Journal of Digital Assets and Monetary Research

Volume 1, Issue 1

Research Article

Date of Submission: 04 May 2026
Date of Acceptance: 28 May 2026
Date of Publication: 01 June 2026

# Summarised: An Analysis and Evaluation of ChatGPT as a Text-Summarisation Tool in an Era of Boundless Information

**Lucas Broszies***
Independent Researcher, Schiller-Gymnasium Berlin, Germany

***Corresponding Author:** Lucas Broszies, Independent Researcher, Schiller-Gymnasium Berlin, Germany.

**Citation:** Broszies, L. (2026). Summarised: An Analysis and Evaluation of ChatGPT as a Text-Summarisation Tool in an Era of Boundless Information. J Digi Assets Monetary Res. 1(1), 01-15.

## Abstract

This study evaluated a total of 10 summaries which were derived from news articles found on the BBC-web page. First, the summaries' retention ratios were calculated by dividing their word count by that of their corresponding original text. Next, the summaries were assessed using the ROUGE-1 metric which is divided into the ROUGE-1 precision score, the ROUGE-1 recall score and the ROUGE-1 F1 score, the harmonious mean of the two others. These were calculated using tokenisation of the original texts and the quantification of unigram overlap of said tokens.

Specifically, the ROUGE-1 precision score measures the amount of words in the summary which are relevant (i.e. keywords present in the original text) relative to the entire summary's length, whereas the ROUGE-1 recall score measures the proportion of words which were used in the summary that are also present in the original text. The ROUGE-1 F1 score is derived from multiplying the product of precision and recall scores divided by their sum by two. After this, the cosine similarity of all the summaries and their corresponding texts was calculated by, again, tokenising the texts, vectorising the two texts within the multidimensional space than arises from expressing each token as a dimension. To calculate the cosine similarity of the two vectors in the multidimensional space, the dot product of the vectors was divided by the product of their magnitudes. This then yielded a decimal representing the overall semantic similarity of the generated summary and the article. Overall, we reached the conclusion that ChatGPT's summary quality declines with longer texts, showing higher precision but lower summary retention, cosine similarity, and ROUGE-1 scores, with shorter texts performing better overall. This shows us the impact that input text length has upon the result.This suggests that the model used in the study may need further refinement for handling longer documents, particularly in terms of retaining key information and maintaining semantic similarity.

## Introduction

In our modern era of constant streams of information and media, it can be overwhelming to keep track of what is relevant or of interest and what is of no or marginal interest. It is this paradox which creates the need for summarisation tools such as ChatGPT and renders them very attractive

for the general public as they seem a perfect solution for the issue at hand: the boundless flow of information. This study aims at scientifically evaluating the summarisation capabilities of ChatGPT and determining, whether it solves the problem of information-overflow or whether it merely further contributes to its exacerbation.

Ever since OpenAI released ChatGPT in the early months of 2022, the word "summary" has been inadvertently linked directly to Artificial Intelligence, suggesting a world in which such language models excel at tasks regarding summarisation and compression of information. But is this a deserved reputation or one hastily put upon such models by a society only too willing to outsource their already minimal time spent reading to such language models? And can ChatGPT comprehensively summarise any given text, while communicating the core information to the reader?

These are only two of the many questions this paper will delve into and attempt to answer in order to either affirm or disprove the still hazy reputation of ChatGPT as a summarisation tool. In essence, this study will assess ChatGPT's ability to summarise text effectively in accordance with our needs as a society almost suffering from this overflow of information.

## Literature Review

AI and specifically OpenAI's language model ChatGPT being a very novel technology, little to no published research has been conducted in the field of summary assessment. Various studies, however, have been published in rather similar fields such as that of Wang et Al. (2023) which sought to evaluate ChatGPT's performance as a natural language generation (NLG) evaluator. The researchers found that the language model produced results competing with human judgement across multifaceted tasks, including summaries. Furthermore, this lead them to the conclusion that the language model showed potential for being a reasonable evaluation metric in terms of assessing summary quality.

Other published studies, on the other hand, detail many of the methods used for this study such as ROUGE-scores or cosine similarity of two vectors in a multidimensional space. Specifically, C. Lin's study (2004) which introduced the ROUGE-score as a metric in determining the lexical overlap between two texts, greatly details the implementation of ROUGE-scores and their computation. In this paper, Lin introduces a whole row of principles which guide the successful usage of the ROUGE-score metric. These include stop word exclusion and n-gram overlap which will both be detailed in the Methodology section of this paper.

"The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents" (2017) is the title of a study authored by Gunawan et Al. that firmly establishes the use of tokenisation and vectorisation to determine the cosine similarity of two texts as a highly useful metric as it facilitates an objective calculation of the semantic similarity between two given documents. This is of special importance to this study as objectively determining the content overlap, regardless of lexical overlap, is a primary challenge in assessing and quantifying the quality of a summary. Additionally, Gunawan et Al.'s paper underscores the importance of text preprocessing—including punctuation removal, case normalisation, stop word elimination—in

enhancing the accuracy of cosine similarity measurements. This finding is particularly relevant for evaluating AI-generated summaries, as meticulous preprocessing ensures that the semantic content is accurately represented, thereby improving the reliability of similarity assessments.

While the lack of specific research into this field makes postulating the results of this study very difficult, there are some broad hypotheses which can be plausibly made. For instance, previous studies have shown that conciseness and accuracy is a field which ChatGPT excels in on the whole, suggesting that especially conciseness of the generated summaries could turn out to be a major strength. Another plausible hypothesis which is largely supported by Wang et Al.'s paper on ChatGPT as an NLG evaluator is the high sensitivity the language model has to prompt design and input. In terms of determining summary quality, it could be said that potential differences between the original texts could greatly impact the quality of the resulting summary. The nature of said differences, however is The nature of these differences, however, remains undetermined as further investigation is required to understand how variations in the input influence the summary quality

## Methodology

To begin with, a total of ten articles were selected from https://www.bbc.com/ (as this is the most visited news website globally with more than 450 million unique users monthly according to https://www.bbc.com/mediacentre/2024/bbc-global-audience-measure?) which greatly varied in length, complexity and topic. A tabular overview of these articles which includes length and topic can be observed below.

**Table 1.: Length and Domain of all articles used in the analysis of ChatGPT's summarisation capabilities**

| Article | Word Count (length) | Domain |
|---|---|---|
| American Airlines resumes flights after technical issue | 302 | Technology |
| When TikTok's underconsumption trend meets festive excess | 1019 | Social Media |
| Scientists unveil baby mammoth remains | 332 | Science |
| Brazil shuts BYD factory site over "slavery conditions" | 435 | Business |
| Zelensky condemns Christmas Day attack | 580 | International relations |
| Twenty years on: "my boat was metres from the shore when the tsunami hit" | 1308 | Natural disasters |
| Spacecraft attempts closest ever approach to Sun | 611 | Science |
| The animals that give each other gifts | 1346 | Science |

| | | |
|---|---|---|
| A slow explosion: the violent birth of the Geminid meteor shower | 1212 | Science |
| Morrisons apologises after discount and delivery issues | 758 | Retail services |

Next, the phrase "Summarise this text please:" followed by the articles was pasted into the ChatGPT interface at https://chatgpt.com. This project consistently utilised the "GPT-4o mini"-model as a means of summarisation.

After the summaries had been generated, they were, along with their corresponding articles, inserted into a spreadsheet into which the word count was also registered after being determined using the word counter at https://wordcounter.net/. After the word counts of all Articles and their summaries had been determined, the Compression ratio, or the length of the summary compared to that of the original, expressed in a percentage.

**ROUGE-Scores**

Next, the so-called "ROUGE-1" scores for each of the articles' relationship to their summary were computed. While many variants of the ROUGE-score exist, this study solely made use of ROUGE-1 scores which mainly measure the overlap of single words (unigrams) between the article and the summary(1). Others include the ROUGE-L and ROUGE-2 scores which both provide more complex evaluations but do not necessarily improve clarity, hence the use of exclusively ROUGE-1. The ROUGE-1 evaluation is divided into 3 sub-scores which are commonly known as: the ROUGE-1 precision score, the ROUGE-1 Recall score and the ROUGE-1 F1 score, this one being a harmonious mean of the two others(2).

The Recall score, as are all other ROUGE-scores, is scored from 0 (no similarity) to 1 (complete overlap)(3). Specifically, the recall score measures the proportion of key-words present in the summary compared to all the key-words from the original text. This can be condensed to the following formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Concretely, recall shows how many details from the original article were included in the summary. While high recall means that the summary includes many relevant information from the original text, when the value is too high it can also point to an overflow of information(4).

The Precision score, on the other hand, measures the proportions of words in the summary that are relevant, compared to the total number of words in it(5). Again, there is a formula which this can be condensed to:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision shows, effectively, how many words in a summary are relevant, i.e. the "information density"(6). While a high precision score does indicate the absence of irrelevant details and a high information density, a score too high suggests that some important information is being missed.

Finally, the ROUGE-F1 score is the harmonic mean of precision of recall. It balances both metrics, thus providing us with a single score that demonstrates the quality of a summary. The Formula to calculate the ROUGE-F1 score is the following:

$$\text{F1Score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

Specifically, the F1-score represents the balance between precision and recall. A high F1 score means the summary is both accurate (few irrelevant words) and complete (most relevant words are included). A low F1 score indicates the summary has either missed important details (low recall), included irrelevant details (low precision) or both(7). Again, a score too high could point to either important information being left out or too much unnecessary information being in the summary. This is visualised by the graph in figure 1.

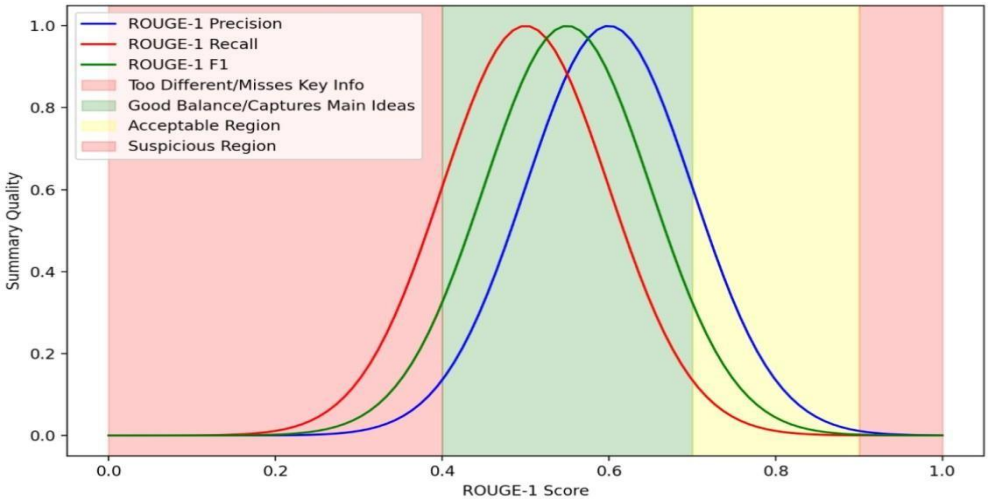



**Figure 1. Relationship of ROUGE-1 Scores and Summary Quality** While this graph may not be based on an entirely objective metric (y-axis), it effectively illustrates the general trend of what constitutes desirable ROUGE-1 scores in relation to summary quality.

**Retention Ratio**

The ratio of retention is an additional parameter in the evaluation of a summary as it shows us how strongly the original text has been compressed, i.e. shortened in the making process of the summary. It reveals nothing however, about the content similarity or dissimilarity, making it a tool solely to be utilised for the determination of the length of the summary in relation to that of the original text. The compression rate is expressed as either a decimal ranging from 0 to 1 or a percentage relating the summary's word count to that of the original text. This can be expressed in the following formula[8]:

$$\text{Retention Ratio} = \frac{\text{Length of Summary}}{\text{Length of Original Text}}$$

This study consistently expressed all retention ratio-values as percentages, thus this very closely

related formula was employed:

$$\text{Retention Ratio} = \frac{\text{Length of Summary}}{\text{Length of Original Text}} \text{ x } 100$$

It is nearly impossible to make general statements as to the ideal retention ratio as this value is highly context and purpose-dependent. For instance in academic contexts, where more details are of the essence, a ratio of 20-40% could be the most reasonable, whereas within media contexts, where brevity and clarity are prioritised, a retention ratio of 12-20% may be more appropriate to ensure the core message is conveyed succinctly. As a general rule, however, one could state that the retention ratio should not exceed ~ 35% and should ideally not be less than ~ 15%[9].

**Cosine Similarity**

Cosine similarity expresses all the key words in two given documents (in our case the original article and the summary) as dimensions in a space. Next, the two texts are expressed as vectors within this multidimensional space and the similarity between the two vectors is calculated be dividing the dot product of the vectors by the product of their magnitudes. Formulaically, this can be expressed as(10):

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$$

This results in the cosine similarity of two documents which is a value mathematically ranging from -1 (diametrically opposed vectors), over 0 (no similarity) to 1 (total overlap). For purposes of assessing two documents' similarity to one another, however, this value never falls below 0, as the vectors are usually non-negative in text-based applications. Thus, this study treats the cosine similarity of two texts as a value ranging from 0 to 1.

This method will now be demonstrated, explored and explained using a simplified example to help illustrate the vectorisation of texts to calculate their overall similarity. Beginning this example with two sentences which represent our article and a summary thereof:

**Table 2.: Cosine Similarity example: Comparison of Detailed and Summarised Sentences**

| Sentence Representation | Sentence |
|---|---|
| (A) **Detailed Sentence (Article)** | The company's new marketing strategy includes a combination of social media campaigns, influencer partnerships, and targeted email newsletters to reach a broader audience and increase brand awareness. |

| | |
|---|---|
| (B) **Summarised Sentence (Summary)** | The company's new marketing strategy focuses on expanding its reach through social media, influencers, and email campaigns. |
| (C) **Detailed Sentence (Article)** | The company's new marketing strategy includes a combination of social media campaigns, influencer partnerships, and targeted email newsletters to reach a broader audience and increase brand awareness. |
| (D) **Summarised Sentence (Summary)** | The company's new marketing strategy focuses on expanding its reach through social media, influencers, and email campaigns. |

The first step to assessing the cosine similarity of these texts is Tokenisation. The first step in doing this is tokenising the two sentences, i.e. determining key words, omitting "stop-words" (common words that carry little meaningful information such as articles, prepositions, conjunctions and pronouns). Next, the keyword frequencies are represented as vectors in the vectorisation step. Finally, the dot product, magnitude and cosine similarity are determined using the calculations in the below table.

**Table 3.: Steps to determining the Cosine Similarity of two Documents**

| Step | Result |
|---|---|
| **Tokenisation** | Vocabulary: "company", "new", "marketing", "strategy", "includes", "combination", "social", "media", "campaigns", "influencer", "partnerships", "targeted", "email", "newsletters", "reach", "broader", "audience", "increase", "brand", "awareness", "focuses", "expanding", "influencers" |
| **Vectorisation** | Vector for sentence A: {1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0}<br>Vector for sentence B: {1,1,1,1,0,0,1,1,1,0,0,0,0,0,1,0,0,0,1,1} |
| **Dot Product Calculation** | A·B = 1·1+1·1+1·1+1·1+0·0+0·0+1·1+1·1+1·1+0·0+0·0+0·0+0·0+0·0+0·0+1·0+0·0 +0·0+0·0+ 0·0+0·0+0·0+0·0 = 9 |
| **Magnitude Calculation** | Magnitude of A: ‖A‖ = √20·1+2·0 = 4,472 Magnitude of B: ‖B‖ = √10·1+12·0 = 3,162 |
| **Cosine Similarity Calculation** | Cos-similarity = 9/4,472·3,162 = 0,636 |

The cosine similarity is of immense importance to assessing the accuracy of a summary because it quantifies how closely the summary reflects the key ideas and principles of the original text, thus enabling an objective judgement of summary quality. In this study, Cosine similarity values were

used alongside ROUGE-1 scores as they provide detailed insights into the semantic similarity of two texts, while the latter primarily focusses upon lexical overlap.

Similar to all the ROUGE-1 scores, the cosine similarity does not directly proportionally correlate to the quality of a summary, as a similarity too high could lead to a summary which merely "parrots" the original text without condensing and rephrasing ideas. Hence a somewhat lower, but still considerably high cosine similarity score between ~0,6 and ~0,9 is desirable. This is illustrated by the following graph in Figure 2.

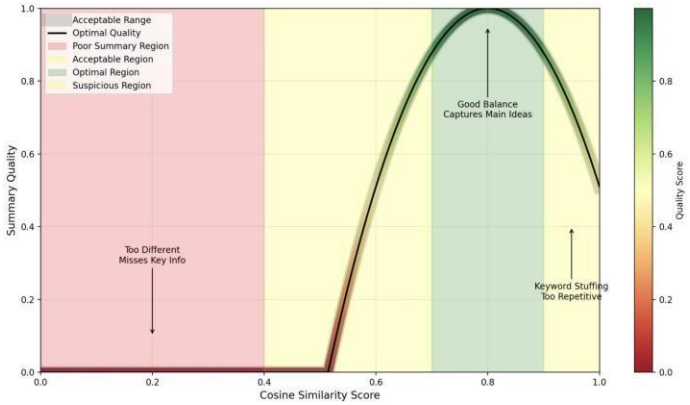



**Figure 2. Relationship Between Cosine Similarity and Summary Quality**

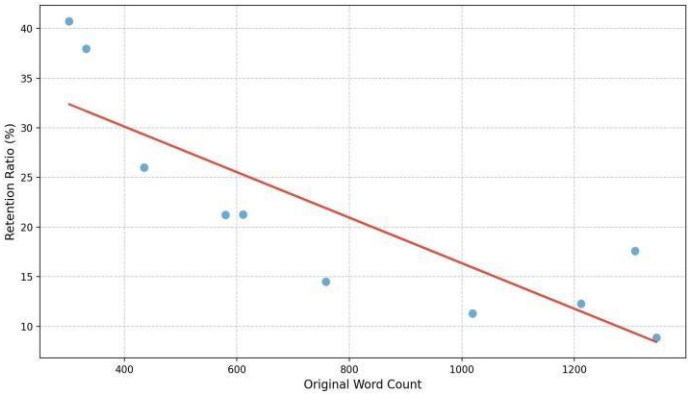
While this graph may not be entirely objective or scientific, it effectively illustrates the general trend of what constitutes desirable cosine similarity scores in relation to summary quality.

## Results

The results of this study provide insights into the relationship between all parameters used and address the central question of whether ChatGPT is an adequate summarisation tool. This section is organised into three subsections, detailing findings regarding ROUGE-1-scores and cosine-similarity findings and. Key results are supported by various graphs to illustrate the observed patterns.

### Findings Regarding the Summary Retention Ratio

The mean retention ration was determined as 21,17%. This number stems from a multifaceted




dataset in which retention ratios spanned from 8,84% to 40,73%. The mean, however, falls within

the range earlier deemed desirable. It was observed that a distinct negative correlation of approximately -0.846 exists between the length of the original article and the retention ratio. This finding implies that as the length of the input document increases, the summaries derived from it capture a smaller proportion of the original text (figure 3.).

**Findings Regarding the ROUGE-1 Scores of the Generated Summaries**

ROUGE-1 scores are divided up into three sub-scores: precision, recall and the first harmonious mean (F1). ChatGPT's language model performed very differently in each of these scores, highlighting its strengths and weaknesses in various areas.

**ChatGPT's Performance in the ROUGE-1 Precision Score**

The mean ROUGE-1 precision score for the AI-generated summaries lies at 0,66. This demonstrates one of the model's primary strengths: conciseness. A higher precision score indicates that more unigrams in the generated summary are relevant compared to its total word count. This showcases that ChatGPT is effective at retrieving important information, and above all, writing in a fashion which allows for high information density, i.e. communicating key ideas clearly and efficiently without unnecessary elaboration.

An interesting trend which was taken note of during this study is the positive correlation of around 0,4 between the word count of the input document and the ROUGE-1 precision score. While this may not be an especially strong, definitive correlation, it suggests that summaries derived from longer documents may achieve slightly higher precision (Figure 4.)

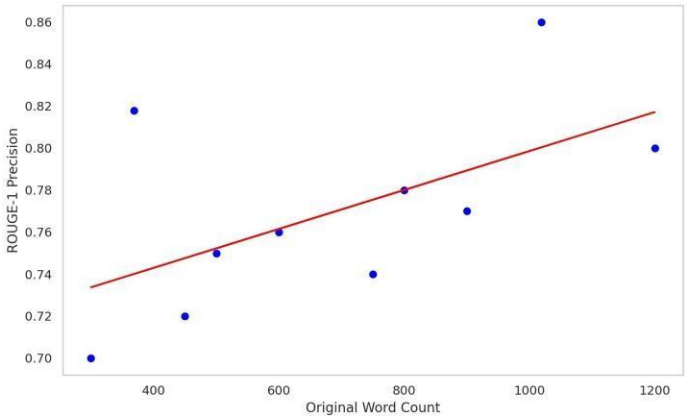



**Figure 4. Original Word count vs. ROUGE-1 precision score**

**ChatGPT's performance in the ROUGE-1 recall Score**

The AI-generated summaries performed considerably worse in the ROUGE-1 recall domain: with a mean score of 0,17, the model often failed to include sufficient detail from the original document. The recall score represents the proportion of key-words used in the summary to all occurring key-words in the input document. This means that ChatGPT frequently left out many important keywords (i.e. details from the original article).

Previously, the length of the Article had shown a direct positive correlation to both ROUGE-1 F1 and precision, the ROUGE-1 recall scores, however, seemed to diminish with increasing input word count. While the recall scores acquired by the ChatGPT-summaries never exceeded the rather low score of 0,24, a strong negative correlation of around -0,83 between the original word count and the recall score was observable. This may seem puzzling at first, but when the following graph relating the original word count and that of the summary is taken into account, it becomes clear why this relation arises.

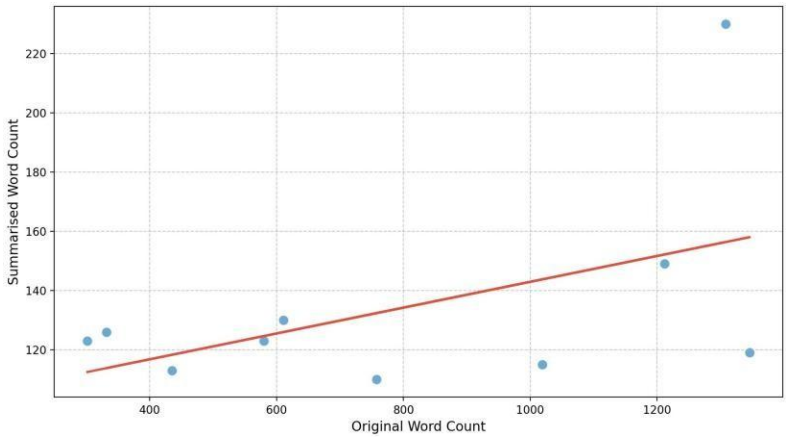



**Figure 5. Original vs. Summarised Word Count**

This scatterplot shows us that the summary's word count does not grow proportionately to that of the input document. Consequently, at higher word counts of the original text, the summary captures fewer details relative to the total content, resulting in lower recall scores.

In other words: as previously established, as the original word count increases, the retention ratio decreases. What importance does this bear upon the topic of the negative correlation between the original word count and the ROUGE-1 recall score?

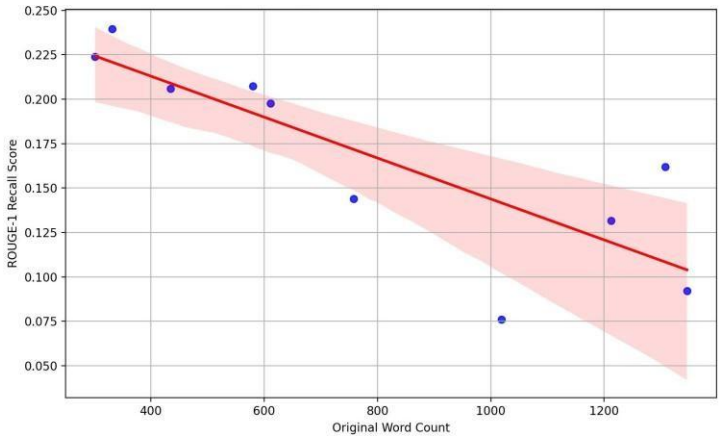



**Figure 6. Original Word Count vs. ROUGE-1 Recall Score**

This scatterplot shows us that the summary's word count does not grow proportionately to that of the input document. Consequently, at higher word counts of the original text, the summary captures fewer details relative to the total content, resulting in lower recall scores.

In other words: as previously established, as the original word count increases, the retention ratio decreases. What importance does this bear upon the topic of the negative correlation between the original word count and the ROUGE-1 recall score?

The fact that as the length of the input document increases, the retention ratio of the summary decreases means that thus, inevitably, the summaries of longer texts will be able to include less detail, resulting in a lower ROUGE-1 recall score and thus in the following graph shown in Figure 6. The correlation between original word count and the ROUGE-1 recall score lies at -0,83, underscoring this distinctly inverse relationship.

**ChatGPT's Performance in the ROUGE-1 F1 Score**

The ROUGE-1 F1 score is the first harmonic mean of the precision and recall scores, which means that it does not generate novel information but rather optimally balances the retrieval of relevant data (precision) with the inclusion of critical content from the source material (recall). As to be expected, the mean of the F1 score lies in between the mean averages of the precision and recall scores at 0,26. The Dataset includes values ranging from 0,14 to 0,33. These scores fall within the range labeled "insufficient detail or key information missing" on the graph shown in Figure 1, which approximately relates ROUGE-1 scores to summary quality.

Additionally, a strong positive correlation of 0,813 was observable between the retention rate and the F1 score, which indicates that summaries that are more elaborate in relation to the length of the input document tend to be better in quality of lexical overlap. This is likely due to the fact that the higher retention ratio allowed for more information to be included in the summary. The following scatterplot (Figure 7.) demonstrates this relation. This also means that longer documents' summaries usually received lower F1 scores, as there is an inverse relationship between the original word count and the summary retention ratio and one of positive nature between the retention ratio and the F1 score. The correlation of around -0,796 between input document length and ROUGE-1 F1 score proves this assertion.

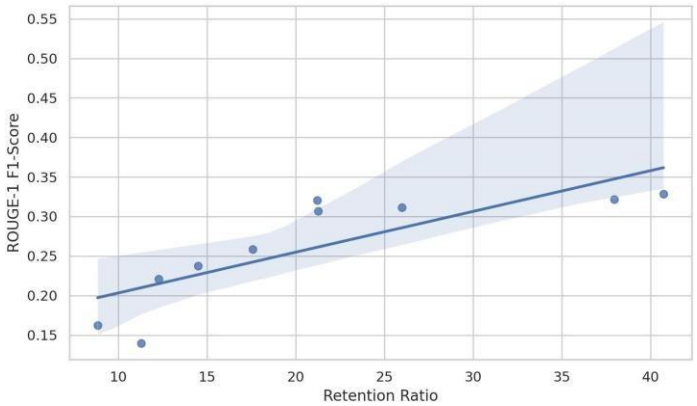



**Figure 7. Retention Ratio vs ROUGE-1 F1 Score**

### Findings Regarding the Cosine Similarity

The study found a mean cosine similarity between the original article and its derived summary of 0,61 which falls right in the middle of the region deemed acceptable in Figure 2. The cosine similarity values across all summaries ranged from 0,36 to 0,79 which highlights ChatGPT's varied performance depending on the original documents's word count. As previously established, a longer input document led to smaller retention ratios; i.e. the longer an original text, the less information the summary includes.

This is further underlined by the following graph relating summarised word count and cosine similarity as it shows a distinct negative correlation of-0,359 between the two. This finding also supports previous ones which also indicated decreasing summary quality with increased original word count.

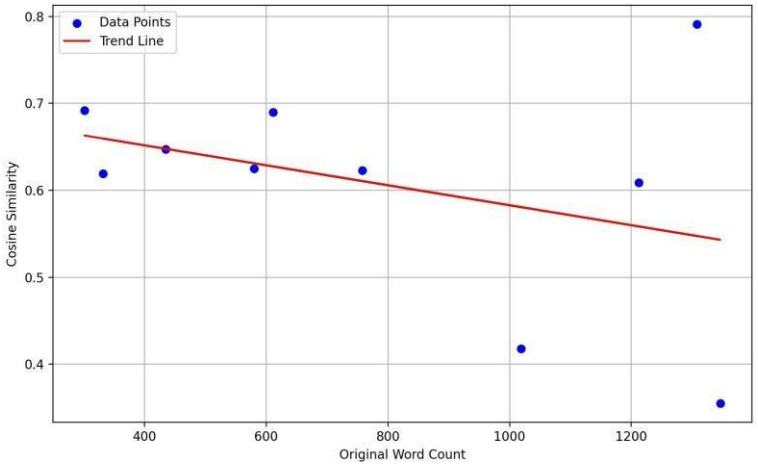



**Figure 8. Original Word Count vs. Cosine Similarity**

The scatterplot in Figure 8 clearly shows that while there may be some exceptions, the input document's length directly, negatively impacts the summary's semantic similarity to the original text. This is likely due to the fact that the model fails to keep a fixed retention rate especially when it comes to longer input, thus negatively impacting summary quality at increased original word count.

## Limitations

While this study faced no major limitations of any physical nature, there are two primary challenges we faced in conducting this project which assesses ChatGPT's summary-writing capabilities. The first one being the small number of input documents surveyed: at 10, we still get a good idea of the main correlations and trends, we do, however acknowledge that a larger sample size would provide a more robust and generalisable analysis of ChatGPT's summary-writing capabilities.

Next, the study focussed solely upon OpenAI's ChatGPT-4o mini, thus all our findings relate only the summarisation skills of this model. While a broader scope of ChatGPT's models may well have provided a more comprehensive understanding of the summarisation capabilities across different versions, our study specifically aimed to analyse the performance of ChatGPT-4o mini in detail.

## Discussion and Analysis

Many trends and relationships were observed and documented over the course of this study. These can be categorised into two groups: those which were hypothesised based on previous research and were found to be correct, and those which either were not considered or were falsely hypothesised. This division allows for an accurate and logical analysis of findings and their implications.

### Hypothesised Findings

A major strength of the language model revealed itself early in the study as its conciseness. This was demonstrated by the relatively high ROUGE-1 precision scores obtained by the language model, showcasing its ability to densely pack essential information into its output. In the literature review, this finding was hypothesised, as previous studies, such as N. Chanchad's work on the accuracy of ChatGPT as a summarisation tool for datasets and L. Giray's paper on enhancing communication using language models, have highlighted conciseness as one of ChatGPT's key strengths. Conciseness is clearly an indispensable attribute in generating summaries of media articles, as it enables the effective conveyance of key information while minimising unnecessary details, ensuring that the content remains accessible and easy to understand for readers. This ability not only saves time but also enhances comprehensibility by presenting information in a clear, deft and structured manner.

### Unforeseen or Falsely Hypothesised Findings

Few concrete research hypotheses were established leading into this study, mainly due to the severe gap in published research on the topic. The absence of such studies made it close to impossible to foresee specific findings such as the rather low ROUGE-1 recall and F1 scores. The low scores reached by ChatGPT in these metrics are very likely to be due to the very low word count relative to the input document's length many summaries had, thus rendering it impossible to reach a level of lexical overlap sufficient for higher scores in these metrics. These findings underscore the challenge of balancing conciseness with comprehensive coverage when summarising large texts, highlighting areas for potential improvement in future models.

One additional result found in this study was the unexpected inverse correlation between input document length and summary retention ratio. This correlation, illustrated in Figure 3, represents a major flaw in ChatGPT-generated summaries as it greatly influences not only the length of the summary but likely also other metrics such as the cosine similarity and the ROUGE-1 F1 and recall scores. This finding holds considerable importance for the general use of ChatGPT as a summarisation tool for media articles, as it strongly suggests that the language model may fail to capture all relevant nuances of the original text, potentially leading to semantic alterations when summarising articles.

## Conclusion

Finally, when taking into account all the results gathered in this study, we can come to the conclusion that ChatGPT's ability to summarise data depends greatly on the length of the input document. We observed that there was a distinct negative correlation of -0,846 between original word count and the summary's retention ratio which implies that, as the input document grows in

length, the summary thereof tends to become shorter (in relation to the input length). This means two things: first and foremost, this was shown to negatively impact ROUGE-1 recall, F1 and cosine similarity scores, suggesting that the lower summary retention ratio leads to more details being omitted and thus the aforementioned scores to also show a mean negative correlation with input length of approximately -0,662.

Next, the second observation made during this study relating to summary quality in relation to original document length is the fact that there was a positive correlation of 0,4 between ROUGE-1 precision score and the original word count. In order to explain this, one must clarify briefly what exactly the precision score demonstrates. In general terms, the precision score can be said to be a quantitative measure of what proportion of words in the summary is relevant (i.e. usage of keywords from original text) compared to its total number of words.

From this (and the fact that there is a negative correlation between retention ratio and original word count) we can surmise that as the input documents grow in length, the model generates shorter summaries in relation to the original text and thus becomes more selective in the words it uses, reaching a higher ROUGE-1 precision score.

While this higher information density results in an increased ROUGE-1 precision score, it also creates a trade-off between precision and retention ratio as in longer texts the summaries likely focus on extracting only the most important information, leaving out much of the content at the expense of the retention ratio.

In conclusion, we can state that summaries generated by ChatGPT's-4o mini model tend to decrease heavily in quality as the input document grows in length. While the precision score does grow slightly, this is not an indicator of increased summary quality but rather the logical result of the summary retention rate drastically declining. Specifically, while shorter input texts (of around 300-400 words) produced acceptable results in the summary retention ratio, cosine similarity and precision scores, they consistently performed comparably worse in ROUGE-1 recall and F1 scores. Longer texts (500+ words) on the other hand, were consistently outscored by their shorter counterparts, rarely reaching scores in the acceptable region in any metric [1-10].

## References


1. Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." Proceedings of the ACL Workshop on Text Summarisation and Question Answering, 2004.
2. Ganesan, Kavita. "An Intro to ROUGE, and How to Use It to Evaluate Summaries." Medium, 2017.
3. Gao, Mingqi et Al. "Human-like Summarisation Evaluation with ChatGPT." Proceedings of the 2023 Conference on Natural Language Processing, 2023.
4. Laskar, Tahmid-Rahman et Al. "A Systematic Study and Comprehensive Evaluation of ChatGPT on Text Summarisation." Proceedings of the 2023 ACL Conference, 2023.
5. Hake, Joel et. Al. "Quality, Accuracy, and Bias in ChatGPT-Based Summarisation of Scientific Literature." PubMed Central, 2023.
6. Santhosh, Sthanikam. "Understanding BLEU and ROUGE score for NLP evaluation." Medium,

2023.
7. Liu, Yiheng. "Summary of ChatGPT-Related research and perspective towards the future of large language models" Meta-Radiology, 2023.
8. Waseemullah, M., et al. "A Novel Approach for Semantic Extractive Text Summarization." Open Research Knowledge Graph, 2022.
9. Rany, Ruby et Al. "An extractive text summarisation approach using tagged-LDA based topic modelling." Multimedia Tools and Applications, 2021.
10. Miesle, Phil. "Exploring the Real-world Applications of Cosine Similarity" Datastax, 2023.