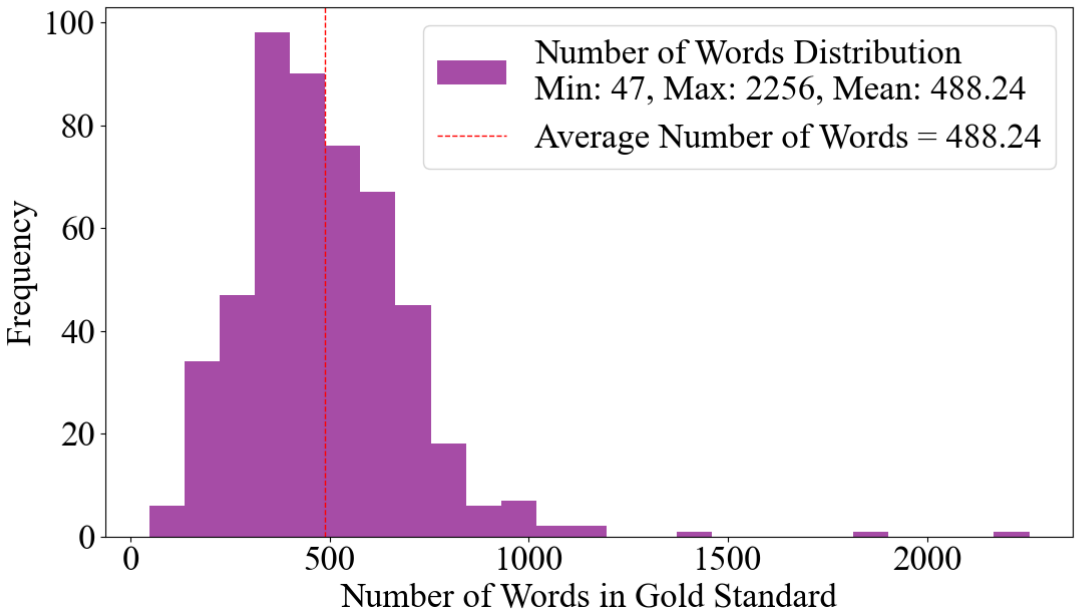

Figure 3. Word Count Distribution for ScienceAlert articles

3.4 SYNTHETIC ARTICLE GENERATION

The baseline method of the article generation is by prompting the LLM to act as if it were a journalist who can read and understand scientific research language. Then, we asked GPT to directly generate a news article given a research paper abstract.

Since the LLM-generated article is directly generated based on a single research paper abstract there might be more than one paper linked to one ScienceAlert article. We explored three different methods to generate news articles.

1. **Random**

   In the random method, we prompt Journalist GPT to generate N number of stories given a scientific abstract. Then, prompt another Journalist GPT to write a news article based on those stories.

2. **Iterative**

   In this method, given a scientific abstract, we prompt a Journalist GPT to generate a news article. Then $N-1$ times ask GPT to read the previous articles and draft another one in a different writing style. We prompt another Journalist GPT to write a news article based on N news articles. In this way, we are trying to cover most of the semantic space.

3. **Boost**

   This method tries to improve at every iteration. We first ask Journalist GPT to read the scientific abstract and write a news article about it. Then $N-1$ times we ask Journalist GPT to write a better story than the last one. We just take the last news article as machine-generated text.

3.5 TERM FREQUENCY AND DOCUMENT FREQUENCY

This section provides an analysis of the distributions of the Term Frequency (TF) and the Document Frequency (DF) in three fields: *annotation*, *abstract*, and *generated_article*. The analysis is visualized through two plots that show the similarities between the TF and DF distributions of the top 20 terms in these fields. The patterns that we observed in these distributions are discussed in the context of Zipf's law.

3.5.1 Term Frequency

Figure 4 shows the top 20 terms by frequency for each field. TF is a measure of how often a term appears in the entire corpus. Common terms such as "the", "of", "and", "in", and "to" dominate the distribution, which is typical of any natural language corpus. This is consistent with Zipf's law, which states that the frequency of any word is inversely proportional to its rank in the frequency table [35].

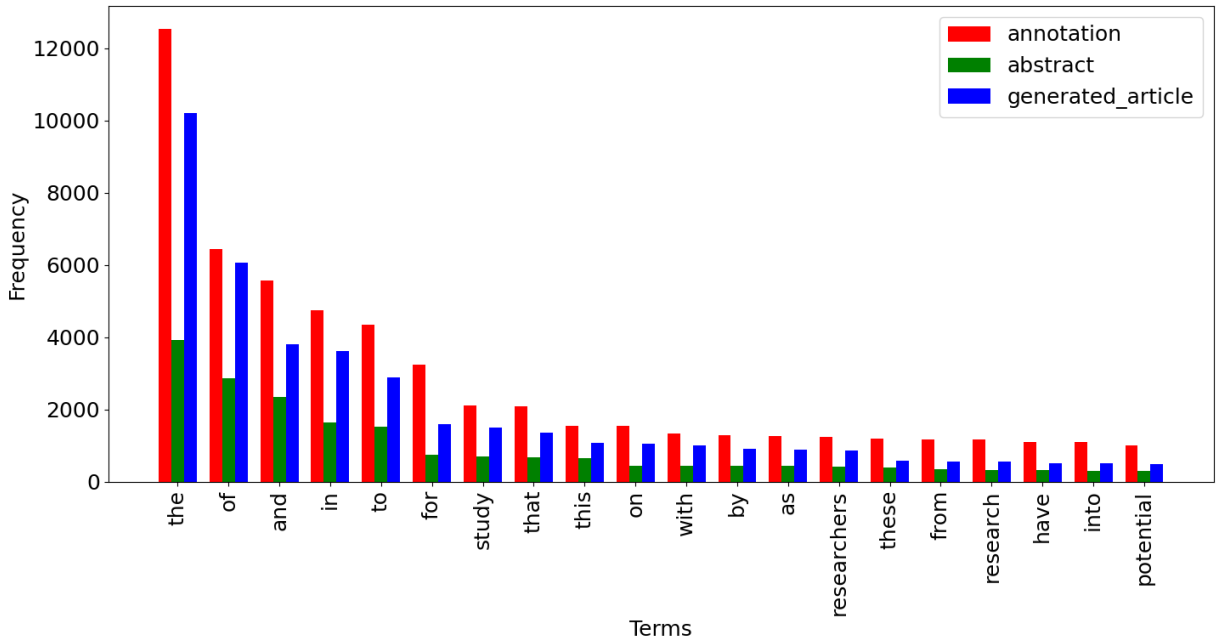

Figure 4. Term Frequency Distribution for *annotation*, *abstract*, and *generated article*.

The terms follow a similar pattern throughout the abstract, annotation and generated article, with the highest frequencies found in the *annotation* as it also has the highest word count shown in Figure 3. It is followed by *generated article*, and then *abstract*. Based on the distribution, the *annotation* contains more descriptive features/words, which is natural for human-written text. From Figure 6, we also know that the length of the ScienceAlert annotation is greater than the lengths of the abstract and the generated articles by at least 100 words on average.

3.5.2 Document Frequency Distribution

Figure 5 displays the top 20 terms by document frequencies for each field. The document frequency DF indicates how many documents contain a particular term. Similar to the term frequency

distribution, the common terms dominate the DF distribution.

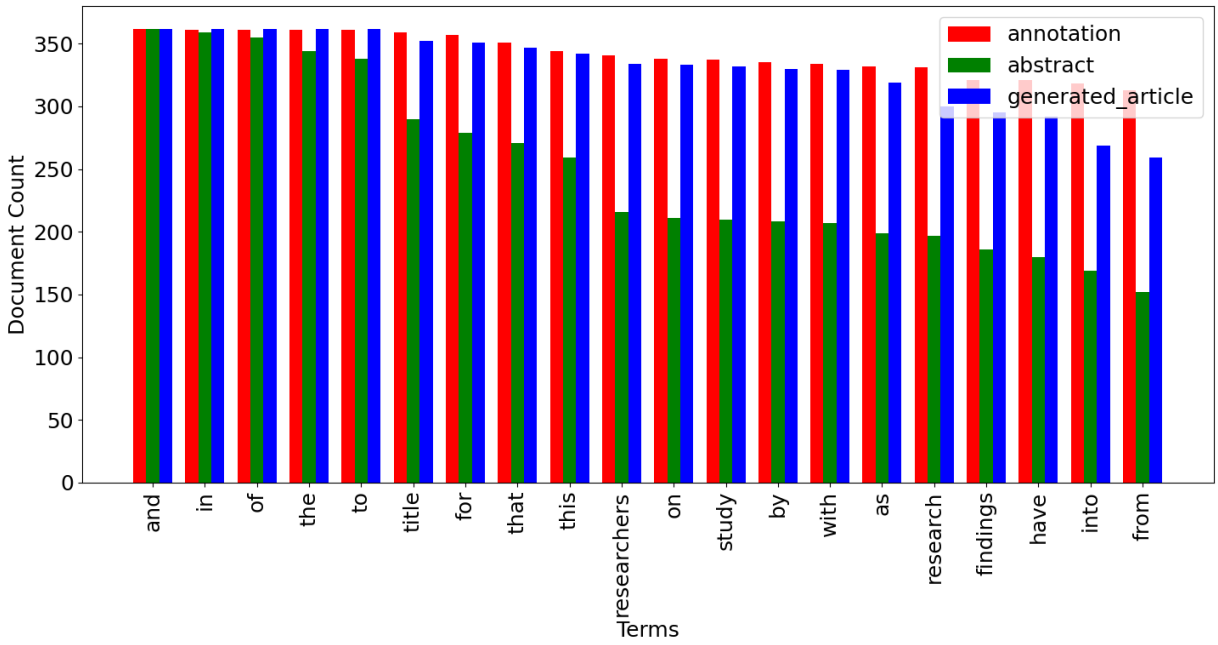

Figure 5. Document Frequency Distribution for *annotation*, *abstract*, and *generated article*.

The DF distribution shown in Figure 5 follows a similar pattern as the TF distribution, with *annotation* having the highest DF for most terms followed by *generated_article* and *abstract*. This indicates that these common terms are not only frequent within documents but are also spread across a majority of documents in the corpus.

CHAPTER 4

METHODOLOGY

4.1 STANDARD EVALUATION METRICS

To comprehensively measure the efficiency of the news article generation, we employed the following evaluation metrics: ROUGE-1, ROUGE-2, ROUGE-L, BlEU, BERTScore, and METEOR. These metrics were chosen because they collectively capture the efficiency of each method.

Recall-Oriented Understudy for Gisting Evaluation or ROUGE is a system that employs the automatic evaluation of the quality of a summary by contrasting it with human-generated ideal summaries. The ideal summary in our study is the annotated ScienceAlert news articles that are relevant to the research paper abstract. ROUGE-1, ROUGE-2, and ROUGE-L primarily measure the generated article and human-edited article quality, capturing unigram, bigram, and longest common subsequences, respectively [20]. The ROGUE n-gram is calculated using the following equation:

$$\text{ROUGE\_N} = \frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{1}$$

where S represents the reference summaries and $\text{Count}_{\text{match}}$ refers to the highest count of n-grams that appear in both the candidate and the set of reference summaries [20].

To evaluate the quality of the human-edited news article or generated news articles, we assume that news articles written by news editors have the highest quality, and use them as the ground truth.

The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of machine-generated translations [27]. It compares the n-grams in the translated text to those in the reference text to quantify the closeness. The precision is then modified by giving it a brevity penalty to account for a text shorter than the reference text. BLEU cores range from 0 to 1, with 1 being a perfect match. METEOR overcomes the lack of recall and it uses unigram matching to outperform BLEU in many languages [4]. However, these metrics do not consider the semantics of text.

$$\text{METEOR} = \left(\frac{10 \cdot P \cdot R}{9P + R}\right) \cdot \text{Penalty} \tag{2}$$

where $P = \frac{\text{Numberofmappedunigrams}}{\text{Totalnumberofunigramsincandidate}}$, $R = \frac{\text{Numberofmappedunigrams}}{\text{Totalnumberofunigramsinreference}}$ and

$$\text{Penalty} = \left(1 - 0.5 \cdot \left(\frac{\#\text{chunks}}{\#\text{matchedunigrams}}\right)^3\right)$$

METEOR penalizes the score for continuous sequences of matched tokens relative to the total number of unigrams present in the reference text [3]. BERTScore calculates a sentence-level similarity score between the machine translation and the referenced text [42]. This evaluation metric is widely used in evaluating text generation tasks where the pre-trained BERT considers contextual and semantic word embeddings of each token in the GPT-generated news with each token in the ground truth. These metrics focus on token-level evaluation of texts, rather than sentences and paragraphs, which is what news text consists of. More robust evaluations like LLMs are required that are more aligned with human evaluations.

4.2 LARGE LANGUAGE MODELS

Large language models (LLMs) are advanced artificial intelligence systems designed to under-

stand, generate, summarize, and predict content. These models have been trained on large-scale datasets of diverse types. They are considered to be large since these models can be hundreds of gigabytes in size [7] as the number of parameters ranging from billions to hundreds of billions. Due to their ability to process and analyze large volumes of text, LLMs have significant potential in evaluating scientific news articles. The capabilities would potentially help users or laypersons read and understand the often complex landscape of scientific papers. We experimented with both commercial and open-weight LLMs using multiple levels. We explore the current state-of-the-art models and test their effectiveness on our data.

4.2.1 Commercial Models

Most state-of-the-art LLMs are commercial and are often closed-source because they offer economic value and competitive advantages to businesses. Training LLMs requires extensive investment in computational resources, which may cost over millions of dollars as the model's size exceeds a billion [33].

State-of-the-art Models

GPT, or Generative Pre-trained Transformers is a family of large language models trained by OpenAI. They have gained popularity because of their increasing ability to generate and evaluate text. Their most recent model is GPT-4o released on May 13, 2024. Claude-3.5 Sonnet by Antrhopic rolled out on June 20, 2024, outperforming all previous LLMs. Claude-3 Opus was shown to beat GPT-4 in ten capability benchmarks [2]. We explored three different models by OpenAI: GPT-4o, GPT-4, and GPT-3.5.

4.2.2 Open-Weight LLMs

Open-weight LLMs have gained popularity in the research community for their free cost and their increasing capabilities for many NLP tasks. They are viewed as a good alternative and cost-efficient way of testing the capabilities of different LLMs. The open-weight LLMs mean that they are open for everyone to download the weights of the neural network and use. The downside is that the users need to be more familiar with theoretical concepts such as neural networks and model inference. In this study, we used state-of-the-art models from Meta and MistralAI. We explore the feasibility of using open-weight LLMs. With some knowledge of their structure, anyone can potentially use them for NLP tasks.

As of the time of this writing, the most advanced open-weight models are LLaMA-3 with 8 billion and 70 billion parameter models. They also offer an instruction model that can be tuned to the user's preferences. Their sensitivity to prompt engineering shows their potential in using them over commercial models for a variety of tasks. We also experimented with MistralAI's newest model [18]. Mistral was chosen for their relatively small number of parameters and their efficiency in many NLP tasks [18].

4.3 PROMPT ENGINEERING

Prompt engineering is a technique used to guide the responses produced by LLMs. This is done by carefully crafting the input queries. By formulating prompts in a specific way, developers can influence the model's output to reduce or eliminate inherent biases present in them. This is of great importance especially when evaluating scientific news, as biases in the LLM toward other LLM-generated content can distort the interpretation and evaluation of the scientific news text.

Through the use of prompt engineering, it is feasible to ensure a balanced evaluation of news text. LLMs can provide a more reliable and unbiased evaluation of scientific news text, enhancing the quality of information accessible to the general public. To demonstrate the significance of prompt engineering, we performed an experiment in which we generated two different batches of news articles given a scientific paper abstract. Figure 6 illustrates the effectiveness of simply instructing the model to control the number of words. We can also see that the means of the two distributions are much closer to each other.

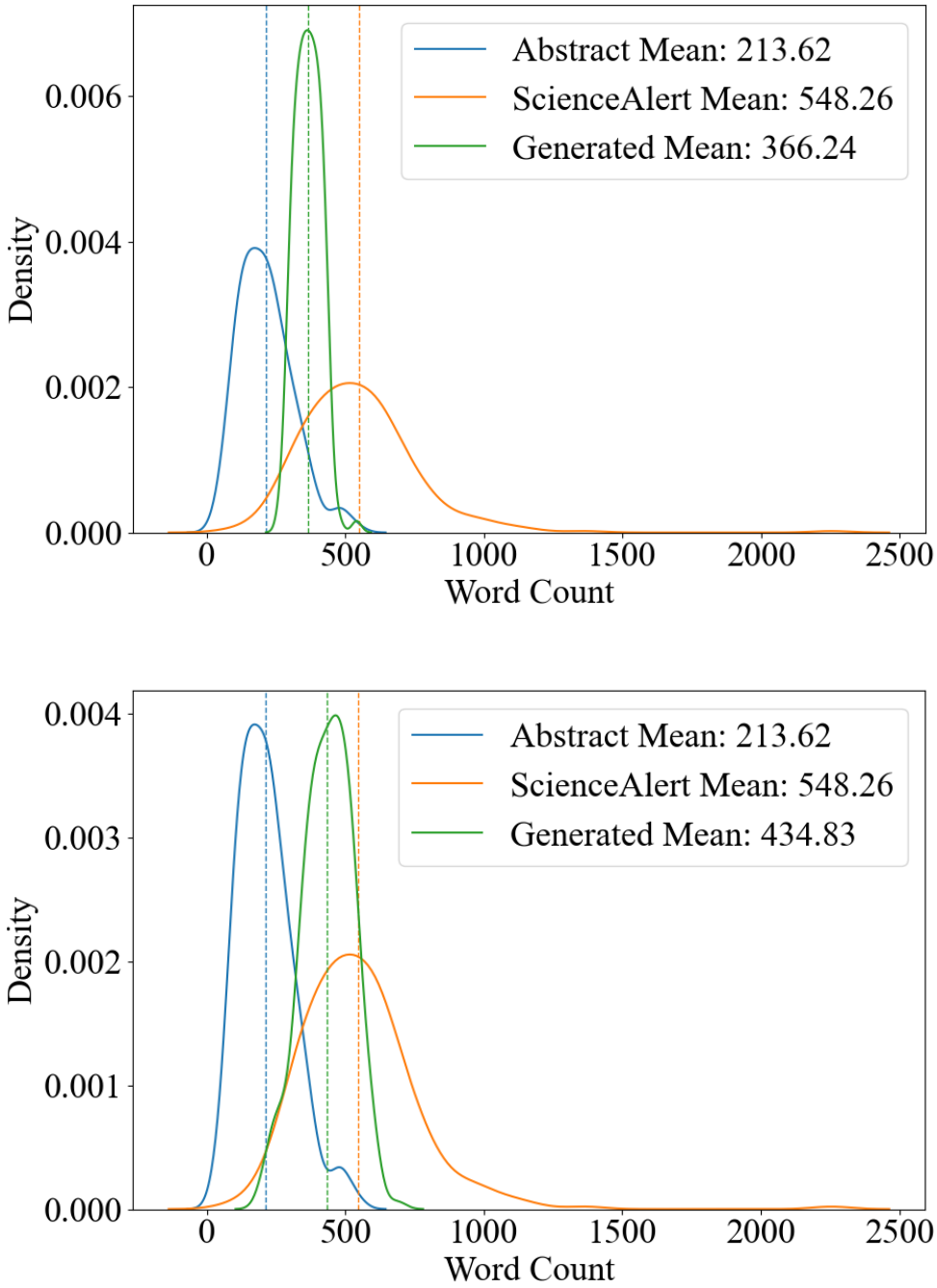

Figure 6. Word Count Distributions of the annotated ScienceAlert news article and the AI-generated article before (top) and after (bottom) instructing LLM to control the number of generated words.

To demonstrate this, we performed two experiments, where we generated multiple news articles using two different prompts. One prompt included a directive for the model to limit the length of the articles. Note that GPT did not always generate the exact number requested in the prompt. However, the average of the word count distribution changed significantly. In Figure 6, the top plot shows the distribution of the article lengths generated without any word limit, resulting in a wider range of article lengths with a mean of 366.24. In contrast, the bottom plot illustrates the distribution when a specific word limit is instructed, leading to a more consistent output with the human-edited news text. We can see that even a small change, can result in a big difference in terms of the distribution of the generated articles. These plots show the impact of prompt engineering in controlling the output of language models, making it a powerful tool for various applications in content generation and evaluation, where adhering to specific guidelines is essential. In the following section, we introduce the main methodologies developed throughout this work.

4.4 MAIN SETTINGS

By leveraging the capabilities of LLM, we can create methods that can improve the model's ability to assess the readability, comprehensiveness, and accuracy of scientific information given in an abstract of a scientific paper. With some guidance, they may also be able to identify if a candidate text was written by another LLM or a human. There are two main settings used in our experiments:

(a) direct scoring

(b) pairwise comparison

4.4.1 Setting (a): Direct Scoring

In this setting, the texts are evaluated by a score system based on different criteria. Previous work focused on directly using GPT as an evaluator of the summarization task they break the overall score into several dimensions, but they did not investigate the rubrics. The methodologies we have used in the study are the following:

a.1 **Single Score**

In this setting, we prompt GPT-4 to evaluate both human-written news articles and LLM-generated articles on a scale of 0-10

a.2 **Component Scores**

Here, we model this comprehensive score as a combination of three component scores: comprehensiveness, accuracy, and readability.

- **Comprehensiveness** measures the coverage of the information that is presented in the ground truth news.
- **Accuracy** measures the correctness of information in generated news compared to the ground truth news.
- **Readility** measures how easy to read the generated news is compared to the ground truth news.

a.3 **Rubric-based Component Scores**

Using this setting, we further improve the component score method by providing rubrics for each component. That is the definition of each component and certain points to deduct for linguistic features

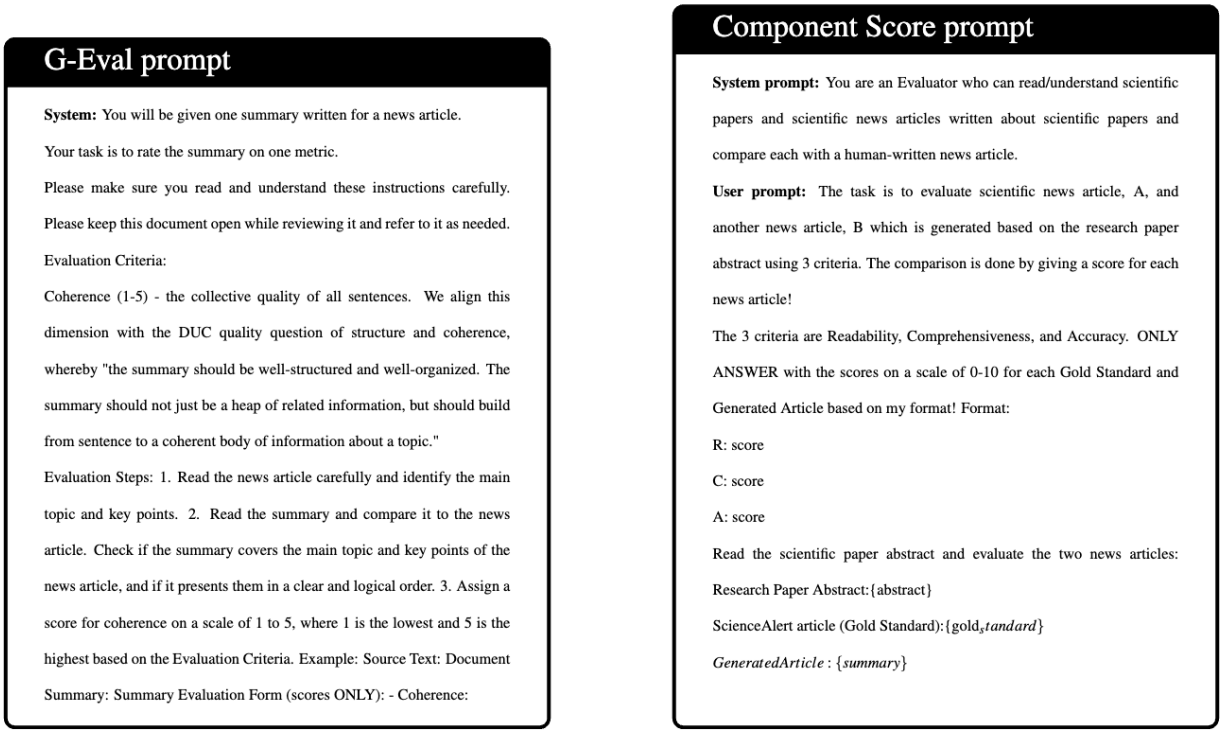

Figure 7. Example prompts comparing G-Eval and Component Score method

The baseline approach (a.1) is to directly get a single score for each article, while the component scores method (a.2) breaks down the evaluation into multiple criteria such as comprehensiveness, readability, and accuracy. G-Eval is a similar example of direct scoring where they use four dimensions, namely coherence, consistency, fluency, and relevance. However, they base their scoring on a smaller scale (0-5), while we use a more fine granular scale (0-10) for each component score. Figure 7 demonstrates the comparison between G-Evals sample prompts to the prompt we use in (a.2).

Our three dimensions, Readability, Comprehensiveness, and Accuracy, make our scoring space more compact while also guiding the model to the fine granular scale. In addition, the prompt is

enriched with the rubrics in (a.3) for each component score to provide a clear explanation for each criterion. In addition to the rubrics, a penalty is also given to a component score for linguistic features that decrease the quality of the article such as repetition of words, or hallucinations of information that is not present in the given abstract.

4.4.2 Setting (B): Pair-Wise Comparison

The pair-wise comparison is a method where we compare a human-written news article with LLM-generated news articles based on a research paper abstract. This section presents the main contribution of this work, which is the prompting techniques used throughout all experiments. Existing methodology in terms of pairwise comparison has been done. The baseline method is the PairEval for pairwise comparison for dialogue-based data for chat models [28]. They used one LLaMA 7 billion model for this task. No further model exploration was done or different types of data such as news text. LLM Comparative Assessment [23] is another method with a zero-shot pairwise approach. We have one-shot, two-shot, and three-shot prompts with guidance to the LLM of the characteristics of text to be evaluated. Ours is a more general approach using multiple examples.

We propose more general approaches where we try to reprogram the built-in rubrics of an LLM by giving it sophisticated prompts such as guiding the model using the characteristics of text and/or providing the LLM one or several examples of both human-edited and LLM-generated texts. We propose four novel pairwise comparison approaches to mitigate the bias of LLMs toward LLM-generated content.

b.1 **Direct Comparison**

The vanilla approach directly compares the human-written human-annotated articles with the LLM-generated content.

b.2 **Guided Zero-shot Comparison**

We improve this technique by guiding the model by giving it the characteristics of both LLM-written news and human-written news.

b.3 **Few-shot Comparison**

We explore the potential of giving models examples since LLMs are few-shot learners [7].

b.4 **Guided Few-shot Comparison**

We combine the characteristics and examples within the prompt to create the method we call *guided few-shot pairwise comparison*, which provides examples to the model first and then guides about the characteristics of the text. We noticed that order makes a difference.

The first method (b.1) provides a baseline method that has been explored. Upon requesting the LLMs the internal rubrics, we noticed that it is looking for the wrong characteristics, which is leading the model to incorrectly identify the characteristics of human-written news. This finding led us to include guides in the prompt to help the model correctly identify the style of text a human would write. Then, we improve this method in (b.4) by giving the model examples. By giving the model one, two, or three examples of both types of text, they can look for patterns and correctly identify which article was written by humans. However, the more examples we give, the longer the prompt will be, which may result in a loss of attention to the initial description of the task. In the following section, we extensively test the performance of both settings (a) and (b) using various experiments and LLMs.

CHAPTER 5

EXPERIMENTS

We test three hypotheses to investigate the feasibility of using GPT-4, as an evaluator for *Scientific News Generation*. We explored various methods within each setting to determine their effectiveness in identifying the source of the articles. We test the following hypotheses to investigate whether assess whether direct scoring or pairwise are suitable evaluators to discern the source of news articles.

Hypothesis 1 (H1) *The scores given by GPT-4 are strongly correlated with scores calculated using standard text summarization evaluation metrics.*

Hypothesis 2 (H2) *Direct scoring is a reliable method to evaluate to discern human-written and GPT-generated news articles.*

Hypothesis 3 (H3) *Pairwise comparisons provide a more consistent and reliable way of evaluating the quality of news text.*

5.1 EXPERIMENT 1: TESTING HYPOTHESIS 1

The standard automatic evaluation metrics are based on token-level overlaps between the generated text and the ground truth. However, a comprehensive evaluation by considering multiple factors beyond token-level representations is beyond the capability of these metrics. Ideally, humans should be involved as evaluators of these metrics. However, human evaluations are time-consuming and financially expensive. In addition, evaluating the quality of scientific news articles

usually requires domain experts, which is challenging to do reliably on common crowdsourcing platforms, such as Amazon Mechanical Turk. Therefore, here, we investigate whether it is possible to use direct scoring to automatically evaluate the text quality.

To test Hypothesis 1, GPT-4 was prompted to generate the overall scores for a given GPT-3.5-generated news article. We then calculated the traditional evaluation metrics and the Kendall correlation coefficients between traditional scores and GPT-4 generated scores. The heat maps in Figure 8 do not exhibit strong correlations between GPT4-generated scores and traditional metrics, which ruled out the null hypothesis and further justified using GPT-4 scores independent of traditional text summarization metrics.

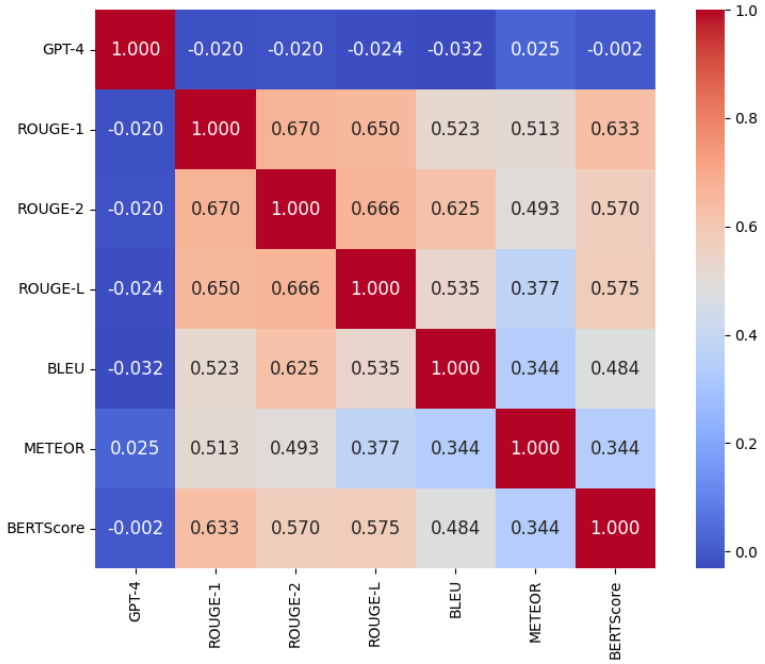

Figure 8. Kendall correlation coefficients heatmap between GPT-4 generated scores and traditional text summarization metrics.

5.2 EXPERIMENT 2: TESTING HYPOTHESIS 2

This experiment tries to show that direct scoring is not consistent and new methods are required to improve the quality of Automated Evaluators. The main goal is to refute Hypothesis 2. We have two main baseline methods, one is G-Eval for direct scoring and the other is direct pairwise comparison. The direct pairwise simply leverages the built-in knowledge and reasoning capabilities of an LLM. We want to justify that the scoring method does not work because GPT-4 gives inconsistent results, limited discrete scores, and makes arithmetic mistakes when calculating scores. The

built-in rubrics that LLMs use may differ from the rubrics set by humans. Humans must explicitly instruct LLMs on the desired rubrics to make them work better. As we know before, LLMs are exceptional few-shot learners so providing examples of characteristics will enhance their performance on this task [7].

In this experiment, we randomly selected 11 samples to test direct scoring for evaluating news articles. We tried three different methods to generate the news articles and control the temperature of the model. Then, we evaluated the quality of these news articles using GPT-4 with the component score method only.

5.2.1 Results of Direct Scoring

While this method is straightforward, it has some inherent limitations. For instance, scoring often fails to account for the relative importance of criteria or the interdependencies between items. The reason they fail is that the scores generated by GPT-4 are not diverse. From Table 2, we can observe that the scores are in a narrow range of $8.5, 9, 9.5$ with 70 instances of 8.5 score out of 121. One way is to let GPT figure out what each of these metrics means. Another would be to give definitions and rubrics to deduct points.

We explored several methods to generate different types of news articles to try to move the evaluation in a certain direction, but GPT-4 generated very repetitive scores. These methods were developed using empirical testing. The *random* method generates *N* stories given a scientific paper abstract and then prompts GPT to summarize it as if it were a journalist. We tried generating using a variety of temperatures in an attempt to cover the entire probability distribution. The *iterative* method knows all previously generated articles, and we ask GPT to generate another one that is unique from the rest. The *boost* method is built on the *iterative* method by trying to

improve the quality of articles at each iteration. The news articles are generated using GPT-3.5 and then evaluated using GPT-4. The current state-of-the-art LLMs cannot evaluate news articles by generating scores.

Table 2. Direct Scores generated by GPT-4 demonstrating low variability.

Setting	temp	stories	**37**	**404**	**435**	**515**	**530**	**936**	**1403**	**1442**	**1590**	**1606**	**1653**	avg	median
random	0.7	3	8.5	8.5	9.5	8.5	9.5	8.5	9	9	8.5	9	8.5	8.818	8.5
iterative	0.7	3	8.5	8.5	9.5	8.5	9.5	8.5	9	9.5	8.5	9.5	8.5	**8.909**	8.5
iterative	0.7	9	8.5	8.5	9	9	9	8.5	8.5	8.5	8.5	9.5	8.5	8.727	8.5
boost	0.7	3	8.5	8.5	9	9.5	8.5	8.5	8.5	9	8.5	9.5	8.5	8.773	8.5
boost	0.7	9	8.5	8.5	9.5	8.5	9	8.5	9	9.5	8.5	9	8.5	8.818	8.5
iterative	0.3	3	8.5	8.5	9.5	9	8.5	8.5	9.5	8.5	9	8.5	8.5	8.773	8.5
iterative	0.3	9	9	8.5	8.5	9	8.5	8.5	9	9	8.5	9.5	8.5	8.773	8.75
boost	0.3	3	8.5	8.5	9.5	8.5	9.5	8.5	9	9	8.5	9	8.5	8.818	8.5
boost	0.3	9	8.5	8.5	9.5	8.5	9.5	8.5	9	9.5	8.5	9.5	8.5	**8.909**	8.5
random	0.3	3	8.5	8.5	9	9	8.5	8.5	9	8.5	8.5	9	8.5	8.682	8.5
random	0.3	9	8.5	8.5	9	8.5	9	9	9	8.5	8.5	9	8.5	8.727	8.5
GroundTruth			9.0	9.0	8.83	7.3	9.0	8.0	9.0	9..0	6.3	9.0	7.8	8.385	8.4

We investigated different settings for generating articles. The scores are highlighted in Table 2 demonstrates the low variability in the scores generated by GPT-4. The scoring results have revealed some fundamental issues that may show the direct scoring methods do not provide optimal results. These are obtained not from human study. By doing the automated evaluation on 11 sam-

ples. Table 2 also shows why scoring methods are not reliable in evaluating the quality of news text. The yellow volume highlighted represents 70 instances when GPT-4 rated the generated article 8.5 which yields 57% of the total ratings. Our observation is consistent with previous works.

In the rubric-based method, we ask for the rubrics back in the prompt to ensure that GPT-4 followed the given instructions. Then, we also query the term frequency distribution to ensure GPT can count the number of terms as it is directly related to the Readability component score. Figure 9 shows that GPT-4 is unable to deduct points from a 10-point scale. For example, *"Therefore, 2.5 points are deducted, resulting in a score of 9.5"* is not correct. GPT fails to give us the correct term frequency because it might not be able to In some cases, GPT cannot correctly count the number of words. For example, the hallucination of "Dr. John Smith" appears twice, but GPT-4 thought it appeared 5 times which resulted in a penalization of 4 points, "giving a score of 7.0". The word "research", highlighted in light blue in Figure 9, appears 9 times.

We can observe from Figure 9, Figure 10 and Table 2 that direct scoring has the following disadvantages:

1. GPT-4 rated scores are not very diverse. Most scores are concentrated on N discrete values.
2. GPT-4 rated scores favor GPT-generated news articles.
3. GPT-4 does not seem to generate arithmetically correct scores under the given scales.
4. GPT-4 does not appear to have the capability to correctly count the n-gram frequency and use it as a condition to rate scores.

For these reasons, we reject Hypothesis 2 because direct scoring is not a reliable method to evaluate news quality and is thus used for discerning human-written and LLM-generated news articles. In

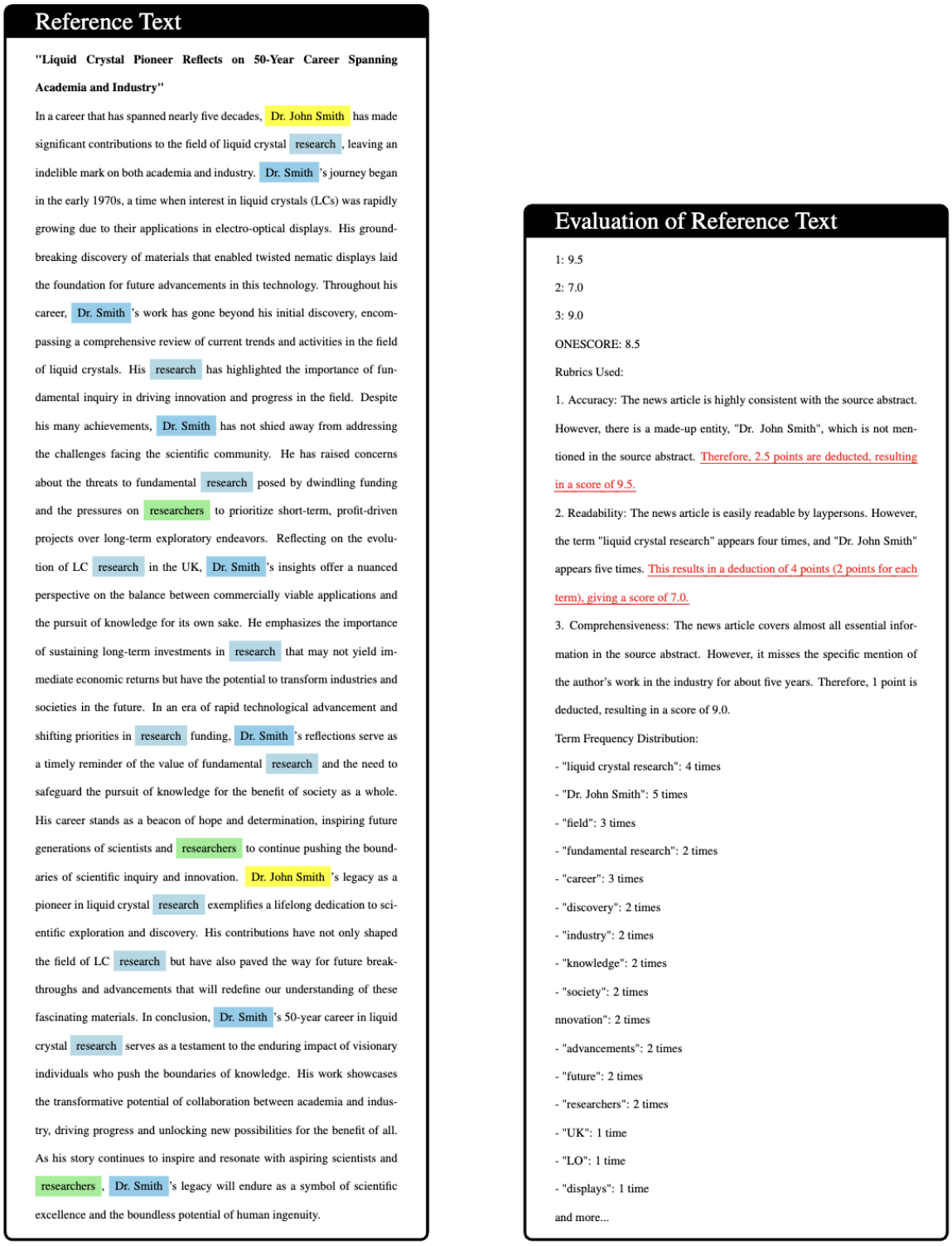

Figure 9. Machine-generated text and its Evaluation using GPT-4 demonstrating arithmetic mistakes and term frequency distribution

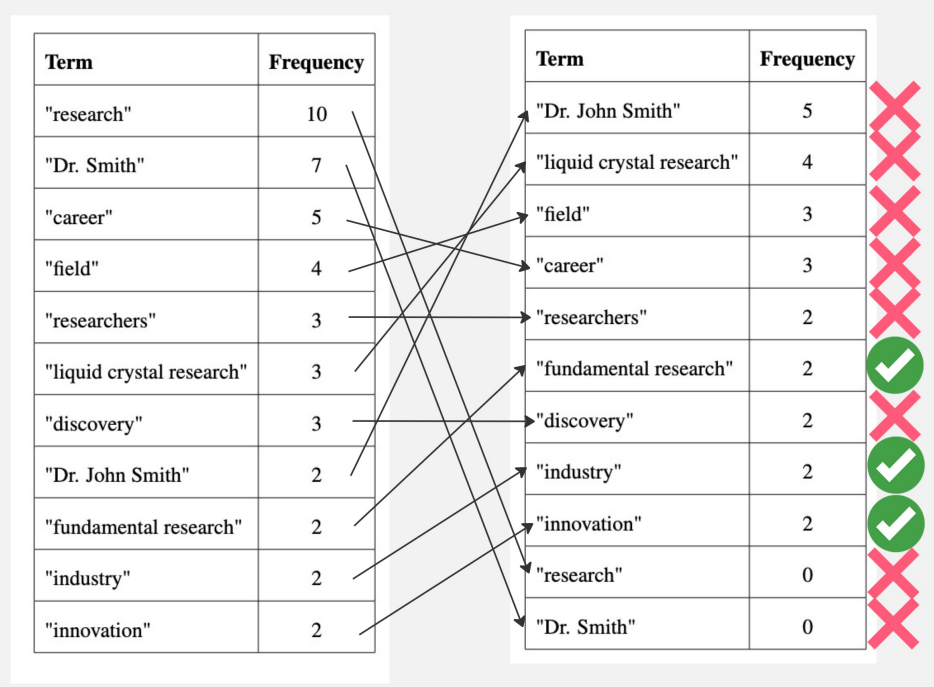

Figure 10. Sorted Correct Term Frequency based on text in Figure 9 compared with sorted LLM-generated Term Frequency

the following sections, we investigate the feasibility of using pairwise comparison as an alternate method of evaluating news articles.

5.2.2 Pair-Wise Comparison

In the previous section, we show the existing problems with direct scoring and the limitations of direct pairwise comparison. In this section, we seek to assess which method of evaluation brings more value and mitigates more bias in the evaluation of scientific news text. For this experiment, we have randomly selected 11 examples. Pair-wise comparison is a method where items are compared against each other in pairs. This method considers the relative performance of items, making it more robust against the limitations of simple scoring. We tested the baseline method (b.1), guided

pairwise (b.2), one-shot pairwise (b.3), and guided one-shot (b.4) methods where the accuracy was calculated by:

$$Accuracy = \frac{\#\ correctly\ discerned\ samples}{\#\ samples}$$

5.2.3 Comparison Between Scoring and Pair-Wise Comparison

Using scores is rigid and can be meaningless, but pairwise comparison methods do not look at exact scores. LLMs' preference for LLM-generated content can be encountered using a new prompting style to correctly discern which one is human-written or LLM-generated. It is a qualitative comparison, therefore there are no rubrics, only guides to the model. LLaMA-3 has a low accuracy score in the direct pairwise comparison, but we can fix that. Providing guides to the model increased performance. One-shot pairwise performed similarly to guided pairwise, getting two more correct than its predecessor. Providing guidance and one example improved the model's performance to a perfect score of 100%.

Table 3. LLaMA-3 Pair-Wise Comparison Results

LLaMA-3 8b	Correct	Wrong	Accuracy
direct pairwise	3	8	27%
guided pairwise	5	6	45%
one-shot pairwise	7	4	63%
guided one-shot pairwise	11	0	100%

5.3 EXPERIMENT 3: HUMAN STUDY

The main goal was to test whether discerning human-written and LLM-generated news is a hard task for humans. In this experiment, we asked three graduate students completing their PhD in Computer Science to evaluate the human-written news and LLM-generated news text. We asked them to determine which one was written by humans and the reason/rubric they used to make those decisions. Figure 11 illustrates a sample questionnaire used throughout the human study.

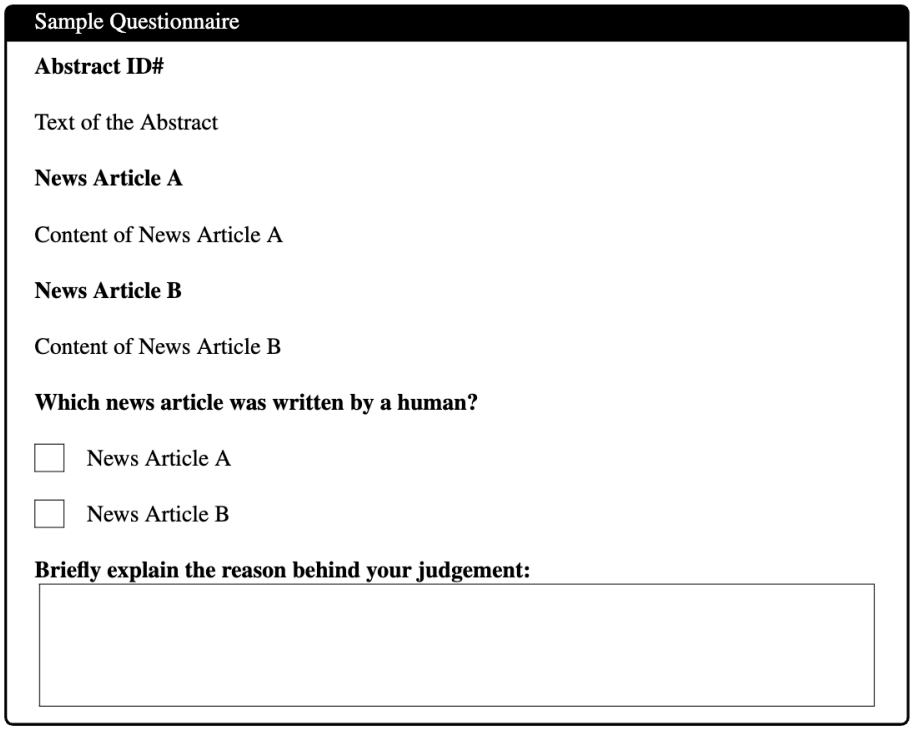

Figure 11. Sample Questionnaire Provided to Participants in Experiment 3

We compared different labelers to see if they used consistent reasoning for their decisions. Evaluation criteria reported by the labelers included the presence or absence of additional content

provided by human-written articles, as human writers may include additional information that is not in the abstract in the news. Instead, LLMs generate news articles only based on the abstracts. Although LLMs may generate additional information beyond the abstract, we observed that the information was simply hallucinations that are inconsistent with the abstracts.

The experiment results indicate that it is relatively straightforward for humans to discern human-written and LLM-generated news while the task exhibits various levels of difficulty for the LLMs we tested.

When comparing the three labelers, we found that their evaluations shared common criteria for readability and context. We then use GPT-4 to conduct the same experiments. We let GPT do the same to see if it can reach human performance It is a limitation of the current state-of-the-art LLMs. To prove that a human's ability to evaluate is superior to LLMs. We compare different labelers if they use similar rubrics to evaluate the quality of news text. Some of their conclusion is based on the same reason where the human-written article provided additional context. When they decide if an article is readable or not, they find that human-written articles are better written. When we look at a particular article, the evaluators have the same criteria.

5.3.1 Compare Pairwise Comparison with Human Study

For this experiment, 11 testing samples were used and shuffled when giving them to the model. Each prompt sample contained a triplet, the scientific paper abstract, the human-edited news article, and GPT-3.5 generated news text. For the entirety of this study, the examples were shuffled and only the labelers were not exposed to the ground truth. Our experiment aimed to determine whether humans or LLMs can better discern human-written to LLM-generated news. The results indicate that while humans outperformed the current open-weight state-of-the-art model, their performance

can be improved to the human level by using guidance and an example. Larger-scale experiments are needed to provide further proof of concept for our method work.

Table 4. Comparison of Human, GPT-4, and LLaMA-3 across different settings.

Settings	Human	GPT-4	LLaMA-3 8b
direct pairwise	**97%** (macro avg)	36%	27%
guided pairwise	NA	**54%**	45%
one-shot pairwise	NA	**100%**	54%
guided one-shot pairwise	NA	**100%**	**100%**

Using the trivial prompt, direct comparison (b.1) achieved only 36% accuracy for GPT-4, indicating its ineffectiveness. Providing guides (b.2) improved accuracy to 54%, showing the value of structured assistance. Providing an example to the model (b.3) and combining guides with examples (b.4) resulted in 100% accuracy, highlighting the critical role of comprehensive guidance and illustrative examples in enhancing the performance of the evaluator LLM. Although the sample is relatively small, the articles used cover different topics and have various lengths, so it is a reasonable representative of the sample. We demonstrated that pairwise comparison is superior to direct scoring and that pairwise comparison can be used as a prompting technique to improve the quality of the evaluation.

5.3.2 Discussion

Identifying human-written news from GPT-generated news is a relatively easy task for humans with higher education backgrounds. The same task is challenging for state-of-the-art LLMs by direct pairwise comparison, but the performance could be improved at the human level by providing guidance and one-shot learning. However, we improved the capabilities of GPT-4 and LLaMA-3 in distinguishing between AI-generated content and human-written articles. Scoring methods provide quantitative data useful for detailed analysis but can be subjective, especially without clear definitions. In contrast, pairwise comparison is more intuitive. The simplicity of pair-wise comparison makes it a preferred method for evaluations. The human study demonstrated that they can better distinguish between human-written articles and LLM-generated articles. LLMs have shown varying success rates using different pairwise methods. The results from the human study clearly show that humans are better at direct pairwise asking (b.1).

5.4 EXPERIMENT 4: SAMPLE SELECTION STUDY

Sample Selection study focuses on the impact of examples selection inside the prompt. This experiment focuses on the impact of example selection on the fine-tuning process. It also aims to further test GPT-4's ability to differentiate between human and LLM-generated content. We investigated what examples to use to fine-tune an LLM for two different settings using 11 testing samples. Each sample contains a triplet of research paper abstract, human-edited news text, and GPT-generated news text. The first setting is the one-shot prompting method, then an improved version of it where we combine providing the characteristics of human-written articles and AI-generated content.

5.4.1 Discussion

Table 5. Comparison of GPT-4 and LLaMA-3 across different news domains and settings.

Settings	News domain	GPT-4	LLaMA-3
One-shot pairwise	Uncategorized	100%	54%
	Uncategorized	100%	54%
	Nature	100%	63%
	Humans	100%	54%
	Space	100%	54%
Guided one-shot pairwise	Uncategorized	100%	81%
	Uncategorized	100%	100%
	Nature	100%	81%
	Humans	100%	81%
	Space	100%	63%

Table 5 demonstrates the potential of combining the few-shot method with the characteristics. We may need to do further experiments to reveal more patterns of the performance difference of such settings. The performance of LLMs to recognize human-written scientific news is very little dependency on the domains of the examples in the one-shot pairwise prompting. Adding the guidance improves the one-shot learning performance by 25% on average. We need to do further experimenting with a larger sample size to discover patterns in the performance of these prompt engineering methods. Let us first discover which domains we would like to use for the one-shot, two-shot, and three-shot settings.

5.5 EXPERIMENT 5: DOMAIN DEPENDENCY STUDY

We split 360 samples into 60 training and 300 testing samples to investigate the domain dependency of testing performance. We selected 3 domains, each containing at least 3 samples in the training data. Then for each domain, we conducted n-shot experiments in two settings:

- Setting #1: training using 3 in-domain samples.
- Setting #2: training using 3 out-domain samples.

Then, we selected 3 domains, denoted as D_1, D_2, and D_3, each containing at least 3 samples in the training set. We tested using different settings in terms of the number of examples given to the model for each domain using the test set and also out of domain that is not D_1, D_2, and D_3. For this experiment, LLaMA-3-8b and LLaMA-3-70b were used to determine which domain is the best. It may also reveal further performance improvements across methods.

5.5.1 Unguided Few-Shot Results

We wanted to see what type of examples to use and how many shots would result in the best performance. From Table 6 shows the performance of using different examples

In Table 6 for two-shot evaluation the out of domain won that round. Both examples' domains came out of the uncategorized categories. Such examples relate to ones from 2014 before they started labeling their news article's categories. The results in Table 6 show that providing more examples to the LLM improves their accuracy. The use of unguided few-shot showed its potential

Table 6. Unguided Few-shot pairwise comparison using LLaMA-3

	LLaMA-3 8b			LLaMA-3-70b		
Domain	One-shot	Two-shot	Three-shot	One-shot	Two-shot	Three-shot
$D_1 = nature$	37.54%	29.24%	**34.22%**	57.81%	77.74%	85.38%
$D_2 = tech$	**40.20%**	28.57%	25.91%	71.43%	77.08%	**85.71%**
$D_3 = health$	29.57%	**36.88%**	23.26%	**72.09%**	79.07%	75.75%
Out	39.87%	33.89%	32.56%	66.55%	**80.40%**	81.40%

5.5.2 Guided Few-Shot Results

For the one-shot setting in Table 7, the out-of-domain was uncategorized. Two-shot had two random examples from the environment domains as out-of-domain outperforming the rest of the domains in this category. For the three-shot setting in row out, there were two uncategorized, and one environment example triplet was given to the model. This study revealed performance comparison between the use of both shots and guides produces better results than either one alone. It seems for the smaller LLaMA-3 model, increasing the number of shots would reduce the accuracy due to the model diverging from the original instructions. The more shots it is given, the less it remembers the initial instructions. It makes sense, as the context window increases, it may forget previous instructions given at the beginning of the prompt. Larger models can remember more.

Table 7. Guided Few-shot pairwise comparison using LLaMA-3

	LLaMA-3 8b			LLaMA-3-70b		
Domain	One-shot	Two-shot	Three-shot	One-shot	Two-shot	Three-shot
$D_1 = nature$	44.52%	33.89%	55.48%	97.01%	88.70%	89.70%
$D_2 = tech$	**54.15%**	56.81%	57.48%	97.67%	86.36%	**90.70%**
$D_3 = health$	46.84%	54.82%	**60.13%**	95.35%	84.39%	89.70%
Out	47.18%	**67.11%**	51.83%	**98.34%**	**96.01%**	**90.70%**

5.5.3 Discussion

The unguided few-shot pairwise comparison results show that providing more examples generally increases accuracy. The guided few-shot results show that adding guidance improves performance significantly across all domains. This experiment highlighted the use of domain-dependence-specific training and further proved the effectiveness of combining characteristics with examples. These findings provide a proof of concept that guided methods and careful selection of examples can significantly improve open-weight LLMs accuracy. The pattern in the experimental results further shows that LLMs are few-shot learners. By also providing guides to the model, we can maximize the capabilities of current LLMs surpassing human performances in the guided one-shot setting.

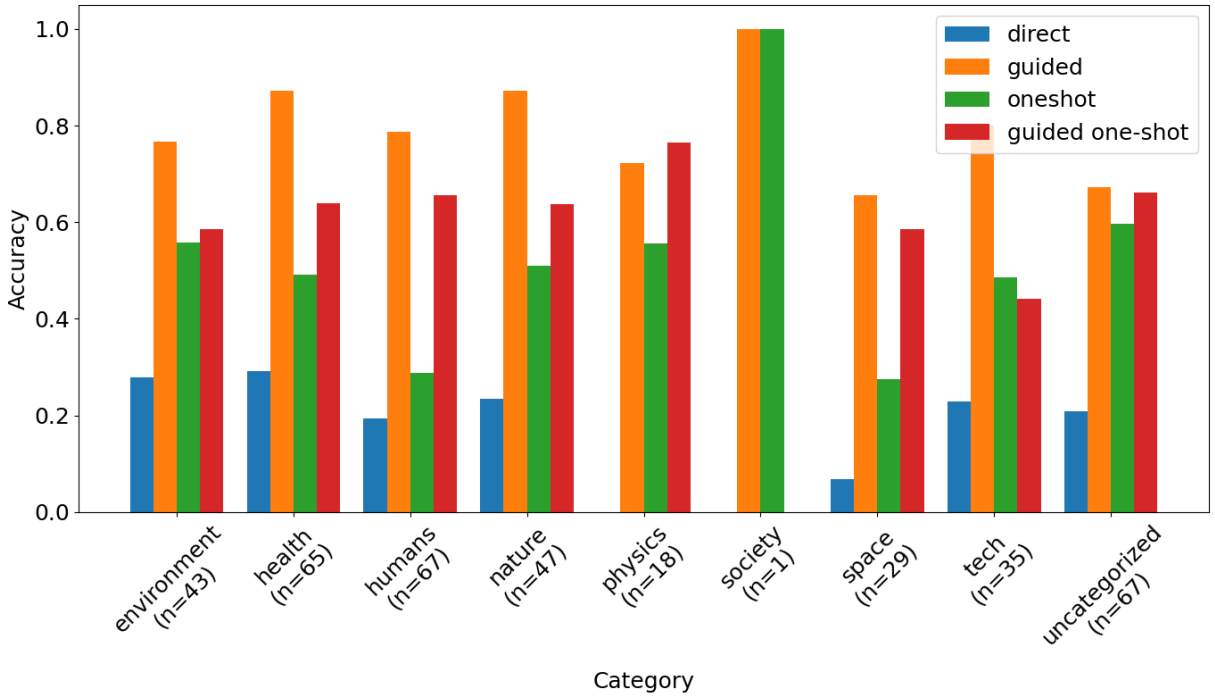

Figure 12. Accuracy in Different Domains across Pairwise Comparison Methods

Figure 12 shows the accuracies across domains and our pairwise comparison methods. The accuracy may vary greatly across methods. For example, the "environment" category and "humans" have high variability in terms of accuracy. Further experimenting is necessary using a collection of LLMs to explore the potential of domains and our methods.

5.6 EXPERIMENT 6: LLM EVALUATION ON SANEWS

We performed an automated pair-wise evaluation on the SANews dataset using the most advanced LLMs such as GPT-4o and Claude-3.5 Sonnet. The methods tested include direct pairwise, guided pairwise, few-shot pairwise, and guided few-shot pairwise comparison. This experiment presents the results for the main contributions of this work, which are the prompting techniques

used in this experiment. We explore different levels of LLMs, using both commercial and open-weight LLMs. Table 8 shows the varying performances of our proposed pairwise comparison using different prompt engineering techniques.

Table 8. Aggregated comparison of different settings across models with tier differentiation

Model	Direct	Guided	One-shot	Guided one-shot
GPT-4o	28.48%	**92.72%**	76.92%	92.31%
Claude-3.5 Sonnet	21.19%	67.63%	**99.32%**	**100%**
GPT-4	23.32%	52.32%	69.87%	93.05%
Claude-3 Opus	14.57%	82.00% (100)	62.00% (100)	**100%** (100)
LLaMA-3 70b	40.38%	79.80%	72.09%	98.34%
GPT-3.5 20b	7.95%	31.46%	27.48%	28.48%
LLaMA-3 8b	21.24%	77.78%	47.15%	60.72%
Mistral 7b	**50.41%**	48.24%	48.90%	49.17%

Most of the LLMs in the direct comparison perform poorly in this setting as they fail to correctly identify the human-written news articles. The results in Table 8 demonstrate the challenges direct comparison proposes without providing guidance and examples to the model. In the guided pairwise comparison setting, GPT-4o achieves the highest accuracy with 92.72%. Due to the costs, we only use 100 examples for Claude-3 Opus, yet it performs well. In future work, we will need to evaluate the entire testing set. Other models show moderate response to the guides. We have

shown that examples help the model's performance significantly. Table 8 also shows that commercial LLMs can reach 100% accuracy when guides and examples are given to them. The most promise was shown by LLaMA-3-70b within the open-weight realm.

CHAPTER 6

DISCUSSION

The reasoning capabilities of LLMs have been a widely discussed topic recently. Previous research has shown that incorporating model-based evaluations can improve alignment with human judgments, but these models often have limitations in understanding and generating complex language structures. Prompting the LLM to provide a single score value is suboptimal due to restrictions in the generated content. Evaluating the use of GPT-3.5 as LLM, they did not explore GPT-4 due to limited access. Open-source LLMs may be suitable for this task since their ability to follow instructions has improved a lot and beat commercial models such as GPT-4 across all settings. From the experimental results, we have discovered that it is much easier for humans to distinguish between human-written articles and LLM-generated content than it is for LLMs. However, we can improve their accuracy by tuning the instructions given to the model. More importantly, we have shown that by providing both examples and guides to the model even open-source models can reach near-perfect accuracy and give reason for the response reciting the guidance that was given to it.

CHAPTER 7

CONCLUSIONS

Since the inception of GPT-3.5, many editorial offices have employed GPT to write news articles. Our work is the first step in discerning the origin of the article. The result implies that we were able to tune LLMs to successfully identify between human-written news text and LLM-generated text. Although scoring can be useful for quick evaluations, it is not always reliable for comprehensive assessments. Pairwise comparison, by considering the experimental performance of different settings, offers a more accurate and reliable method to evaluate news articles.

7.1 LIMITATIONS

The main limitations are due to the lack of resources to carry out a more comprehensive human study and more extensive annotations. For this reason, we only used 506 annotated samples. We think that the sample size is a sufficient number to demonstrate meaningful observations. Another limitation was the annotation tool that we used, but I have overcome those challenges. Despite their advantages, direct score methods and pairwise comparisons still face challenges. However, we were able to improve their performance.

7.2 FUTURE WORK

We showed that pairwise comparison can reach near-perfect accuracy. In the future, we also want to involve the exploration of novel approaches to generate news articles using LLMs that better bridge the gap between the general public and the scientific community. In this study, we

have looked at document-level assessments. We may want to dig into the sentence level to estimate the fraction of news articles written by AI.

In the future, we would like to address the challenges above in addition to conducting more extensive human studies. We are trying to find some statistical evidence demonstrating that more people are reading scientific news than scientific papers. Then, we might be able to reduce the amount of time people spend to achieve the same level of understanding by reading scientific news than by reading scientific papers.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic, *Claude*, `https://www.anthropic.com/news/claude-3-family`, Accessed: 2024-07-11, 2023.

[3] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," *Proceedings of ACL-WMT*, pp. 65–72, 2004.

[4] S. Banerjee and A. Lavie, "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*, Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: `https://www.aclweb.org/anthology/W05-0909`.

[5] S. Banerjee and A. Lavie, "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*, 2005, pp. 65–72.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.14165.

[9] Y. Chen, R. Wang, H. Jiang, S. Shi, and R. Xu, "Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study," *arXiv preprint arXiv:2304.00723*, 2023.

[10] C.-H. Chiang and H.-y. Lee, *A closer look into automatic evaluation using large language models*, 2023. arXiv: 2310.05657 [cs.CL].

[11] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[12] S. Dai, Y. Zhou, L. Pang, W. Liu, X. Hu, Y. Liu, X. Zhang, and J. Xu, "Llms may dominate information access: Neural retrievers are biased towards llm-generated texts," *arXiv preprint arXiv:2310.20501*, 2023.

[13] Y. Ding, W. Fan, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meets llms: Towards retrieval-augmented large language models," *arXiv preprint arXiv:2405.06211*, 2024.

[14] M. Fomicheva and L. Specia, "Taking mt evaluation metrics to extremes: Beyond correlation with human judgments," *Computational Linguistics*, vol. 45, no. 3, pp. 515–558, 2019.

[15] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024.

[16] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "Sumpubmed: Summarization dataset of pubmed scientific articles," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021, pp. 292–303.

[17] L. Halawi, A. Clarke, and K. George, "Evaluating predictive performance," in *Harnessing the Power of Analytics*. Cham: Springer International Publishing, 2022, pp. 51–59, ISBN: 978-3-030-89712-3. DOI: 10.1007/978-3-030-89712-3_4. [Online]. Available: https://doi.org/10.1007/978-3-030-89712-3_4.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[19] K. S. Kalyan, "A survey of gpt-3 family large language models including chatgpt and gpt-4," *Natural Language Processing Journal*, p. 100 048, 2023.

[20] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[21] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "Gpteval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[22] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, *et al.*, "Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arxiv," *arXiv preprint arXiv:2304.01852*, 2023.

[23] A. Liusie, P. Manakul, and M. J. Gales, "Zero-shot nlg evaluation through pairware comparisons with llms," *arXiv preprint arXiv:2307.07889*, 2023.

[24] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, and D. Roth, *Recent advances in natural language processing via large pre-trained language models: A survey*, 2021. arXiv: 2111.01243 [cs.CL].

[25] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[26] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[28] C. Park, M. Choi, D. Lee, and J. Choo, "Paireval: Open-domain dialogue evaluation with pairwise comparison," *arXiv preprint arXiv:2404.01015*, 2024.

[29] D. Pu, Y. Wang, J. Loy, and V. Demberg, "Scinews: From scholarly complexities to public narratives–a dataset for scientific news report generation," *arXiv preprint arXiv:2403.17768*, 2024.

[30] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

[31] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[32] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[33] O. Sharir, B. Peleg, and Y. Shoham, "The cost of training nlp models: A concise overview," *arXiv preprint arXiv:2004.08900*, 2020.

[34] G. Sharma and D. Sharma, "Automatic text summarization methods: A comprehensive review," *SN Computer Science*, vol. 4, no. 1, p. 33, 2022.

[35] E. I. Sicilia-Garcia, J. Ming, F. J. Smith, *et al.*, "Extension of zipf's law to words and phrases," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[36] S. Singh and N. Singh, "GPT-3.5 vs. GPT-4, Unveiling OpenAI's Latest Breakthrough in Language Models," Nov. 2023. DOI: 10.36227/techrxiv.24486214.v1. [Online]. Available: https://www.techrxiv.org/articles/preprint/GPT-3_5_vs_GPT-4_Unveiling_OpenAI_s_Latest_Breakthrough_in_Language_Models/24486214.

[37] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[38] S. Tan, S. Joty, K. Baxter, A. Taeihagh, G. A. Bennett, and M.-Y. Kan, "Reliability testing for natural language processing systems," *arXiv preprint arXiv:2105.02590*, 2021.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[40] X. Wang and C. Yu, "Summarizing news articles using question-and-answer pairs via learning," in *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, Springer, 2019, pp. 698–715.

[41] Z. Wang, X. Shan, X. Zhang, and J. Yang, "N24news: A new dataset for multimodal news classification," *arXiv preprint arXiv:2108.13327*, 2021.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[43] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, 2024.

[44] W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger, "On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation," *arXiv preprint arXiv:2005.01196*, 2020.

[45] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

VITA

Dominik Soós

Department of Computer Science

Old Dominion University

Norfolk, VA 23529

EDUCATION

* **Master of Science in Computer Science**, Old Dominion University 2023 - 2024

  Major Courses: Intro to AI, Machine Learning, NLP, HPC and Big Data

* **Bachelor of Science in Computer Science**, Old Dominion University 2020 - 2023

  Overall Cumulative GPA: 3.85, Minor in Mathematics *Magna Cum Laude*

* **Associate in Science, Computer Science**, City College of San Francisco 2018 - 2020

  Cumulative GPA: 3.73, *Major Courses: Data Structures and Algorithms, Calculus I-III*

EXPERIENCE

Graduate Research Assistant May 2023 - Present

* **LAMP-SYS:** Lab for Applied Machine Learning and Natural Language Processing Systems Scientific News Generation using Large Language Models

* **HiPSTERS:** High Performance Scientific Computing Team for Efficient Research Simulations, Objective: Global Optimization on CPU using Genetic Algorithm for multidimensional non-convex mathematical test functions. The main focus was the integration of forward automatic differentiation and utilization of the power of GPUs.