Long Document Text Summarisation

A thesis submitted to The University of Manchester for the degree of Master of Philosophy in the Faculty of Science and Engineering

2023

Jennifer A Bishop

Department of Computer Science,
School of Engineering,
Faculty of Science and Engineering

List of Contents

Word count: 25, 774

List of Figures

List of Tables

Abstract

Text summarisation is the task of converting a longer piece of text into a shorter text whilst communicating the same key points from the original document. The value of automatic text summarisation is derived from the efficiency saving gained through distilling long documents into shorter text. Despite this, most studies researching text summarisation, and the automated metrics used to assess its efficacy, have primarily been focused on short documents. Recently, Pretrained Language Models (PLMs) have been used to improve the performance of text summarisation. However, PLMs are limited by their need of large corpora of data for pretraining (ideally in the domain of any anticipated downstream tasks), labelled training data for fine-tuning, and by their attention mechanism, which often makes them unsuitable for use on long documents. The computational complexity of their attention mechanisms means that if a document is long, it generally must be truncated to be computationally feasible to process. This work aims to develop methods which adapt PLMs in ways to make them suitable for summarisation of long documents.

Three main novel contributions are proposed in this work. Firstly, GenCompareSum, a hybrid, unsupervised, abstractive-extractive method is developed, which cycles through a document generating salient textual fragments and uses these to guide an unsupervised extractive summarisation. This hybrid approach can be easily extended to any document length and out-performs existing unsupervised methods, as well as state-of-the-art supervised methods, despite not needing labelled training data for the summarisation task. Secondly, since most long document data sets are highly domain-specific, a framework for injecting domain knowledge into PLMs is proposed: KeBioSum. Evaluation of this method shows that using an adapter-based framework to inject domain knowledge into PLMs improves performance of text summarisation. Lastly, maintaining factual consistency is a critical issue in abstractive text summarisation, but it cannot be assessed by traditional metrics, such as ROUGE scores. Recent efforts have been devoted to developing improved metrics for measuring factual consistency using PLMs. However, there is a lack of research on automatic metrics which can assess the factual consistency of long document summarisation. To this end, LongDocFACTScore (LDFACTS) is proposed. This metric extends an existing evaluation metric, BARTScore, by comparing each sentence in the generated summary with the most similar sections of the source document. It is designed to be extendable to any length document and demonstrates a strong correlation with the human judgement of factual consistency on long document summarisation data sets. In addition to these three main novel contributions, both intrinsic and extrinsic evaluation of different methods for the abstractive summarisation of long documents are conducted and discussed.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

List of Thesis Revisions

General Revisions

Tables of results, references to tables and figures have been standardised across the thesis.

Section 1 Revisions

The reference to substantiate the claim that the number of controlled trials published every year has increased has been updated to cite a paper published in 2021, which analysed data up to 2019, rather than citing a paper published in 2010, which analysed data up to 2007. Furthermore, the statement "most PLMs restrict their input to a small number of tokens" has been clarified to refer to the context window of a PLM's input.

Section 3 Revisions

Paragraphs 1-3 of this section have been re-worded to clarify that, by generating queries or document titles, the summarisation method is trying to capture the key points of an article. Additionally, the caption of Figure 1 has been updated to include explanations of each enumerated step in the diagram. Lastly, the statement "GenCompareSum (t5-s2orc-title), outperformed LexRank by a large margin" has been toned down.

Section 4 Revisions

In order to reduce the redundant statements in the beginning of Section 4.2 and Section 4.1.3, the previous section 4.1.3 "This Work" has been removed and more detail on the methodology has been included in Section 4.2. In Section 4.2.2, further justification has been provided on the selection of PICO elements as the knowledge source to inject into KeBioSum. In Section 4.4, the implications of the performance improvement obtained by the KeBioSum method have been further discussed.

Section 5 (and 6) Revisions

Section 5 has been split into two Sections (now Section 5 and Section 6). The former focusses on evaluation of methods for abstractive summarisation of long documents, whilst the latter assesses the efficacy of different evaluation metrics for long document abstractive summarisation and proposes a new metric for assessing factual consistency in long document summarisation, LDFACTS. Further detail has been provided on the methodology and experiments in this section,

such as detail on the beam search configuration in the DANCER method, and the noise in the ArXiv data set.

Section 7 Revisions

Section 7 has now been updated to restate the objectives outlined at the beginning of the thesis, and to discuss how the work in the thesis has addressed these objectives. It has also been updated to include ideas for future work, as discussed in the oral examination, such as using LLMs, extending to other domains and settings (such as multi-document summarisation) and an idea around improving the pre-processing of training data to attempt to improve factual consistency in generated summaries.

Acknowledgements

I would first like to thank Professor Sophia Ananiadou for supervising me on this work and providing me with great insight into the world of natural language processing and guidance on conducting my research. Additionally, I would like to thank Dr Qianqian Xie for supporting me with this work and for providing me with technical guidance throughout it. It has been a pleasure to work with both of you and I am grateful for our collaboration and research outputs over the last couple of years.

The author

Jennifer Bishop studied an MEng in Engineering Science at the University of Oxford for her undergraduate degree, achieving a 1st class degree. In her final year of study, she conducted a project which used machine learning to predict whether a patient was ready for hospital discharge. As an outcome of this project, she authored a research paper which was published in PLOS ONE and was named as an inventor on a patent. Since University, Jennifer has is working as a Machine Learning Engineer. For the last two years, alongside her work, she has studied part-time for an MPhil in Text Summarisation at Manchester University. During this time, Jennifer was supervised by Professor Sophia Ananiadou and has authored one research paper, which she presented at the BioNLP workshop, as part of the ACL 2022 conference held in Dublin and was a named author on a second research paper, first-authored by Dr Qianqian Xie, which was published in the journal 'Knowledge-Based Systems'. She has additionally authored a second research paper, which is under review for publication.

1 Introduction

1.1 Background

In the digital era through which we are living, there is an ever-increasing quantity of textual content available to us online. Although it is beneficial in providing us with sources of knowledge to help us with our work and daily lives, it also presents challenges in identifying and comprehending the relevant literature for a given task. As an illustrative example, in the biomedical domain, systematic reviews are generated for a given research topic to synthesise all research in a particular area, however, in the years between 2000 and 2019, there has been a 20-fold increase in the numbers of systematic reviews published on PubMed[1] daily, from 4 to 80 per day (Hoffman et al., 2021), making the task of synthesising all relevant research extremely difficult.

Machine learning techniques are enabling the automation of many natural language processing (NLP) tasks, and in the future will be necessary tools for us to be able to retrieve and comprehend documents as the quantity of available resources to us continues to grow. In this context, an important NLP task for automating the comprehension of documents is text summarisation. Text summarisation is the transformation of a longer piece of text to a more succinct version, where the shortened version still conveys the key points of the original document.

Text summarisation is generally divided into two types of approach, extractive and abstractive text summarisation, where the former directly selects sentences from the original document to form the summary, and the latter rewords the original document in a shorter format, in a similar fashion to the way a human would summarise a document. Traditional text summarisation approaches tend to be extractive and use features, such as sentence position and word frequency, to select the most relevant sentences (Luhn, 1958, Radev et al., 2004; Erken and Radev; 2004).

State-of-the-art (SOTA) approaches generally use transformer (Vaswani et al., 2017) based models, often in the form of pretrained language models (PLMs), such as BERT (Devlin et al., 2019). Extractive, abstractive and hybrid methods have been proposed using these types of models. These modern approaches have shown dramatically increased performance on text summarisation tasks. This is likely due to their ability to learn knowledge, linguistic, and semantic structures during their pretraining. However, these approaches have also introduced new limitations into NLP tasks.

[1] https://pubmed.ncbi.nlm.nih.gov

The attention mechanism which enables these models to learn relationships between words and phrases is extremely computationally expensive to train, thus, most PLMs restrict the context window of their input to a small number of tokens (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020). For this reason, most recent studies on text summarisation using PLMs focus on short documents from news domains (Lewis et al., 2020; Liu and Lapata, 2019). Since the value of text summarisation comes from making long documents more succinct, by summarising only short documents, or truncating a long document before summarisation, much of the value of the process is lost. The aim of this work is to harness the advancements made by use of PLMs for text summarisation, but to address some of the outstanding challenges and to explore practical methods for summarising long documents, without requiring documents to be truncated to only the first few hundred tokens or needing huge computational resource. In addition to the hardware and energy resources required to train large neural models, they additionally require vast amounts of training data, which are often unavailable or costly and resource-intensive to create. To this end, this work also explores unsupervised approaches for text summarisation of long documents, which do not need large, annotated data sets for the summarisation task.

Most long document summarisation data sets are domain specific (Koh et al., 2022), however, most prior literature considers text summarisation for the general domain, rather than other domains such as science or biomedicine. Since these domains could greatly benefit from automated text summarisation to support tasks like the generation of systematic reviews, there is a need to explore the best methods for summarisation of domain-specific documents. In response, this work focuses on long document summarisation for the scientific and biomedical domains and discusses methods to improve performance of the task on domain-specific literature.

Although abstractive summarisation has the potential to be more succinct and readable than its extractive counterpart, prior literature has shown that, in its current state, it cannot be trusted to be factually consistent (Wallace et al., 2021). This makes it unsuitable in many practical applications, such as summarisation of biomedical articles for use by clinicians. One issue which is preventing SOTA methods from advancing in this space is that the primary evaluation metrics used to assess the efficacy of these methods consider only word-overlap (Lin, 2004), and do not correlate well with other important measures, such as factual consistency (Kryściński et al., 2019). Although modern methods have been proposed to measure factual consistency (Scialom et al., 2021; Kryściński et al., 2020; Yuan, et al., 2021), these metrics are only designed to evaluate summarisation of short documents and do not extend well when they are used to evaluate long document summarisation (Koh et al., 2022). To address this gap in research, different abstractive methods, which provide coverage of entire long documents in their generated summaries, are

evaluated. Furthermore, LDFACTS, a metric for evaluating the factual consistency of long document abstractive summarisation, is proposed, and is shown to correlate better with human judgement than existing SOTA metrics for evaluating long document summarisation.

1.2 Research Aims

The aims of this MPhil were as follows:

1. To evaluate existing SOTA extractive and abstractive approaches for long document text summarisation, and to propose novel summarisation methods to improve performance over these existing approaches in long document settings.

2. To investigate text summarisation of domain-specific documents, and to propose novel methods to improve performance over existing SOTA approaches in these contexts.

3. To conduct a study into the suitability of automated metrics for evaluation of long document text summarisation, specifically focusing on their ability to measure the factual consistency of summaries generated from long documents.

1.3 Overview of this work

Here, the structure of the thesis is outlined. Section 2 provides a literature review. An overview of existing extractive and abstractive methods for summarisation of long documents is given, along with a review of methods which consider domain knowledge in their approach. Additionally, an overview of existing automated metrics for evaluating text summarisation is provided.

Section 3 looks to address Research Aim 1, with a focus on extractive summarisation of long documents. In this section, a novel, two-step, unsupervised, hybrid abstractive-extractive summarisation method is proposed, which generates salient textual fragments (specifically, queries or document titles) that represent sections of a document and uses these to guide the extractive summarisation step. The method fuses state-of-the-art PLMs with unsupervised approaches, to achieve a summary which harnesses the semantic knowledge of transformer-based models, whilst being extendable to any length document, without requiring a large corpus of training data. Evaluation results demonstrate this hybrid method outperforms both existing unsupervised methods and state-of-the-art supervised methods, both on long and short documents. The ideas for the work in this section were generated, and experiments undertaken by the author of this MPhil, Jennifer Bishop. This research was supervised by Qianqian Xie and

Sophia Ananiadou. The research was published as part of the BioNLP workshop at ACL 2022 in Dublin (Bishop et al., 2022).

Section 4 looks to address Research Aim 2 and explores summarisation methods for domain-specific documents, with a focus on biomedical literature. In this section, an extractive approach is proposed which uses adapters (Houlsby et al., 2019) to infuse domain knowledge, specifically information about PICO elements, into PLMs during their fine-tuning for the summarisation task. The models developed in this research were shown to outperform SOTA summarisation methods for documents of the biomedical domain. The conceptualisation and methodology for the work in this section were created by Qianqian Xie, who led this research. Qianqian Xie, along with the author of this MPhil, Jennifer Bishop, wrote the software and ran the experiments for this work. Prayag Tiwari additionally ran some experiments for this work. The work was supervised by Sophia Ananiadou and was published in the Journal of Knowledge-Based Systems (Xie et al., 2022). To the best of our knowledge, this is the first work incorporating PICO domain knowledge into PLMs for extractive summarisation on biomedical literature.

Section 5 further addresses Research Aim 1, focusing on abstractive summarisation of long documents. For this research, intrinsic and extrinsic evaluation of a range of abstractive summarisation methods was conducted. Baseline and SOTA abstractive methods were compared to an abstractive method proposed in Section 5.2 of this work, which makes use of document zones for its prediction. For the intrinsic evaluation, a range of automated metrics were used, and for the extrinsic evaluation, human annotations were collected which assessed the measures coherence, fluency, and factual consistency. SOTA methods for abstractive summarisation of long documents are computationally expensive; however, the research in this section restricts computational consumption to a practical level, making the methods accessible to a wide range of researchers and institutions. The ideas for the work in this section were generated by, and the experiments for this work were conducted by the author of this MPhil, Jennifer Bishop. This research was supervised by Qianqian Xie and Sophia Ananiadou.

Section 6 of this work evaluates the efficacy of different automated metrics for evaluating long document summarisation. In this section, it was found that existing automated metrics do not correlate well with human measures of factual consistency in long document settings. In response, a novel metric is proposed which is shown to correlate better with human annotations of factual consistency than any existing metric evaluated in the long document setting. The ideas for the work in this section were generated by, and the experiments for this work were conducted by the author of this MPhil, Jennifer Bishop. This research was supervised by Qianqian Xie and Sophia Ananiadou.

Section 7 of this work provides a conclusion, and suggestions for future directions of work to further research into text summarisation of long documents.

2 Background

2.1 Pretrained Language Models in NLP

The proposal of the attention mechanism (Bahdanau et al., 2014) has transformed NLP research over recent years. In neural models which implement this mechanism for NLP tasks, textual pairs are tokenised, i.e., divided into smaller units such as words or punctuation, and are provided as training data to the model. During the training process, the attention mechanism learns the relationship between tokens in the two texts. Bahdanau et al., (2014) originally proposed this in the context of machine translation, where the textual pairs were made up of one text in the source language and another, equivalent, text in the target language.

Building on this work, Vaswani et al., (2017) proposed the transformer model. This is an encoder-decoder model which uses an attention-based architecture, making use of both self-attention mechanisms, where the input representation is compared to itself in the encoder and the target representation is also compared to itself in the decoder, and an encoder-decoder attention mechanism, which compares the input and target representations to each other.

The transformer model is the foundation architecture used by many pretrained language models (PLMs). PLMs are large neural models, which are pretrained using large quantities of texts, generally using a self-supervised approach, such as masked language modelling (MLM), where words are randomly masked and the model's training objective is to predict the hidden words. The approach taken with these models is to pretrain a PLM over a large corpus of data so that it learns a base level of knowledge and language understanding, and then to fine-tune it on a downstream task. These models have shown significant performance increases on a broad range of NLP tasks. PLMs tend to fall into one of three categories: auto-encoding models like BERT (Devlin et al, 2019), which make use of encoders and tend to be used for classification-style tasks, such as extractive summarisation; sequence-to-sequence models with encoder-decoder architectures, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), which are principally used for tasks which convert one piece of text to another, such as abstractive summarisation or machine translation; and models which make use of decoder architectures, also called auto-regressive models, such as GPT-2 (Radford et al., 2019), which are used for text generation tasks.

However, even the latest, best performing PLMs, such as OpenAI's ChatGPT[2], have limitations on the number of tokens they can process, due to the quadratic computational complexity of the models' attention mechanisms. More efficient attention mechanisms have been developed including LongT5 (Guo et al., 2021), BigBird (Zaheer et al., 2020), Longformer and LED (Beltagy et al., 2020) models; these all utilise different strategies to adapt their attention mechanisms to scale linearly, rather than quadratically, with sequence length. However, these models are still computationally expensive, requiring GPUs to train sequences of any significant length, and are therefore unable to fit textual inputs over a few thousand tokens onto a single GPU. Thus, running NLP tasks with PLMs on long documents remains a challenge.

2.2 Extractive summarisation

The earliest works researching text summarisation focus on extractive methods (Luhn, 1958). These methods tend to be unsupervised, i.e., they do not require labelled training data for the extractive summarisation task. Several early extractive methods are graph-based: LexRank (Erkan and Radev, 2004), and TextRank (Mihalcea and Tarau, 2004) are both based on Google's PageRank algorithm (Brin and Page, 1998), which assumes that the sentences with the highest centrality are the most important and use these to form a summary. In contrast, SumBasic (Nenkova and Vanderwende, 2005) simply assumes that sentences containing the words which are used with the highest frequency across the whole document will be the most important, while LSA (Steinberger and Jezek, 2004) applies SVD to term-frequency matrices to identify sentences which capture the most important topics within a document. These techniques have the advantage that they are not restricted by document length and do not require training data, however, they do not benefit from the semantic knowledge instilled within modern neural models. Despite this, LexRank has been shown to have competitive performance when compared with modern, supervised approaches (Cohan et al., 2018; Pilault et al., 2020).

More recently, PLM-based models have been proposed for supervised approaches to extractive summarisation. These methods generally treat the extractive summarisation task as a classification problem. BERTSum (Liu and Lapata, 2019) is one of the earlier PLM-based models proposed for text summarisation. In this work, BERT is adapted for the extractive and abstractive summarisation tasks by introducing a document-level encoder, and a positional embedding encoder, which enable supervised training to select the sentences to include in the summary, in the context of the surrounding document. MatchSum (Zhong et al., 2020) extends the extractive

[2] https://chat.openai.com/chat

implementation of BERTSum (BERTSumExt) by first predicting candidate sentences to be included a summary using the BERTSumExt model and then training a secondary model to select the best sentences from the candidate list to include in the final summary by maximising similarity of the candidate summary to the overall document. In these works, to overcome the memory implications of using a transformer architecture, the source documents are truncated to 512 tokens. Thus, when used in a long document setting, these methods essentially only use the introduction of the source document to form the predictive summarisation.

Several models have been proposed to extend PLM-based extractive summarisation methods to be suitable for long documents. As mentioned in Section 2.1, Beltagy, et al., (2020) propose Longformer, an autoencoder with an efficient attention mechanism which could be used for this purpose, for example, by replacing BERT in the BERTSum model with Longformer. Zhang et al., (2019a) propose HIBERT, a hierarchical pretraining strategy for long documents where two encoder layers are used: a first which follows BERT, using a pretraining strategy where individual tokens are masked, whilst a second encoder layer is trained by masking and predicting entire sentences. They then fine-tune this PLM on a downstream extractive summarisation task and show it to be a highly effective method for summarising long documents. Grail et al., (2021), Rohde et al., (2021), Xiao and Carenini (2019) and Ruan et al., (2022) similarly propose hierarchical approaches for the extractive summarisation task. Xiao and Carenini (2020) focus on a different aspect of improving models for long document summarisation, exploring methods to reduce redundancy in extractive summaries. These methods have all demonstrated improved performance in long document settings. However, being supervised methods, they all still require large amounts of labelled training data specific to the task and domain for which they are intended to be used.

Modern unsupervised extractive approaches tend to build upon traditional graphical approaches: Liang et al., (2021) uses a PageRank-based model and incorporates a weighting between each sentence and the overall document, as well as weightings between the sentences to better represent documents which contain multiple facts. PacSum (Zheng and Lapata., 2019) and HipoRank (Dong et al., 2021) both use transformer-based embeddings to better represent the input document and incorporate additional information about the sentences' positions when calculating centrality; the latter publication extends this by also incorporating document hierarchy into its calculations. Xu et al., (2020) trains a model to generate sentence and document embeddings hierarchically using unlabelled data, and then selects sentences based on several criteria which use the attention-based embeddings.

2.3 Abstractive summarisation

Research into abstractive methods began relatively recently, with the popularity of research into this area growing in line with the proliferation of deep, neural machine learning models. Abstractive methods generally use encoder-decoder architectures, with early abstractive methods using feed-forward networks or RNNs for their models (Rush et al., 2015; Chopra et al., 2016).

Since the introduction of PLMs, there has been a surge of research into abstractive summarisation. BART (Lewis et al., 2020), T5 (Raffel and al., 2020), and PEGASUS (Zhang et al., 2020a) utilise different pretraining strategies for sequence-to-sequence models which can be fine-tuned for the abstractive summarisation task. LongT5 (Guo et al., 2021), BigBird (Zaheer et al., 2020), and LED (Beltagy, et al., 2020) all contain more efficient attention mechanisms in their encoder-decoder architectures, but are still limited in the number of tokens they can process due to hardware limitations. Whereas for extractive summarisation, the main consideration in regards to memory consumption is the token limit of the source document, for abstractive text summarisation, both the input document length and the target generation length have implications for the memory consumption of the model, thus there is often a trade-off between the two for transformer-based encoder-decoder models.

Akin to the extractive approaches mentioned earlier, there is an abundance of research into abstractive models which extend upon simply fine-tuning a PLM with the abstractive summarisation objective. Liu and Lapata (2019) propose BERTSumAbs and BERTSumExtAbs, which are summarisation models trained with an abstractive objective, where the latter uses weights initialized by pretraining the model with an extractive objective first. Liu and Liu (2021) propose a two-step approach using BART. They first fine-tune BART for the abstractive summarisation task which, given a source article, they use to generate a few canditiate summaries. They then apply a second model which uses contrastive learning with a ROUGE-based objective to select the best summary. As with extractive methods, there is prior literature which explores hierarchical approaches to summarise longer documents. HAT-BART (Rohde et al., 2021) adapts the traditional transformer by incorporating additional hierarchical learning steps within it, and by inserting additional special tokens between sentences. However, HAT-BART still limits the input token length to 3072 and the generation token length to 512 for the long document data sets. Hie-BART (Akiyama et al., 2021) and 'Top Down Transformer' (Pang et al., 2022) are other examples of hierarchical abstractive summarisation models. Other works propose different strategies for dealing with long documents in an abstractive setting. DANCER (Gidiotis and Tsoumakas, 2020) uses the concept of document zoning to split a longer text into smaller sections

and selects only the most relevant sections to feed into their model. DANCER is trained to learn to summarise the short, relevant, document sections, using beam search decoding to generate the summaries. The short summaries are then concatenated to form a longer summary. Liu et al., (2022) propose a similar method, making 'local' predictions from document sections which are then combined into a 'global' document prediction downstream. They experiment with spatial locality (i.e., splitting a document into pages), and discourse locality (i.e., splitting a document into its predefined sections).

Factual inconsistency is a known limitation of abstractive text summarisation (Maynez et al., 2020; Wallace et al., 2021) and several works to date have attempted to address this. A relatively early work exploring approaches to improve factual consistency used a pointer-generator network to copy words directly from the source text (See et al., 2017). Similarly, Mao et al., (2020) selects entities and key phrases from the source document which they require that the generated summary must include. Cao et al., (2020) propose a method to correct factual inconsistencies in a secondary step after the summaries have been generated. Zhu et al., (2021) use knowledge graphs to help guide their summaries to be factual and also have a secondary stage, which is designed to correct factual inconsistencies after the initial generation. Cao et al., (2022) show that some hallucinated facts in abstractive summaries generated using PLMs are in fact correct and are a product of the knowledge learnt during their pretraining. They therefore propose a method to separate hallucinated true facts from incorrect ones. FACTPEGASUS (Wan and Bansal, 2022) is a pretraining strategy proposed to improve factual consistency in downstream tasks such as summarisation.

2.4 Evaluation metrics

ROUGE scoring (Lin et al., 2004) has long been the popular automated metric used for evaluation of text summarisation. ROUGE scoring compares a predicted summary to a reference summary and uses word overlap between the two summaries to calculate their similarity. Thus, to evaluate generated summaries using the ROUGE metric, test data sets which include reference summaries are required. ROUGE-1, -2 and -L F1 scores are the standard metrics reported in the evaluation of text summarisation in literature. ROUGE-1 compares individual word overlap between the predicted and target summaries, ROUGE-2 compares bigram overlap and ROUGE-L compares the overlap of the longest common subsequence. Despite being, by far, the most popular metric for evaluation of automatic text summarisation, ROUGE is well-known to be flawed (Yuan et al., 2021; Huang et al., 2020; Kryściński et al., 2019). This is due to it only assessing whether the same words appear in two summaries, regardless of their ordering. It is unable to capture other

important measures when evaluating the quality of a generated summary, such as coherence, fluency, and factual consistency.

As with advancements in the methods to generate the summaries, modern metrics for evaluation of text summarisation also use PLMs and have been shown to align better with human judgement than ROUGE. BERTScore (Zhang et al., 2019b), is a popular metric for evaluating text summarisation. In a similar vein to ROUGE, this metric also compares the predicted summary to a reference summary, and measures agreement at a token level between the two by calculating the cosine similarity between BERT-based token embeddings. BARTScore (Yuan et al., 2021) can also be used to compare a predicted summary to a reference summary but can additionally be used in a 'reference-free' setting, where the predicted summary is compared to the source document. BARTScore works by calculating the log probability of generating a sequence of text, given a second sequence. The 'reference-free' setting of BARTScore can be used to assess factuality of a predicted summary. Building on this work, T5SCORE (Qin et al., 2022) uses T5-based models and combines the generative approach taken by BARTScore with a discriminative approach - i.e., fine-tuning a model to predict a quality score for a summary using human annotated summaries as training data. FACTGRAPH (Ribiero et al., 2022) builds knowledge graphs of source documents and predicted summaries and compares the two to assess factual inconsistencies. FactCC (Kryściński et al., 2020) is a reference-free metric designed to assess factual consistency of summaries. It uses a BERT-based model which is fine-tuned to make a classification of factual correctness for each sentence of the predicted summary, given the source document. QuestEval (Scialom et al., 2021) takes a question-answering approach, also aiming to measure factual consistency. This method uses T5-based models to generate questions from the entities in each of the predicted summary and source document and assesses whether the other text is able to answer the generated question. It does this by measuring the similarity between an answer generated by a question-answering model and the true answer (i.e., the entity used to generate the question). Several other works also propose question-answering based approaches for measuring factual consistency (Fabbri et al., 2021; Durmus et al., 2020). However, Kamoi et al., (2022) claims that these approaches are flawed as they miss factual errors which are not associated with the predefined entities from which the questions are generated.

In prior literature these metrics have been evaluated on human annotations of short document data sets such as CNN/DM (Hermann, 2015), Newsroom (Grusky et al., 2018), XSUM (Narayan et al., 2018), and FRANK (Pagnoni et al. 2021). Despite only being evaluated on short documents, some of the existing reference-free metrics for evaluating factual consistency were reported to be very computationally expensive. QAGS (Wang et al., 2020a) running on a single

NVIDIA v100 GPU will take over 4 days to process the CNN/DM data set (Nan et al, 2021). Although the automated metrics described here align better with human judgement than traditional metrics such as ROUGE scores, their PLM-based architectures and expensive computations restrict them from being applied to long documents; this is particularly problematic when attempting to assess factual consistency in reference-free settings, where the source documents are often well over the token limits of these PLMs (Koh et al., 2022). This results in them not extending well when being used directly to evaluate factual consistency of long documents (Koh et al., 2022).

3 A hybrid abstractive-extractive approach

3.1 Extractive approaches for long document summarisation

Whilst PLMs have contributed to great advancements in NLP due to their ability to learn semantic knowledge, they have also introduced new limitations which are particularly apparent when processing long documents. They are restricted by the number of tokens that they can process at any one time and the computational cost of fine-tuning their attention mechanisms is expensive. Thus, to achieve text summarisation with PLMs, generally the input document is truncated (Liu and Lapata, 2019; Xu et al., 2020; Zhong et al., 2020). Since summarisation should be able to succinctly capture the meaning of very long documents in a few sentences, the requirement to truncate a document before summarisation is a disadvantage. As a result, recent works have shifted their attention towards addressing the issue of long document summarisation (Xiao and Carenini, 2020; Grail et al., 2021; Rohde et al., 2021; Xiao and Carenini, 2019). However, these are mostly supervised methods, requiring large amounts of labelled training data, which are often unavailable or time-consuming and costly to produce. In this section a hybrid abstractive-extractive unsupervised approach is proposed, where the PLMs are required only to act on short sections of the document at any time, meaning that this method can be extended to any document length. Furthermore, because it is an unsupervised approach, it does not require manually labelled training data for the extractive summarisation task. This method is named GenCompareSum.

To-date, unsupervised methods for text summarisation have generally used graph-based approaches (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Liang et al 2021; Zheng and Lapata, 2019; Dong et al., 2021), the more recent of these using transformer-based embeddings to encode the source text (Zheng and Lapata, 2019; Dong et al., 2021). The method proposed in this section differs from these previous approaches as it does not use a graph-based model, and

instead applies a novel approach: generating and using salient textual fragments to guide the extractive summarisation. Moreover, earlier unsupervised, graph-based methods have been criticised in their ability to effectively represent documents which present multiple facts (Liang et al., 2021). GenCompareSum addresses this by generating multiple salient texts per document, thus enabling it to represent multiple facts per document. Furthermore, this work differs from existing hybrid extractive-abstractive approaches as it uses transformer-based abstractive models for the generation of salient points, but ultimately, generates an extractive summarisation to ensure factual consistency.

3.2 The GenCompareSum model

This work proposes a hybrid abstractive-extractive model called GenCompareSum. This method makes use of transformer-based architectures but is extendable to any document length, can represent multiple facts, and does not require vast amounts of training data. The method is comprised of two steps: GenCompareSum first splits a document into sections of several sentences and walks through them, generating salient textual fragments which represent each section. In the second step, the salient text fragments are used to guide the selection of sentences to form an extractive summary.

In this work, different models for generating salient text fragments are evaluated, which are fine-tuned to either predict queries or document titles that best represent a section of the document. The aim of these models is to capture the key points of a section of text. In the case of models which generate document titles, a document section is entered to the model, and the model predicts titles which it thinks best represent the section of text. In the case of models which generate questions, a document section is used as an input and the model generates questions, the answers to which should be the key points of the input document section.

GenCompareSum uses these generated textual fragments to guide an unsupervised extractive summarisation by calculating the BERTScore similarity between each of the generated texts and each of the sentences in the source document. Since the generated textual fragments aim to represent the key points of the input document, document sentences with the highest similarity to these fragments should be of high importance, and therefore, in GenCompareSum, are used to form the predicted extractive summary.

A representation of GenCompareSum can be seen in Figure 1. In this diagram, the steps are as follows: (a) The document is split into sentences. (b) Sentences are combined into sections of several sentences. (c) Each section is fed into the generative text model to generate several text fragments per section. (d) The questions are aggregated, and redundant questions are removed by

using n-gram blocking. Where aggregation occurs, a count is applied to represent the number of textual fragments which were combined, and this count is used as a weighting going forwards. The highest weighted textual fragments are then selected to guide the summary. (e) The similarity between each sentence from the source document and each selected textual fragment is calculated using BERTScore. (f) A similarity matrix is created from the scores calculated in the previous step. These are then summed over the textual fragments, weighted by the values calculated in step (d), to give a score per sentence. (g) The highest scoring sentences are selected to form the summary. The code for GenCompareSum has been made publicly available[3].

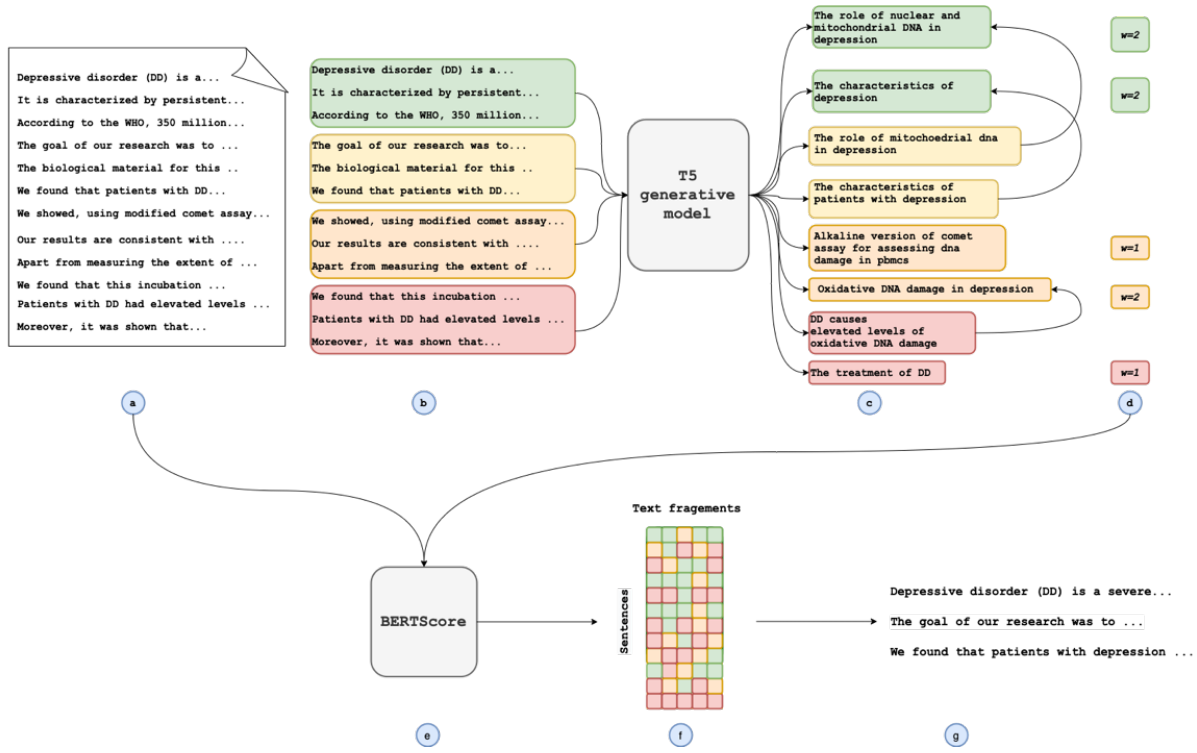


Figure 1. An overview of the GenCompareSum process, from the original document (a) to the extractive summary (g): (b) illustrates the document being split into sections, (c) the generation of salient text fragments from each section, (d) the weightings given by n-gram blocking, (e) and (f) the weighted BERTScore calculation between the sentences from the original document and generated salient text fragments.

3.2.1 Text Splitting

[3] https://github.com/jbshp/GenCompareSum

Given a document D, it is first split it into sentences s, such that $D = \{s_1, \dots, s_n\}$, using the Stanford CoreNLP software package (Manning et al., 2014). The sentences are then combined into document sections, p, of x sentences, i.e., $D = \{p_1, \dots, p_m\}$; $m = ceil\left(\frac{n}{x}\right)$. A design decision was made not to use any predefined sections already existing within the documents as, during inspection of the data sets, it was found that the documents sections were not always extracted well. Splitting the document into a consistent number of sentences per section removes the requirement for high quality text extraction into document sections. The number of sentences x used to form the short text sections was decided via experimentation on the validation data sets, as detailed in section 3.3.2.

3.2.2 Salient Text Generation

T5 (Raffel and al., 2020) is a sequence-to-sequence model, pretrained on a cleaned and pre-processed version of the Common Crawl[4] data set – a data set consisting of textual content scraped from the internet. T5-based models have been shown to be high performing sequence-to-sequence models across a range of generative tasks, from question generation (Nogueira and Lin, 2019), to graph-to-text generation (Ribeiro et al., 2021), generative common-sense reasoning (Yuchen Lin et al., 2020), to abstractive text summarisation (Goodwin, 2020). The T5 model uses an encoder-decoder architecture and is pretrained via an unsupervised task in which 15% of tokens are masked; the masked words can be individual words or a span of words; the target of the training objective is to predict these masked words, given the un-masked tokens and their respective positions. For downstream tasks, the pretrained T5 model is fine-tuned using pairs of input and output sequences. A diagram of the T5 architecture and its pretraining and fine-tuning settings can be seen in Figure 2. In this figure, the left diagram shows the unsupervised pretraining task, in which a tokenized text containing masked spans is passed to the encoder and the output target of the decoder is the prediction of the masked spans. The right diagram shows the supervised downstream task, where the pretrained model is fine-tuned on pairs of tokenized sequences for a query generation task.

[4] https://commoncrawl.org

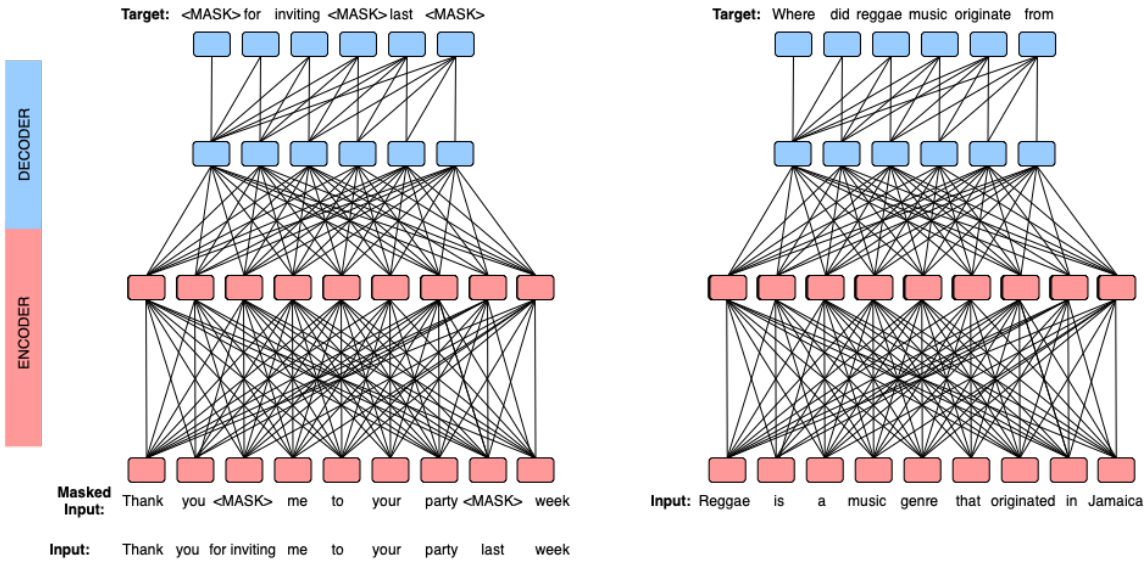


Figure 2. A diagram of the T5 architecture and its pretraining and fine-tuning settings.

For the generation of the textual fragments, first, an existing open source T5-based model, docTTTTTquery (Nogueira and Lin, 2019), was evaluated. This is a question generation model of the general domain, which was fine-tuned on a question-answer data set generated from Bing's[5] search query logs called the MS-MARCO data set (Bajaj et al., 2018). Surita et al., (2020) showed this pretrained model to be effective at generating questions for long, biomedical texts.

Second, the approach taken by Nogueira and Lin (2019) was followed and a T5 model was fine-tuned on data consisting of long-answer - query pairs curated from the biomedical domain. For this, four biomedical data sets were combined to make a large corpus of text-question pairs, where the questions can be answered by the long textual input. From the BioAsq data set (Nentidis et al., 2021) 3,433 'ideal answer'-question pairs were used, 2,720 text-question pairs from COVID-QA (Möller et al., 2020) were included (where the paragraph containing the answer was used as the textual input), 61,244 context-question pairs from PubMedQA (Jin et al. 2019) were used (where the 'context' refers to the abstract without its 'conclusion' section), and finally 27,722 long answer-question pairs from the MASH-QA data set (Zhu et al., 2020) were included. The t5-base[6] model was loaded and fine-tuned on this data set for 5 epochs, with a batch size of 8. In this work, this model is referred to as 't5-med-query'.

[5] https://www.bing.com
[6] https://huggingface.co/t5-base

A third open source T5-based model was also included in the evaluation, which had been trained with abstract-title pairs from the scientific domain[7]. The approach of fine-tuning a PLM to predict a title given an input document has shown to be an effective method for proxying highly abstractive summaries (Cachola et al., 2020). The model was applied to the GenCompareSum method by generating potential 'titles' for each document section. In this work, this model is referred to as 't5-s2orc-title'.

3.2.3 N-gram blocking

N-gram blocking is a technique which is applied to reduce redundancy and improve coverage in summarisation models (Liu and Lapata, 2019). In the GenCompareSum method, N-gram blocking is applied to the generated textual fragments, resulting in only a subset of the original generated texts remaining: $T^* \subseteq T$, where $T^* = \{t_1, \dots t_{l,l \leq mk}\}$. Where generated texts are removed by applying this technique, a count is kept of how many times a similar textual fragment was seen before n-gram blocking. This results in each remaining generated textual fragment having an associated count representing the number of similar textual fragments before n-gram blocking. These counts are then treated as weights, which can be described by $w = \{w_1, \dots, w_l\}$, such there is one weight associated with each generated textual fragment remaining after n-gram blocking. A visualisation of this can be seen in steps c and d of Figure 1. The top $q; q < l$ generated texts are then selected after ordering by the weight.

3.2.4 Text vector comparison

BERT-based comparisons have been shown to outperform traditional sentence comparative metrics like TF-IDF when used in unsupervised summarisation tasks (Dong et al., 2021). Furthermore, they have been shown to align better with human judgement of text similarity than n-gram matching approaches during evaluation, likely due to their ability to match based on semantic meaning and their penalisation of word reordering which changes a text's meaning (Zhang et al,. 2019b). BERTScore (Zhang et al., 2019b) uses BERT-based token embeddings, calculates the cosine similarity between them and uses greedy matching to match each token in the first text to its most similar token in the second; these scores are averaged across the sentences to give precision, recall and F1 scores which quantify the similarity between two texts. The BERTScore between each sentence in the document and each selected generated text fragment is

[7] https://huggingface.co/doc2query/S2ORC-t5-base-v1

calculated. The score is weighted by w, the count representing the number of textual fragments which were aggregated during n-gram blocking to give:

$$score_i = \sum_{1}^{Z=q} w_z * BERTScore(s_i, t_z) \quad (2)$$

The sentences with the highest score are selected to form the predicted summary S^* and they are reordered back into the sequence which they originally appeared within the source document.

3.3 Experimental set-up

3.3.1 Data sets

The efficacy of the developed hybrid summarisation model was evaluated with four publicly available long document data sets from the biomedical and scientific domains. All four data sets consist of full-article research papers and their corresponding abstracts. In line with previous literature, their abstracts were used as the target summaries. The data sets included in the experiments are CORD-19 (Wang et al., 2020b), PubMed and ArXiv (Cohan et al. 2018), and S2ORC (Lo et al., 2020). The CORD-19 data set used is the version released on 2020-06-28, containing 57,037 articles relating to COVID-19. The S2ORC data set is a large corpus of scientific literature across several domains; from the subset of articles tagged as being from the biological and biomedical domains, 63,709 were randomly sampled. The PubMed and ArXiv data sets are from the biomedical and scientific domains respectively.

For the S2ORC and CORD-19 data sets, the data set was split by sampling randomly to create training/validation/test sets using the ratio 75/15/10. For the PubMed and ArXiv data sets, the train/validation/test sets given in the resources associated with the original paper were used.

Since most previous literature using transformer-based summarisation models either evaluates them on short or truncated texts (Liu and Lapata, 2019; Xu et al., 2020; Zhong et al., 2020), short data sets were also created and used in this evaluation for comparison. These data sets were created by truncating documents to the end of the sentence which contains their 512th token. The models were evaluated both on the short and full-text versions of the four data sets described above. Table 1 gives, for each data set, the mean number of tokens and sentences for the documents and their target summaries.

As training is not required for unsupervised models, for these methods only the test data sets were used. BERTSumExt (Liu and Lapata, 2019) was implemented as a strong supervised baseline for comparison. To train this method, the training and validation data sets were used to train the model and to select the best performing epoch for evaluation on the test set.

Data set	Instances			Input length: Truncated		Input length: Full		Target length	
	Train	Val	Test	Tokens	Sentences	Tokens	Sentences	Tokens	Sentences
PubMed	117108	6631	6658	525	20	3209	124	208	9
S2ORC	47474	9490	6631	523	19	4312	154	250	9
CORD-19	31505	6299	4202	525	18	5240	206	232	8
ArXiv	202917	6436	6440	528	20	6515	249	279	11

Table 1. A description of the four data sets used in the extractive summarisation experiments.

3.3.2 Parameter selection

To select the optimal parameters for the GenCompareSum models, a seeded random sample of 1000 articles from the PubMed validation data set were selected. Different combinations of the parameters were experimented with, details of which can be found in Table 2.

Different methods were compared for calculating the similarity between the generated salient text fragments and the document sentences. BERTScore, a method which uses word embeddings to calculate the similarity between texts, was compared with two other methods to calculate the similarity between texts using sentence embeddings. Sentence Transformers (Reimers and Gurevych, 2019) is trained with a triplet / siamese bert-based architecture and a training objective designed to minimise distances between similar sentences. This method was implemented with their python package[8]. Their suggested base model for the general domain 'all-mpnet-base-v2' and a model trained to calculate document-level similarity for scientific documents 'allenai-specter' (Cohan et al., 2020) were compared. SimCSE (Gao et al., 2021), a method which generates sentence embeddings with a model trained using contrastive learning was also implemented and evaluated. For this method, the general-domain base model which is suggested to be the most performant in SimCSE's documentation[9] was used. For the BERTScore method, base models from the general domain were evaluated, namely 'bert-base-uncased'[10], and a base model pretrained on data from the scientific domain (Beltagy et al., 2019), 'allenai/scibert_scivocab_cased'. For this experiment, GenCompareSum was implemented with the s2orc-title generative model and the different methods for the text comparison step were evaluated. In Table 3, the results for the ROUGE metrics are reported, calculated for the extractive summarisation task on the PubMed 'Short Document' data set. In Table 3, and tables of results

[8] https://github.com/UKPLab/sentence-transformers
[9] https://github.com/princeton-nlp/SimCSE
[10] https://huggingface.co/bert-base-uncased

throughout this work, bold font indicates the best result and underlined text indicates the second-best result. The similarity method is given in the table's first column, with the base model used in its implementation given in brackets. Despite the data set on which the different methods were evaluated being of the biomedical domain, BERTScore, implemented with a 'bert-base-uncased' PLM, outperformed all other methods compared. This is a promising result as it suggests the method is more likely to generalise to other domains.

Parameter	**Parameter Definition**	**Experimental Range**	**Selected Parameter**
T5 model temperature	Controls randomness of generative text model predictions	0.2-1	0.5
T5 input size (x)	# of sentences in input section	2-12	4
T5 predictions (k)	# of salient texts generated per section	2-6	3
T5 prediction n-gram blocking	# of consecutive word matches used to determine whether a generated text should be removed due to redundancy	No n-gram blocking, n=3, n=4	4
T5 generated texts for comparison (q)	# of generated texts used for comparison to the original document sentences	4-12	10
BERTScore embedding model	Base model used in BERTScore package for word-embedding comparison	bert-base-uncased[11], facebook/bart-large-mnli[12], allenai/longformer-large-4096[13], allenai/scibert_scivocab_uncased[14]	bert-base-uncased
Score weighting	Option to multiply scores by frequency of question occurrence	True/False	True
Sentence selecton n-gram blocking	# of consecutive word matches used to determine whether a selected sentence should be removed due to redundancy	No n-gram blocking, n=3, n=4	4

Table 2. Parameters experimented with, and selected for use, in the GenCompareSum models.

[11] https://huggingface.co/bert-base-uncased
[12] https://huggingface.co/facebook/bart-large-mnli
[13] https://huggingface.co/allenai/longformer-base-4096
[14] https://huggingface.co/allenai/scibert_scivocab_uncased

Text similarity method	R1	R2	RL
BERTScore (bert-base-uncased)	**39.19**	**14.35**	**35.65**
BERTScore (allenai/scibert_scivocab_cased)	37.78	13.40	34.45
SentenceTransformer (all-mpnet-base-v2)	<u>39.03</u>	<u>14.20</u>	<u>35.45</u>
SentenceTransformer(allenai-specter)	38.20	13.41	34.67
SimCSE (princeton-nlp/sup-simcse-roberta-large)	38.62	13.73	35.07

Table 3. A comparison of different methods for calculating text similarity between generated salient texts and the document's sentences.

3.3.3 Implementation details

All experiments requiring GPUs were run on NVIDIA Quadro RTX 6000 hardware. The results are reported in terms of ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004), which were calculated using the pyrouge[15] python package.

Several extractive text summarisation methods were compared across the short and full-text versions of the four scientific data sets. For the short-text data sets, 6 sentences were selected to generate the predictive summary. Results are given on a short-text summary for a fair comparison against supervised methods, which are restricted by the length of document that they can easily summarise. For the full-text articles, the number of sentences that were selected for the predictive summary is the same as the average number of sentences in the target summaries for a given data set, shown in Table 1. e.g., for the PubMed data set, 9 sentences were selected to summarise the full text article.

ORACLE summaries indicate the upper bound for extractive text summarisation. ORACLE summaries were generated by adapting code from Liu and Lapata (2019), which applies greedy sentence selection to maximise ROUGE scores. As baseline methods for comparison, the LEAD method, taking the first n sentences to form the summary, and the RANDOM method, taking a random sample of n sentences to form the summary, were implemented. Unsupervised extractive methods, LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004) and SumBasic (Nenkova and Vanderwende, 2005), were also implemented as baselines, all of which were implemented using the sumy[16] package. LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are both graph-based models, based on Google's PageRank algorithm (Brin and Page, 1998), which assume that the sentences with the highest centrality are

[15] https://github.com/bheinzerling/pyrouge
[16] https://github.com/miso-belica/sumy

the most important and use these to form a summary. SumBasic simply assumes that sentences containing the words which are used with the highest frequency across the whole document will be the most important. Additionally, BERTSumExt (Liu and Lapata, 2019), a state-of-the-art supervised method using BERT-based transformer models was implemented for comparison. For evaluation on the short data sets, where the documents were truncated at the end of the sentence containing the 512th token, the original implementation of BERTSumExt was used without modification to train and evaluate the models. For the full-text article, this code was adapted (and is thus denoted BERTSumExt*) to cycle through the article in 512 token-length blocks and predict the best sentences to select from across this cycle. However, due to hardware limitations and the computational intensity of the attention calculation, truncation of the document to its first 1024 tokens was required to evaluate this method. Lastly, GenCompareSum was implemented, and its performance was evaluated using the different generative text models described in Section 3.2.2: docTTTTTquery, t5-med-query, and t5-s2orc-title.

3.4 Evaluation of the unsupervised hybrid abstractive-extractive approach

3.4.1 Automatic Evaluation

The results of the unsupervised hybrid abstractive-extractive method on the extractive summarisation task are reported in Table 4.

For the short documents, GenCompareSum (t5-s2orc-title) performed best across three out of four of the data sets, and second-best for the fourth data set. There was no clear 'second-best' model out of the methods compared for the short data sets. Interestingly, for the S2ORC data set, the method that outperformed all others was LEAD, i.e., taking the first sentences from the document as the predictive summary. However, in evaluation of the full-text version of the S2ORC data set, it did not hold that LEAD was the best method, and it was outperformed by several other methods.

For the long document data sets, GenCompareSum (t5-s2orc-title) outperformed all other unsupervised models. A strong unsupervised baseline, LexRank, has been shown in prior literature to give competitive performance when compared to supervised approaches (Cohan et al., 2018; Pilault et al., 2020). In-line with these works, LexRank was shown to be the best-performing unsupervised method after GenCompareSum.

The method proposed in this work, GenCompareSum (t5-s2orc-title), outperformed LexRank by an average of $\Delta R1, \Delta R2, \Delta RL$ of 2.35, 1.47, 2.27 across the four data sets. A slight performance increase over BERTSumExt* (the strong supervised baseline which was adapted to run over longer documents) can be seen, with an $\Delta R1, \Delta R2, \Delta RL$ of 0.36, 0.56, 0.06 across the

four data sets. Given that GenCompareSum is unsupervised, and therefore does not require labelled training data and can be extended to any document length, it is arguably a favourable method.

Considering the different implementations of GenCompareSum, as expected, the results show that using a generative model fine-tuned on in-domain data gives notable performance increases. Table 4 shows that on the CORD-19 biomedical data set, the $\Delta R1$ between an out-of-domain query generation model (docTTTTTquery) and a query generation model trained on biomedical data (t5-med-query) was as high as 3 and 2.49 for the short and long articles respectively. However, for the ArXiv data set, which consists of predominantly physical and computer science related research articles, the performance decreased when using the t5-med-query generative model instead of the general domain docTTTTTquery model.

The best-performing GenCompareSum model, t5-s2orc-title, uses a generative PLM fine-tuned on document-title pairs from the S2ORC data set to guide the extractive summarisation. In many ways, a title can be considered as a highly abstractive summarisation (Cachola et al., 2020). A major advantage of this finding is that, although it does require training data to fine-tune this generative model, document-title pairs are readily available across many domains, thus a model can easily be trained for a specific task without needing extensive manual labelling effort. Furthermore, this model, although fine-tuned on biomedical and scientific data, is fine-tuned on a very broad range of documents within these fields. This work demonstrates that, despite the broad coverage of fields in its training data, it performs very well when applied to data from a more specific domain, e.g., biomedicine in the PubMed and CORD-19 data sets.

Lastly, it can be observed that there are big differences in ORACLE scores between the short and full text data sets. Although GenCompareSum outperformed all other methods evaluated for both short and full text documents, the gap between the best predictive scores in the experiments and the ORACLE upper bound is large for long documents, suggesting that much more research could be done in this space. Furthermore, based on this observation, one can hypothesise that predicting summaries from short documents is a significantly easier task than doing the same for long documents. This is supported by TextRank performing worse on the long documents than on the truncated versions. This may be explainable by the fact that there are much fewer sentences to choose from within a shorter document (approximately 32% of all sentences in truncated documents are selected to form the summary vs 5% of sentences in full documents), thus less room for error.

Model	PubMed			S2ORC			CORD-19			ArXiv		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Short Document												
ORACLE	47.27	22.85	43.20	49.29	25.42	45.52	43.47	17.75	39.28	47.29	18.49	41.90
RANDOM	34.98	10.82	31.37	34.69	11.06	31.30	31.64	7.91	28.10	34.53	8.88	30.26
LEAD	35.39	12.07	32.28	**40.50**	**16.72**	**37.68**	34.80	10.17	31.67	34.35	8.75	30.61
LexRank	38.48	13.05	34.92	39.44	14.57	36.13	35.65	10.17	32.11	38.98	11.44	34.64
TextRank	38.15	12.99	34.77	40.17	14.84	36.63	36.25	10.61	32.53	37.97	11.58	33.53
SumBasic	36.11	11.06	32.67	35.99	11.99	32.87	33.63	8.82	30.22	37.14	9.83	33.06
BERTSumExt	38.78	14.47	35.43	39.41	16.14	36.38	34.68	10.34	31.42	39.36	11.74	35.09
GenCompareSum (docTTTTTquery)	37.82	13.12	32.41	38.31	14.27	35.17	33.77	9.73	30.66	38.59	11.49	34.50
GenCompareSum (t5-med-query)	38.54	13.67	35.06	38.96	14.78	35.80	36.77	11.24	33.29	38.92	11.59	34.76
GenCompareSum (t5-s2orc-title)	**39.19**	14.35	**35.65**	40.16	15.84	36.91	**36.84**	**11.35**	**33.35**	**39.66**	**12.30**	**35.38**
Long Document												
ORACLE	61.76	36.78	57.61	64.11	39.21	60.16	59.10	32.09	54.63	60.16	32.17	54.97
RANDOM	37.26	11.19	33.66	37.12	10.23	33.73	33.37	7.70	29.98	34.20	8.70	30.64
LEAD	37.23	11.11	33.67	40.50	16.72	37.68	34.61	10.17	31.68	34.70	10.27	31.37
LexRank	41.02	15.83	37.18	42.60	15.84	38.97	39.50	12.65	35.68	33.94	12.09	30.62
TextRank	34.53	12.98	30.99	36.58	13.23	33.10	32.99	10.39	24.47	26.57	9.20	23.74
SumBasic	40.61	12.42	36.54	36.63	10.43	33.68	33.88	8.24	30.86	33.18	7.75	30.29
BERTSumExt*	41.87	16.01	**38.51**	**43.56**	**17.85**	**40.40**	38.95	12.17	35.48	**40.65**	14.01	**36.89**
GenCompareSum (docTTTTTquery)	40.54	14.77	36.83	40.78	14.24	37.43	36.84	11.19	33.51	38.19	12.76	34.55
GenCompareSum (t5-med-query)	41.60	15.67	37.79	41.84	15.10	38.35	39.33	12.31	35.74	37.17	11.97	33.95
GenCompareSum (t5-s2orc-title)	**42.10**	**16.51**	38.25	43.39	16.84	39.82	**41.02**	**13.79**	**37.25**	39.96	**15.15**	36.19

Table 4. Results of the extractive summarisation task on the PubMed, ArXiv, S2ORC and CORD-19 data sets. Bold font indicates the best results and underlined font indicates the second-best results.

3.4.2 Qualitative analysis

Table 5 shows a randomly sampled PubMed document, the associated generated salient fragments, and the predicted extractive summary given by each of the three GenCompareSum methods. The gold summary (the document abstract) is also given for comparison. Table 6 gives the same for a randomly sampled document from the ArXiv data set. In this subsection, the

differences between the texts generated by the various T5-based models are compared and hypotheses are given about how they may influence the extractive summary.

The docTTTTTquery model produced questions which were relatively general and implied little biomedical knowledge when given the PubMed document as input., producing textual fragments such as “what is nlrp3”. Interestingly, it did manage to produce more complex texts from sections of the ArXiv data set, such as: “what is the contribution of the spiral arm to the resonant structure in the solar neighborhood?”.

In comparison, the t5-med-query model, whilst also generating questions, better encapsulated biomedical concepts when given a document from the PubMed data set, e.g., “what is the role of nuclear and mitochondrial dna damage and repair in people with depression?”. However, in line with the ROUGE results given in Section 4.1, it seemed to perform less well on out-of-domain (i.e., scientific rather than biomedical) literature, and appeared to default to a more general question generation model, generating texts for the ArXiv document such as “what is the effect of a spiral arm?”.

The t5-s2orc-title model generated texts which read much more like very short, highly abstractive summaries. E.g., for the PubMed article, it generated the textual fragment: “the role of the nuclear and mitochondrial dna in depression” and for the ArXiv article it generated: “the spiral arm contribution to the resonant structure of the solar neighborhood”.

Although outperformed by the title-generation model t5-s2orc-title in the automatic evaluation, on analysis of the generated textual fragments, the query generation models did seem to effectively represent the important facts from an article, especially in the biomedical domain. It is hypothesised that the use of BERTScore to calculate the similarity between salient texts and document sentences favours the title generation model due to it calculating the similarity between words in different texts and not being designed to answer questions. In future work, it would be interesting to experiment further with the combination of the query generation models and extractive question answering approaches for the extractive summarisation task.

3.5 Discussion of GenCompareSum

In this section, existing approaches for extractive summarisation of long documents was discussed and an unsupervised extractive method for the summarisation of long documents using PLMs was developed. Experiments were conducted on long document data sets from the biomedical and scientific domains and the results showed that the unsupervised method proposed, GenCompareSum, outperformed both strong supervised and unsupervised baselines on long and short documents. This method can be extended to any length of document and does

not require a corpus of annotated training data for the summarisation task. Furthermore, it was shown that the best-performing model used title-document pairs for the generative task, which are readily available across many domains without the need for manual labelling effort. Future work could be done to evaluate different PLMs for the generation task, including different base models, such as BART (Lewis et al., 2020). This method could also be extended to, and evaluated on, articles from other, non-scientific, domains. An additional interesting direction for this work would be to combine it with the concept of zoning, which is discussed in Section 5.

PubMed Sample Document and Predictions	
PubMed Sample Document	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4329942/
PubMed Sample Abstract (Target Summary)	depressive disorder (dd) , including recurrent dd (rdd) , is a severe psychological di sease , which affects a large percentage of the world population . although pathogene sis of the disease is not known , a growing body of evidence shows that inflammation together with oxidative stress may contribute to development of dd . since reactive o xygen species produced during stress may damage dna , we wanted to evaluate the ex tent of dna damage and efficiency of dna repair in patients with depression. material / we measured and compared the extent of endogenous dna damage single - and doubl e - strand breaks , alkali - labile sites , and oxidative damage of the pyrimidines and p urines in peripheral blood mononuclear cells isolated from rdd patients (n = 40) and healthy controls (n = 46) using comet assay . we also measured dna damage evoked by hydrogen peroxide and monitored changes in dna damage during repair incubation . we found an increased number dna breaks , alkali - labile sites , and oxidative modif ication of dna bases in the patients compared to the controls . exposure to hydrogen p eroxide evoked the same increased damage in both groups . examination of the repair kinetics of both groups revealed that the lesions were more efficiently repaired in the controls than in the patients. the first time we showed that patients with depression , compared with non - depresses individuals , had more dna breaks , alkali - labile sites , and oxidative dna damage , and that those lesions may be accumulated by impairme nts of the dna repair systems . more studies must be conducted to elucidate the role of dna damage and repair in depression .
Salient Texts - GenCompareSum (docTTTTTquery)	• what is nlrp3 • how long does it take for dna damage to be repaired • what is the oxidative modification of purines • what is the main activator of nlrp3 • what is the damage caused by dna repair • what is the role of mitochondrial dna in depression • what is oxidative damage in dna • does oxog cause depression • what is the dna damage response • what is the oxidative damage of pyrimidines and purines?
Predicted Summary - GenCompareSum (docTTTTTquery)	• since the findings described above are inconsistent , we wanted to determine if the oxidative modification of purines , like 8 - oxog , and pyrimidines are present in a higher degree in patients with depression than in controls . • to achieve these objectives , we measured and compared the extent of endog enous dna damage single - and double - strand breaks , alkali - labile sites , a nd oxidative damage of the pyrimidines and purines in pbmcs isolated from dd patients and healthy controls . • we evaluated the level of basal endogenous dna damage by subjecting pbmc s to comet assay procedure immediately after isolation from blood . • moreover , we estimated the extent of oxidative dna damage by employing modified comet assay with 2 glycosylases : nth removing oxidized pyrimidi nes and hogg1 excising oxidized purines . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • the goal of our research was to examine the susceptibility of rdd patients to dna damage induced by oxidative stress by measuring the level of endogeno us dna damage , including oxidative dna damage , the amount of dna damag e induced by h2o2 , and efficiency of dna damage repair in the patients as co mpared to the controls without psychological disturbances . • apart from measuring the extent of endogenous dna damage , we also estima ted the amount of dna damage induced by the incubation of pbmcs with h2o 2 and efficiency of its repair .

	• additionally , we monitored the repair efficiency of the induced dna damage • moreover , nlrp3 inflammasome , activation of which was detected in the pat ients pbmcs , was also found to inhibit dna repair after induction of oxidativ e stress .
Salient Texts - GenCompareSum (t5-med-query)	• what was the purpose of the study? • what is the alkaline version of the comet assay? • what is the effect of pbmcs on basal endogenous dna damage? • what is the incubation time for dna repair? • what is the role of nuclear and mitochondrial dna damage and repair in peop le with depression? • is it possible to study the susceptibility of rdd patients to dna damage induce d by oxidative stress? • what is recurrent depressive disorder? • what is the association between 8 - oxog and depression in japanese office w orkers? • which is the most versatile nlr? • what enzymes are bifunctional glycosylases?
Predicted Summary - GenCompareSum (t5-med-query)	• moreover , we also wanted to know if the patients have elevated levels of ot her kinds of dna damage , such as strand breaks . • we evaluated the level of basal endogenous dna damage by subjecting pbmc s to comet assay procedure immediately after isolation from blood . • figure 2 shows basal endogenous dna damage and the damage induced after 10 - min incubation with 20 m h2o2 in pbmcs isolated from the patients and controls without psychiatric disturbances . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • figure 5 compares basal endogenous dna damage and the level of this param eter at the end of the repair incubation in pbmcs of the patients and the contr ols measured by the alkaline version of comet assay . • the goal of our research was to examine the susceptibility of rdd patients to dna damage induced by oxidative stress by measuring the level of endogeno us dna damage , including oxidative dna damage , the amount of dna damag e induced by h2o2 , and efficiency of dna damage repair in the patients as co mpared to the controls without psychological disturbances . • apart from measuring the extent of endogenous dna damage , we also estima ted the amount of dna damage induced by the incubation of pbmcs with h2o 2 and efficiency of its repair . • additionally , we monitored the repair efficiency of the induced dna damage . • there is a need for further studies to define the role of nuclear and mitochond rial dna damage and repair in people with depression , and their implications for clinical outcome .
Salient Texts - GenCompareSum (t5-s2orc-title)	• dna damage in patients with depression. • oxidative dna damage in depression • the oxidative dna damage in patients with renal failure • activation of nlrp3 by oxygen species in pbmc patients. • activation of mitochondrial nlrp3 in patients with pbmcs. • urinary 8-oxog in japanese office workers • the use of the alkaline version of comet assay for assessing dna damage in p bmcs • the role of the nuclear and mitochondrial dna in depression. • the role of the dna repair rate in the repair of pbmcs in patients with squamo us cell carcinoma.
Predicted Summary - GenCompareSum (t5-s2orc-title)	• in agreement with this , activation of nlrp3 in pbmcs of the patients was acco mpanied by increased lipid peroxidation , which can be attributed to increas ed oxidative stress and elevated mitochondrial ros (mtros) production .

	• moreover , we induced oxidative dna damage in those pbmcs by incubating t hem with hydrogen peroxide , measured the kinetics of removing of such da mage , and compared the results between the patients and the controls . • we evaluated the level of basal endogenous dna damage by subjecting pbmc s to comet assay procedure immediately after isolation from blood . • figure 2 shows basal endogenous dna damage and the damage induced after 10 - min incubation with 20 m h2o2 in pbmcs isolated from the patients and controls without psychiatric disturbances . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • it is possible that increased oxidative dna damage occurs only in patients wit h more severe forms of depression , or in later stages of the disease develop ment . • these results indicate that in the patients , oxidative dna damage is less effici ently removed than in the controls . • moreover , nlrp3 inflammasome , activation of which was detected in the pat ients pbmcs , was also found to inhibit dna repair after induction of oxidativ e stress . • for the first time , we showed that patients with depression had elevated leve ls of dna breaks , alkali - labile sites , and oxidative dna damage , and that th ese lesions may be accumulated by impairments of dna repair pathways .

Table 5. Comparison of salient texts and extractive summaries generated by different implementations of GenCompareSum on an article sampled from the PubMed data set.

ArXiv Sample Document and Predictions	
ArXiv Sample Document	https://arxiv.org/abs/0906.4682
ArXiv Sample Abstract (Target Summary)	we study the phase space available to the local stellar distribution using a galactic pot ential consistent with several recent observational constraints . we find that the induced phase space structure has several observable consequences . the spiral arm contribution to the kinematic structure in the solar neighborhood may be as important as the one produced by the galactic bar . we suggest that some of the stellar kinematic groups in the solar neighborhood , like the hercules structure and the kinematic branches , can be created by the dynamical re sonances of self - gravitating spiral arms and not exclusively by the galactic bar . a structure coincident with the arcturus kinematic group is developed when a hot stel lar disk population is considered , which introduces a new perspective on the interpre tation of its extragalactic origin . a bar - related resonant mechanism can modify this kinematic structure . we show that particles in the dark matter disk - like structure predicted by recent lcd m galaxy formation experiments , with similar kinematics to the thick disk , are affect ed by the same resonances , developing phase space structures or dark kinematic grou ps that are independent of the galaxy assembly history and substructure abundance . we discuss the possibility of using the stellar phase space groups as constraints to no n - axisymmetric models of the milky way structure .
Salient Texts - GenCompareSum (docTTTTTquery)	• what is the role of the bar in the local kinematic structure • what is the effect of the non axisymmetric galactic structure on the solar nei ghborhood kinematic distribution? • what is the shape of the solar structure at @xmath27 • what is the structure of the hercules branch • what is the effect of a spiral arm • what is the hercules structure • how does the kinematics of the disk affect the galaxy? • which of the following structure is a contribution to the solar neighborhood kinematics? • what type of spiral arm is used to measure observations made in the solar ne ighborhood • what is the contribution of the spiral arm to the resonant structure in the sola r neighborhood?
Predicted Summary - GenCompareSum (docTTTTTquery)	• however , it is unclear whether there is any dependence of the induced local solar neighborhood kinematics on the detailed galactic structure . • in order to study the effect of the non - axisymmetric galactic structure on th e solar neighborhood kinematic distribution , we have performed numerical i ntegrations of test particle orbits on the galactic plane , adopting the initial c onditions discussed in sect . • the induced kinematic distribution at the end of the simulation is studied by considering the particles inside a circle of radius @xmath7 centered at the so lar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood . • with these initial conditions , we can study the relatively rapid induced effec ts of the non - axisymmetric component on the local kinematics . • we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • in our simulations the positions of these kinematic arches are modified when the bar is added to the model . • furthermore , these simulations show the important role of the bar in the dev elopment of the local kinematic structure . • the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .

Salient Texts - GenCompareSum (t5-med-query)	• what is the effect of dark matter kinematics on the bar - and spiral arm - ind uced phase space structure? • what is the main argument of @xcite? • what is the structure of the hercules? • what is the solar neighborhood? • what is the kinematic distribution of the particles? • what is the relationship between spiral arms and stellar behavior? • what is the galactic potential? • what is the required condition for a thick disk? • what is the difference between ic3 and ic2? • why is the observed velocity field a useful parameter for predicting the beha vior of galaxies?
Predicted Summary - GenCompareSum (t5-med-query)	• however , it is unclear whether there is any dependence of the induced local solar neighborhood kinematics on the detailed galactic structure . • moreover , the initial conditions hardly consider the evolution of the mw . • the induced kinematic distribution at the end of the simulation is studied by considering the particles inside a circle of radius @xmath7 centered at the so lar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood . • we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • in our simulations the positions of these kinematic arches are modified when the bar is added to the model . • another unexpected aspect of the bar - and spiral arm - induced phase space structure is the effect on the local dark matter kinematics . • the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • the main differences to previous studies are the arm force contrast and force field shape? • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .
Salient Texts - GenCompareSum (t5-s2orc-title)	• dark matter kinematics in the solar neighborhood • a note on the arcturus structure in a $xmath26$ plane • dark kinematic groups in the dark disk • the spiral arm contribution to the resonant structure of the solar neighborhoo d • the birth of stars in the disk with small velocity dispersion • the solar neighborhood kinematics and the spiral arms • spiral arms in the mw-type galaxies • the hercules branch of a galactic model using only a bar • theoretical study of the bar and spiral arm perturbations in the xci model • dark matter currents in the galactic dark disk
Predicted Summary - GenCompareSum (t 5-s2orc-title)	• in @xcite we presented a study of the solar neighborhood kinematic groups using a sample of 24,190 stars . • lastly , we investigate effects on the local dark matter kinematics , in particul ar in the disk - like dark matter structure recently predicted by lcdm models . • the induced kinematic distribution at the end of the simulation is studied by considering the particles inside a circle of radius @xmath7 centered at the so lar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood . • we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • another unexpected aspect of the bar - and spiral arm - induced phase space structure is the effect on the local dark matter kinematics .

	• our results show that these models generate dark matter currents inside the g alactic dark disk . • the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • we show that the galactic non - axisymmetric potential develops dark kinem atic groups in the dark disk predicted in cosmological simulations of galaxy formation . • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .

Table 6. Comparison of salient texts and extractive summaries generated by different implementations of GenCompareSum on an article sampled from the ArXiv data set.

4 Injecting external knowledge into summarisation

4.1 Domain-specific knowledge in NLP

Whilst short document summarisation data sets are largely from general domains such as news articles (Hermann et al., 2015; Grusky, et al., 2018; Narayan, et al., 2018), long document summarisation data sets are often highly domain specific. Scientific research articles are common examples of long document data sets (Wang et al., 2020b; Cohan et al. 2018; Lo et al., 2020), in addition to data sets of patent data (Sharma et al., 2019), US legislative documents (Kornilova and Eidelman, 2019), and government reports (Huang et al., 2021). Since long document data sets are generally more domain specific, they contain a relatively high proportion of out-of-vocabulary words for PLMs trained with data of the general domain (Gu et al., 2020). In the biomedical domain, much of the domain-specific vocabulary used can be tied to the PICO annotation model.

4.1.1 PICO annotations

PICO is a popular framework in the biomedical field for defining four key elements of biomedical knowledge: Population, Intervention, Comparison and Outcome. It is commonly used to tag and structure biomedical documents as well as for formulating search queries over documents in the domain (Brockmeier et al., 2019).

In the field of NLP, several works have been proposed to detect PICO elements. Demner-Fushman and Lin (2007) proposed rule-based methods and simple classifiers to identify PICO elements at the sentence level. Modern approaches use deep-learning models and formulate the problem as an entity recognition task, identifying PICO spans in randomised controlled trial documents (Kang et al., 2019) and biomedical abstracts (Brockmeier et al., 2019). Figure 3 gives an illustrative example of span-level tagging of PICO elements in biomedical literature.

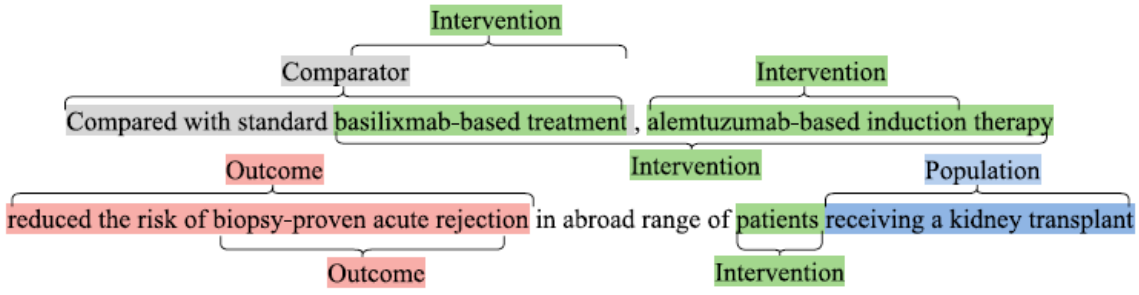


Figure 3. An example of PICO element detection in biomedical text (Stenetorp et al, 2012).

4.1.2 Existing methods for injecting domain knowledge into downstream NLP tasks

There have been several prior studies which use pretraining strategies with in-domain data to adapt PLMs for use in domain-specific applications. These PLMs have been shown to be highly effective for downstream domain-specifc NLP tasks (Gu et al., 2021; Beltagy et al., 2019). However, pretraining is a hugely computationally expensive task. For example, PubMedBERT (Gu et al., 2021) required 10 GPUs and 5 days to train. In the biomedical domain, strategies have been proposed to inject biomedical knowledge graphs (Meng et al., 2021) or PICO elements (Wallace et al., 2021) for downstream NLP tasks to improve performance on domain-specific applications. There is some prior literature combining PICO elements with the text summarisation task: Bui et al., (2016) use a rules-based approach to create a tabular structured summary of information relating to PICO elements. Afzal et al., (2020) propose an extractive multi-document and query-focused summarisation approach which extracts PIO elements from relevant documents, and then uses Word2Vec embeddings (Mikolov et al., 2013) of sentences containing these elements to calculate similarity with the original query and to select the most relevant sentences to form the extractive summary. Zhang et al., (2020b) use an active learning approach to train a deep-learning model to detect PICO sentences, and then use TextRank (Mihalcea and Tarau, 2004) and informativeness metrics to select which of these sentences to include in the final extractive summary. However, there is little research exploring how PICO elements can be incorporated into PLMs to improve the performance of text summarisation of biomedical documents.

4.2 The KeBioSum Model

In this section, methods to incorporate domain specific knowledge into summarisation models for long documents from the biomedical domain are explored. Specifically, a novel method, KeBioSum is proposed[17]. This method uses adapters (Houlsby et al., 2019) to infuse SOTA PLMs with span-level PICO element information for the text summarisation task. This differs from prior approaches by making use of PLMs rather than traditional models for injecting fine-grained biomedical knowledge. The proposed approach shows improved performance over strong baselines for extractive text summarisation on the biomedical data sets evaluated.

[17] https://github.com/xashely/KeBioSum

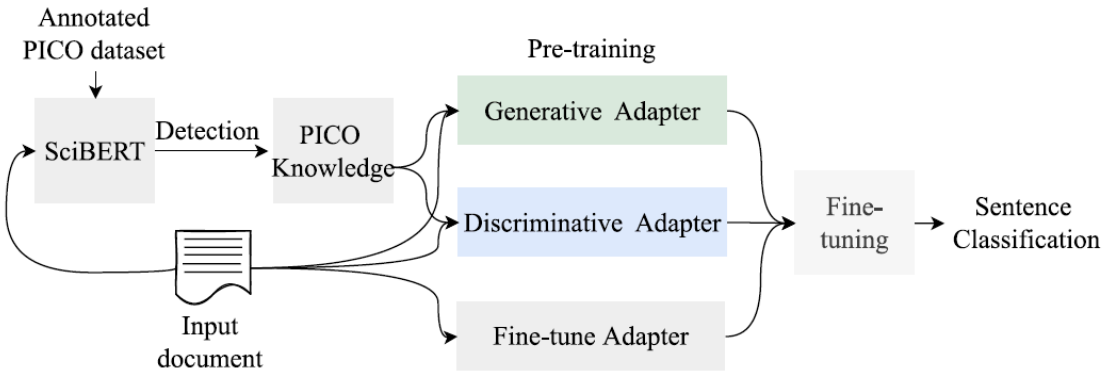


Figure 4. An overview of KeBioSum (Xie et al., 2022).

4.2.1 Method overview

Figure 4 gives a high-level view of the method. The fine-grained biomedical knowledge is injected via generative and discriminative adapters. In KeBioSum, the fine-grained biomedical knowledge injected into the PLM consists of PICO element detections, output from an independent SciBERT-based model (Beltagy et al., 2019), fine-tuned on the EBM-NLP data set (Nye et al., 2018). The overall extractive summarisation task is structured as a binary classification task, where each sentence in the source document is classified as to whether it should be included in the summary (class 1) or not included (class 0).

As with the unsupervised extractive approach outlined in Section 3, the method starts by dividing a document D it into its sentences s, such that $D = \{s_1, \ldots, s_n\}$, using the Stanford CoreNLP software package (Manning et al., 2014). The problem is formulated as a classification task, such that each sentence s_i in D is classified with a label $r_i \in \{0,1\}$, resulting in a predicted summary S^* that consists of the m sentences selected from D where $r = 1$, and $m \ll n$.

4.2.2 PICO detection model

Knowledge derived from PICO elements was injected into KeBioSum because PICO elements are known to have a high correlation with the information that clinicians look for when reading a new piece of research. For example, in documents they shared with us in 2021, the National Institute of Clinical Excellence[18] had fields for ‘Population’, ‘Study setting’, ‘Treatment’, and ‘Outcome’ on the forms they used to summarise new Random Controlled Trial research articles.

[18] https://www.nice.org.uk

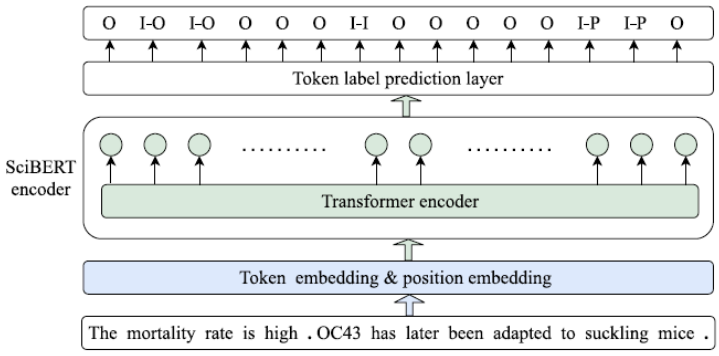


Figure 5. PICO detection model (Xie et al., 2022)

Figure 5 gives the model architecture used for making the PICO span predictions. To train the model, SciBERT (Beltagy et al., 2019) is used to encode the input document. These representations are then fed into a SoftMax classification layer to predict the label. The model was trained using the EBM-NLP data set (Nye et al., 2018), which consists of 5000 PubMed abstracts, annotated with P, I, and O tags. In this data set, a single class I represents both Intervention and Comparison and the P, I, O elements are represented by 'I-P', 'I-I' and 'I-O' tags. An additional tag 'O' was used to represent non-PICO tokens. The data set was split into train, validation, and tests splits with the number of documents in each being 4300, 500, and 200 respectively. The model was optimised with a cross-entropy loss and the best model was selected using the performance on the validation data set. As shown in Figure 4, in KeBioSum, the resulting fine-tuned model is used to predict PIO tags which are fed into the extractive summarisation model.

4.2.3 Adapter fusion for extractive summarisation

Adapters (Houlsby et al., 2019) are lightweight-frameworks for fine-tuning transformer-based models and provide an alternative to directly fine-tuning a PLM with many parameters. They work by introducing a small number of trainable parameters into the PLMs, whilst the rest of the weights within the PLM are frozen. They have been shown in prior literature to be able to successfully provide the ability to transfer knowledge into PLMs (Wang et al., 2021; Pfeiffer et al., 2020). Inspired by prior works (Clark et al., 2019; He et al., 2020) knowledge is injected via complimentary generative and discriminative adapters.

In KeBioSum, the generative adapter works to predict a partially masked input. Given an input sentence made up of t tokens, such that $t_i = \{x_1, \dots, x_t\}$, $w \ll t$ tokens are masked. All tokens, masked and unmasked, are then fed into the generative adapter, which is trained to predict the masked tokens.

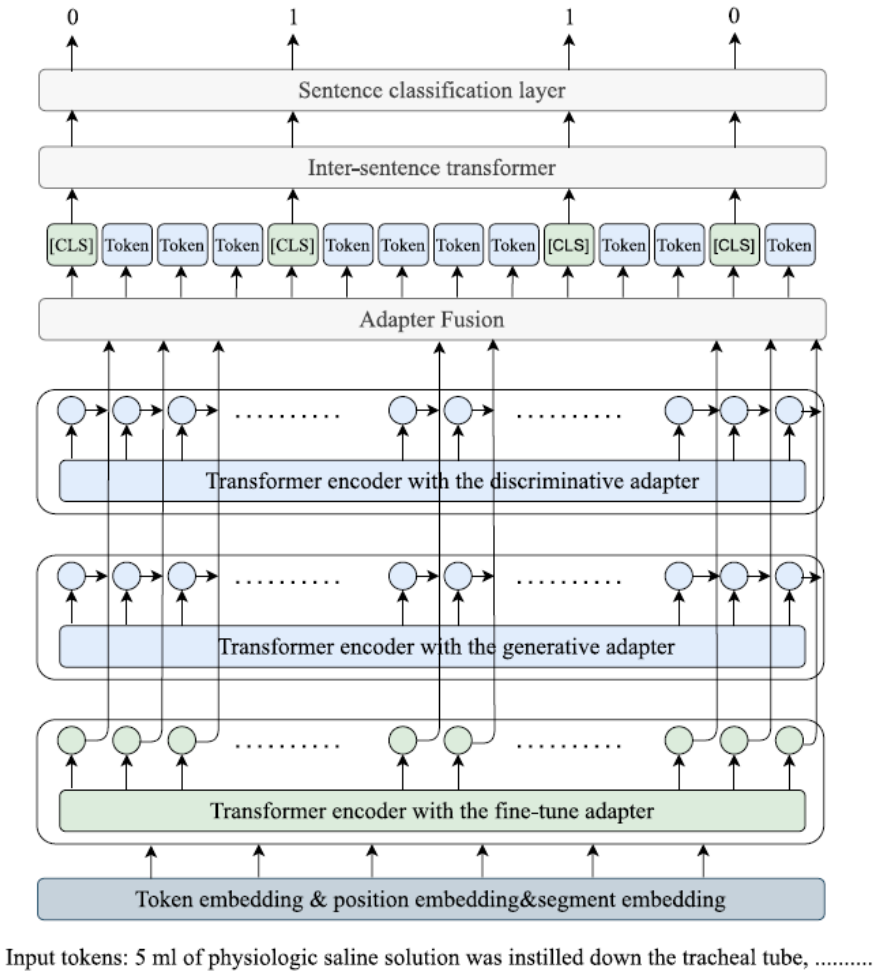


Figure 6. The architecture of the KeBioSum extractive summarisation model (Xie et al., 2022).

To train both the generative and discriminative adapters, all tokens detected to be PICO elements, and 15% of non-PICO tokens were masked. The approach taken is different to the random masking strategy of BERT (Devlin et al., 2019). The intention of this masking strategy was to guide the model to focus on learning the biomedical knowledge pivotal to the comprehension of the biomedical document. The loss of the generative adapter is the negative log-likelihood of predicting a token, given the masked input $\hat{x}$:

$$\mathcal{L}_G = -\sum_{j=1}^{t} \log\left(x_j \middle| \hat{x}\right) \qquad (3).$$

In contrast, the discriminative adapter works to predict the I-P, I-I, I-P and O tags for each token of the same partially masked input $\hat{x}$. In this model, the masked input is fed into the discriminative adapter to get a contextualised embedding for each token and a linear layer with a SoftMax function is then used to calculate the probability of each token being of PIO

class. The adapter maximises the probability of the expected category for each token in the input sentence:

$$\mathcal{L}_G = -\sum_{j=1}^{t} \hat{y_j} \log(y_j) \qquad (4),$$

where $\hat{y_j} \in \{I-P, I-I, I-O, O\}$ is the ground truth PICO label for token x_j. In addition to the discriminative and generative adapters which capture PICO information, an additional fine-tune adapter is used to retain the contextual information learnt in the PLM's pretraining. Following Pfeiffer et al., (2021) the adapter fusion strategy was implemented using an attention-based approach and implemented this using the adapter-transformers library[19].

The overall architecture of the summarisation model can be found in Figure 6. As shown in this figure, [CLS] tokens are inserted to separate sentences. Token embeddings (vectors representing the tokens), position embeddings (indices representing a token's position) and segment embeddings (indices used to represent a tokens segment position within the overall document) were calculated and fed into the PLMs. Sentence representations were calculated by taking the outputs of the [CLS] tokens from the adapter fusion layer. A final sentence classification, to predict the m sentences to be included in the predictive summary S^*, was calculated by feeding the sentence representations into an inter-sentence transformer, and then applying a sigmoid classification layer over these final representations. Thus, in KeBioSum, the weights being learnt are only of the adapter, adapter fusion and additional transformer layers, rather than the entire large PLM models.

4.3 Experimental set-up

4.3.1 Data sets

The same three biomedical data sets used to evaluate GenCompareSum in Section 3 were used to evaluate the efficacy of KeBioSum against baseline extractive methods: CORD-19 (Wang et al., 2020b), PubMed (Cohan et al. 2018), which is referred to in this section as PubMed-Long, and S2ORC (Lo et al., 2020). To effectively evaluate this method against baseline method MatchSum (Zhong et al., 2020), a PubMed-Short data set was additionally included in the evaluation. This data set consists of the PubMed data released with the MatchSum code, which uses a subset of articles of PubMed-Long and truncates the articles to use the introduction only. As in Section 3, the input to the summarisation model was a scientific publication, and the article abstract was used as the target summary. The numbers of samples included in each data set's

[19] https://github.com/adapter-hub/adapter-transformers

splits, in addition to the number of sentences selected to form the extractive summary, can be found in Table 7.

Data set	Train	Valid	Test	Ext
CORD-19	31,162	6,232	4,155	3
PubMed-Long	119,924	6,633	6,658	3
PubMed-Short	83,233	4,946	5,025	6
S2ORC	47,782	9,556	6,371	3

Table 7. Statistics of the biomedical summarisation data sets used to evaluate KeBioSum and other extractive methods.

4.3.2 Experimental set up

The KeBioSum model was implemented with Pytorch[20], HuggingFace[21] and the adapter-hub[22] software packages. To train the SciBERT-based PICO detection model, the 'scibert-scivocab-uncased model'[23] was used, which was fine-tuned on the EBM-NLP data set for a maximum of 75 epochs, with a learning rate of 0.002, a batch size of 32, and a dropout of 0.1.

For the KeBioSum extractive summarisation model, code was adapted from BERTSumExt (Liu and Lapata, 2019) to implement the adapter fusion method. A variety of different base PLMs were experimented with: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), which are pretrained on corpora of the general domain; BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021), which are pretrained on corpora from the biomedical domain. For the generative adapter, the hyperparameters were set as follows: learning rate: 1e-4, warm-up steps: 500, epochs: 12, weight decay: 0.001, batch size: 24. For the discriminative adapter, the hyperparameters were: learning rate: 5e-5, warm-up steps: 500, epochs: 12, weight decay: 0.001, batch size: 24. Perplexity was the metric used to select the best generative adapter and F1 was the metric used to select the best discriminative adapter. Two-layer transformers were used to create sentence representations in the final layers of the model. As in Section 3, the summarisation models were evaluated using ROUGE metrics (Lin, 2004), implemented using the pyrouge[24] package.

[20] https://pytorch.org
[21] https://github.com/huggingface
[22] https://github.com/adapter-hub/adapter-transformers
[23] https://huggingface.co/allenai/scibert_scivocab_uncased
[24] https://github.com/bheinzerling/pyrouge

4.3.3 Baseline methods

As in Section 3, ORACLE was implemented as an upper bound, LEAD as a crude baseline and BERTSumExt (Liu and Lapata., 2019) as a strong extractive baseline. BERTSumExt was additionally implemented with a PubMedBERT, rather than a BERT-based model, which is referred to as PubMedBERTSumExt in this work. The results were also compared to those of MatchSum (Zhong et al., 2020), a SOTA extractive summarisation technique which uses BERTSumExt to generate candidate sentences for the summaries, then generates candidate summaries by using different combinations of the selected sentences and selects a final summary from the candidate summaries by comparing the similarity of each candidate summary with the source document. In these experiments, all documents were truncated to 512 tokens, as the focus of the work in this section was on evaluating the efficacy of injecting knowledge into summarisation models for domain-specific documents, which long documents tend to be, rather than explicitly the summarisation of long documents.

4.4 Evaluation of the KeBioSum model

The experimental ROUGE F1 results are reported in Table 8, 9 and 10. Table 8 gives the performance on the CORD-19 and PubMed-Long data sets. Table 9 gives results for the S2ORC data set and Table 10 gives the results for the PubMed-Short data set. In Table 10, the results are additionally compared to the results reported by Zhong et al., (2020) for their MatchSum model. For completeness, scores are given for different adapters when used independently: '-fine-tune' indicates that only a fine-tune adapter was used in the model, '-gen' indicates that only a generative and fine-tune adapter were used, '-dis' indicates that a discriminative and fine-tune adapter were used. Where '-all' is written, this indicates that all adapters were used, and 'all-full' indicates where all adapters were used, and all PLM weights were updated during fine-tuning. '$*$' indicates where a model significantly outperformed BERTSumExt ($p < 0.05$) and '†' indicates where the "-all" model outperforms models with a subset of the adapters ($p < 0.05$).

Overall, in these tables of results, it can be observed that the '-all-full' KeBioSum models fine-tuned using a PubMedBERT base model outperformed all other models. Additionally, the KeBioSum '-all' models can be seen to outperform SOTA BERTSumExt, as well as its implementation with the PubMedBERT base, for all data sets. This indicates that the inclusion of the adapter fusion approach, which uses adapters to inject PICO knowledge, increases performance over SOTA approaches for summarisation of domain-specific documents, even for SOTA approaches where a base model which has been pretrained on in-domain data is used. This suggests that PICO elements capture important information in biomedical documents and that

PICO elements are seen in a document's target summary. This follows as the abstracts of biomedical literature, which are used as the target summaries in this experiment, are often structured into 'Background'. 'Methods', 'Results' and 'Conclusion' sections, which are aligned to PICO concepts. The advantage of the '-all' method is that a smaller number of parameters need to be fine-tuned and most of the weights in the PLM are frozen.

On inspection of the results of the '-gen', '-dis', and '-all' models, both the generative and discriminative adapters, which inject the PICO domain knowledge, increased performance even when isolated, over only using the fine-tune adapter. This further supports the theory that injecting domain knowledge improves summarisation on domain-specific documents. However, using all three adapters together gave increased performance over using any subset of adapters experimented with, thus indicating that the different adapters are complimentary to one another, and each inject different information to one another.

Although unsurprisingly the ROUGE scores when using PubMedBERT were seen to be higher than when using base PLMs from the general domain (e.g., BERT, RoBERTa), the greatest performance increases when using the adapters to inject domain knowledge were seen for the BERT and RoBERTa models. This suggests that injecting domain knowledge through adapters can significantly help summarisation of domain-specific articles, when a domain-specific PLM is unavailable.

Lastly, it is worth noting that the ROUGE scores on PubMed-Short are higher in Table 10 than they are for PubMed-Long in Table 8. This is simply because, following the work by Zhong et al., (2020), 6 sentences rather than 3 were selected for the predictive summary. However, as these models truncate a document to 512 tokens, by taking 6 sentences rather than 3, this often results in the entire truncated article being used as the summary. In this case, the results are no different to the LEAD method, hence the closeness of LEAD's scores to the other, more advanced, methods in Table 10 than in Table 8 and 9. This therefore highlights the need for study into summarisation of full long documents, which Section 3 and 5 of this work investigate.

	CORD-19			PubMed-Long		
Metrics	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**
LEAD	24.90	7.11	22.29	27.92	8.57	24.99
ORACLE	32.40	12.97	30.30	36.62	16.54	33.44
BERTSumExt	30.01	9.86	26.86	34.00	13.42	30.69
PubMedBERTSumExt	30.65	10.17	27.11	34.98	14.22	31.37
BERT-fine-tune	28.15	8.74	22.99	32.42	12.60	29.38
BERT-gen	29.75	9.25	24.02	34.12	13.54	30.78
BERT-dis	29.67	9.23	24.01	33.76	13.35	30.78
BERT-all	**30.79**†	**10.37**†	**25.13**†	**35.12**†	**14.54**†	**31.80**†
RoBERTa-fine-tune	29.58	9.53	26.40	34.00	13.44	30.69
RoBERTa-gen	30.01	9.61	26.77	34.06	13.54	30.76
RoBERTa-dis	30.02	9.63	26.76	34.07	13.69	30.78
RoBERTa-all	**30.10**†	**10.72**†	**27.81**†	**35.08**†	**14.69**†	**31.78**†
BioBERT-fine-tune	29.53	9.36	26.46	34.01	13.42	30.71
BioBERT-gen	30.04	9.67	26.72	34.05	13.59	30.80
BioBERT-dis	30.08	9.71	26.78	34.06	13.59	30.79
BioBERT-all	**31.11**†	**10.74**†	**27.82**†	**35.09**†	**14.62**†	**31.82**†
PubMedBERT-fine-tune	29.77	9.53	26.56	34.31	13.86	31.03
PubMedBERT-gen	30.75	9.98	27.42	34.47	13.99	31.16
PubMedBERT-dis	30.77	9.97	27.43	34.47	13.98	31.16
PubMedBERT-all	30.85	11.01	28.53	35.94	15.39	32.59
PubMedBERT-all-full	**32.04**†$*$	**12.61**†$*$	**29.10**†$*$	**36.39**†$*$	**16.27**†$*$	**33.28**†$*$

Table 8. Rouge F1 results of different models on CORD-19 and PubMed-Long data sets. Bold font indicates the best results, underlined font indicates the second-best results; '$*$' indicates where a model outperformed BERTSumExt and '†' indicates where "-all" model outperforms models with only a subset of the adapters.

Model	R1	R2	RL
LEAD	26.45	10.35	24.19
ORACLE	37.62	17.72	34.21
BERTSumExt	32.94	14.02	29.08
PubMedBERTSumExt	34.69	15.06	32.30
BERT-fine-tune	29.61	12.07	26.99
BERT-gen	<u>33.25</u>	14.25	<u>30.25</u>
BERT-dis	33.19	<u>14.27</u>	30.20
BERT-all	**33.27**†	**14.33**†	**30.29**†
RoBERTa-fine-tune	**33.61**	14.48	**30.56**
RoBERTa-gen	33.52	<u>14.59</u>	30.51
RoBERTa-dis	<u>33.57</u>	14.55	<u>30.54</u>
RoBERTa-all	33.45	**15.59**†	30.46
BioBERT-fine-tune	32.53	14.41	30.27
BioBERT-gen	33.12	14.45	30.41
BioBERT-dis	<u>33.20</u>	<u>14.57</u>	<u>30.49</u>
BioBERT-all	**34.47**†	**15.62**†	**31.51**†
PubMedBERT-fine-tune	34.01	14.81	30.94
PubMedBERT-gen	34.06	14.84	30.97
PubMedBERT-dis	34.07	14.85	30.98
PubMedBERT-all	<u>36.58</u>	<u>15.75</u>	<u>33.19</u>
PubMedBERT-all-full	**37.44**†$*$	**16.72**†$*$	**34.08**†$*$

Table 9. Rouge F1 results of different models on the S2ORC data set. Bold font indicates the best results, underlined font indicates the second-best results; '$*$' indicates where a model outperformed BERTSumExt and '†' indicates where "-all" model outperforms models with only a subset of the adapters.

Model	R1	R2	RL
LEAD	37.58	12.22	33.44
ORACLE	45.12	20.33	40.19
MATCH-ORACLE	42.21	15.42	37.67
BERTSum	41.05	14.88	36.57
-3gram-Blocking	38.81	13.62	34.52
-4gram-Blocking	40.29	14.37	35.88
PubMedBERTSum	42.14	16.17	37.86
MatchSum	41.21	14.91	36.75
BERT-fine-tune	40.20	14.60	36.42
BERT-gen	40.41	14.74	36.61
BERT-dis	40.43	14.76	36.63
BERT-all	40.48	14.76	36.68
RoBERTa-fine-tune	40.52	14.81	36.70
RoBERTa-gen	40.55	14.85	36.74
RoBERTa-dis	40.56	14.87	36.76
RoBERTa-all	40.61	14.89	36.81
BioBERT-fine-tune	40.52	14.78	36.70
BioBERT-gen	40.59	14.87	36.78
BioBERT-dis	40.58	14.84	36.76
BioBERT-all	40.60	14.89	36.80
PubMedBERT-fine-tune	41.25	15.42	37.40
PubMedBERT-gen	41.31	15.50	37.47
PubMedBERT-dis	41.32	15.51	37.57
PubMedBERT-all	42.36	16.54	38.66
PubMedBERT-all-full	**43.98**$^{\dagger *}$	**18.27**$^{\dagger *}$	**39.93**$^{\dagger *}$

Table 10. Rouge F1 results of different models on the PubMed-Short data set. Bold font indicates the best results, underlined font indicates the second-best results; '$*$' indicates where a model outperformed BERTSumExt and '†' indicates where "-all" model outperforms models with only a subset of the adapters.

4.5 Future work

Long document data sets are often highly domain specific, however, to-date, most research focuses on short document data sets from the general domain (Koh et al., 2022). In this section methods for injecting domain-specific knowledge into SOTA PLM models for extractive text

summarisation were explored. It was shown that using an adapter-based method can be effective, particularly in cases where a PLM pretrained on domain-specific data is unavailable. Future steps would be to combine the adapter-methods proposed in this section with methods for extending models to long documents, as explored in Section 3 and 5 of this work.

5 Abstractive text summarisation methods

This section details an evaluation of current state-of-the-art methods for long document abstractive text summarisation implemented in a realistic setting. Specifically, a single NVIDIA V100 GPU was used to implement a range of abstractive summarisation methods on long documents of the scientific domain. Intrinsic (i.e., automated) and extrinsic (i.e., human) evaluations were conducted of the generated abstractive summaries. In this section, a novel method for summarisation of long documents is proposed, which makes use of zoning and conditioning to apply structure to a generated summary. It is shown that this method is regarded highly by human annotators in terms of fluency and coherence but does not perform particularly well on measures of factuality. It is also shown that an existing method, DANCER (Gidiotis and Tsoumakas, 2020) performs well across several automated and human annotated measures.

5.1 Background and existing methods

As stressed throughout this work, the value derived from text summarisation comes from the timesaving in making a long document more succinct and therefore more easily consumable by a reader. Despite this, to-date, most studies researching abstractive text summarisation, and the automatic metrics used to evaluate its efficacy, have focussed on the CNN/DM (Hermann, 2015), Newsroom (Grusky et al., 2018) and XSUM (Narayan et al., 2018) data sets, all of which consist of relatively short documents. However, documents in long document data sets (Cohan et al., 2018; Sharma et al., 2019; Kornilova and Eidelman, 2019; Koupaee and Yang Wang., 2018) are on average 8.3x longer than their short document counterparts (Koh et al., 2022).

As with the extractive methods mentioned previously, transformer-based PLMs are used in the majority of SOTA abstractive summarisation methods (Liu and Lapata., 2019; Lewis et al., 2020; Zhang et al., 2020a; Beltagy et al., 2020). One can speculate that the lack of research into abstractive summarisation for long documents is largely due to the expense associated with fine-tuning the attention mechanisms embedded in the PLMs used in SOTA abstractive text summarisation methods. This is even more problematic for abstractive than extractive methods as, while extractive methods are often framed as classification tasks where each sentence can be

treated semi-independently, abstractive summarisation is a generative task which requires sentences within the text to be considered together.

To overcome the often prohibitively expensive computation of fine-tuning an attention mechanism of a PLM on a long-document summarisation data set, LongT5 (Guo et al., 2021) BigBird (Zaheer et al., 2020) and LED (Beltagy et al., 2020) models all have their attention mechanisms efficiently adapted to support both the local and global contexts of a document. However, although they can in theory support much longer sequence lengths than the default 512 tokens which traditional PLMs are restricted to (Devlin et al., 2019), in practise, it is often still necessary to truncate documents significantly due to GPU memory limitations. Other approaches for abstractive summarisation of long documents include hybrid extractive-abstractive methods (Gehrmann et al., 2018; Liu and Lapata, 2019), which have been shown to provide good coverage of the source document but suffer badly from factual inconsistency (Huang et al., 2020).

Text zoning is an NLP task which aims to classify a larger body of text into different zones or sections. This concept has been shown to be effectively applied across a range of domains, such as emails (Repke and Krestel, 2018), job advertisements (Gnehm, 2018) and scientific literature (Teufel, 2002) and has been utilised in various downstream NLP tasks. For the summarisation task, Stede et al., (2006) used zones to ensure the most important sections of movie reviews were included in the summary, whilst Contractor et al., (2012) used zone classification of a sentence as an input feature when predicting extractive summaries of scientific documents. More recently, for abstractive summarisation, text zoning has been used to select sections of documents in order to generate summaries from sections at a time, thus providing a strategy for extending abstractive summarisation to long documents without requiring huge computational resource. Gidiotis and Tsoumakas (2020) use only certain sections of scientific articles for their summary, whilst Liu et al., (2022) explore different methods for sectioning documents and find that for long documents of the scientific domain, summarisation using discourse locality (i.e., document zones) performs well.

5.2 Abstractive summarisation method using conditional zoning

Here, a method for abstractive summarisation of long documents is proposed. This method generates a predictive summary with a consistent structure and, by using zoning techniques, reduces the memory requirements on the training hardware by shortening both the length of the input text and the target text per training step. The proposed training and inference methods can be seen in Figure 7 and 8 respectively. Figure 7 shows that at training time, both the source document and target summary are split into sections. Sections which belong to the document

zones deemed most important to the summary are matched between the source and target and used as training pairs to a single LED model. Each input section has text prepended to it which conditions the summary prediction on the zone of the document. Figure 8 shows that at inference time, a source document is split into sections. Important sections are selected and are used as inputs to the fine-tuned LED model. Again, each input section has text prepended to it which conditions the summary prediction on the zone of the document. Predicted summary sections are combined, and the name of the zone is prepended to the prediction for each section. In both settings, non-important sections are discarded.

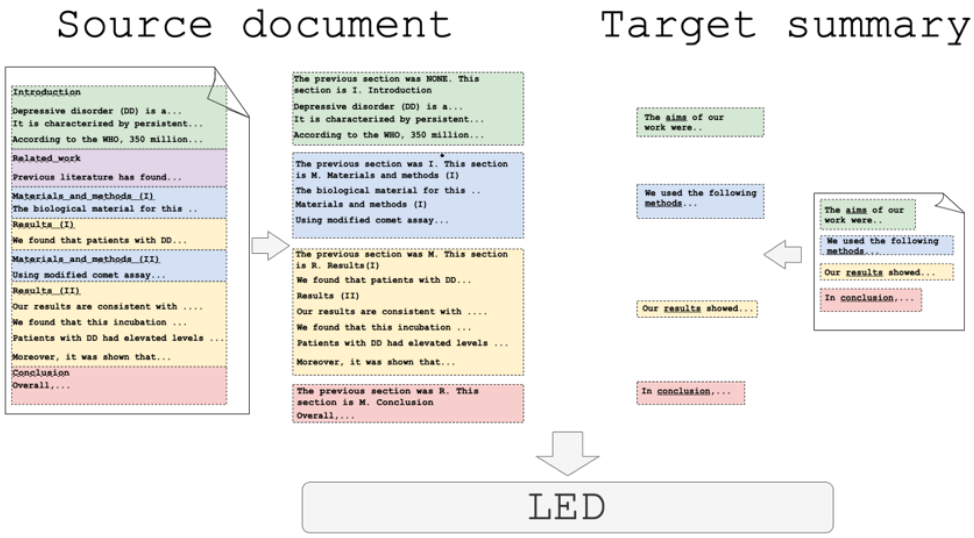


Figure 7. The abstractive method using zoning during training.

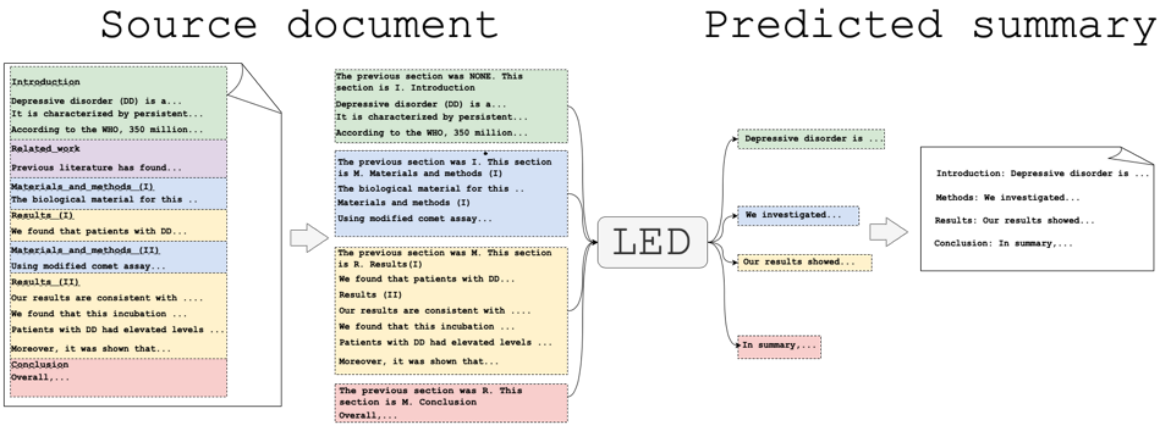


Figure 8. The abstractive method using zoning during inference.

5.2.1 Source document zoning

Zone analysis has been shown to be an effective mechanism for improving information extraction in long documents (Mizuta et al., 2006). The method proposed in this section extends this idea for use in abstractive summarisation. The method creates training pairs from zones of a document and what is determined to be their corresponding zone in the summary. At inference time, the predicted summaries for each zone are combined to create one overall document summary.

Given a document D, it is first split it into its sections S, such that $D = \{S_1, \dots, S_n\}$, and each section S_i has associated section heading h_i. The words in each of the section headings h_i are matched against a pre-defined list to classify the section S_i into one of five zone classes: introduction and aims I, methods M, results and discussion R, conclusion C, none O. Sections with the same zone classification are then concatenated, in the order they appear in the original document, resulting in a maximum of four larger sections: $X_t ; t \in \{I, M, R, C\}$ which represent the most useful parts of the document. Any sections with a classification of O are discarded.

5.2.2 Reference summary zoning

To find corresponding sections in the document's reference summary R, to use as the target section summaries, the reference summary is divided into these sections A_I, A_M, A_R, A_c again using keyword matching and by assuming that sections will appear in the order I, M, R, C. In Figure 7, the type of keywords used to predict the zones within the summary are underlined. Document sections X_I, X_M, X_R, X_c are mapped to their corresponding abstract summary sections. Any section X_n without a corresponding abstract section A_n is discarded, leaving document section-abstract section training pairs. At training time, each section-abstract training pair is treated individually.

5.2.3 Conditional generation

A design choice was made to include all section-summary pairs (regardless of their zone classification) as training data for fine-tuning a single PLM model, rather than fine-tuning separate models for each zone class. This is due to the improved generalisability that comes with using large amounts of training data for PLMs. Thus, if A^* is the predicted summary section, the generative task can be described as:

$$A^* = PLM(X) \qquad (5).$$

However, this design choice also presents challenges at inference time in ensuring the correct zone is selected for the type of input section. For example, without knowing the input section zone, a model may start the summary with "in conclusion…" despite the input zone being the methods section. When the predicted summary sections are combined at inference time, this would result in a lack of fluency and coherence in the target document, despite potentially achieving high ROUGE scores.

To overcome this challenge, a prior is introduced on the zone summary generation, meaning that the generation is conditioned on the current and previous document sections, i.e.,

$$A_t^* = PLM(X_t | X_{t-1}, t) \quad (6).$$

In practise, this is achieved by prepending each section X with its heading h before concatenating any sections from the same zone together to form input X. In addition, the method makes use of the overall document structure, assuming it to always follows the pattern I, M, R, C, by prepending X with the phrase "The previous section was $[t-1]$. This section is $[t]$.", as illustrated in Figure 7 and 8. This is implemented regardless of whether a section X_{t-1} is found in the input document, as each pair is treated independently.

5.2.4 Inference

At inference time, the same steps defined in 5.2.1 and 5.2.3 are followed, however, the steps in 5.2.2 to find matching reference summary sections are not required. As shown in Figure 8, each zone X_t which is deemed to be of importance is fed into the fine-tuned sequence-to-sequence model. Predicted summary sections A_t^* are concatenated, to create an overall predicted document summary S^*. The name of the zone category is also prepended to the beginning of each predicted section of the summary, as illustrated in Figure 8.

5.3 Generation of abstractive summaries

5.3.1 Data sets

The abstractive summarisation methods were evaluated against the same PubMed and ArXiv data sets (Cohan et al., 2018) as were used in Sections 3 and 4. In line with the extractive methods, their abstracts were used as the reference summaries. For the test data sets, the subset of documents for which it is possible to find at least three out of the four zones in the source document were taken, using the method described in 2.1. Training, validation and test data set sizes can be found in Table 11. To allow for a fair comparison, the test data sets for evaluation

of all summarisation methods were restricted to this subset, hence the differences between Table 11 and the counts of documents reported in the extractive methods in previous sections.

	Train	Val	Test
PubMed	117108	6631	4011
ArXiv	202917	6436	2875

Table 11. Counts of numbers of articles in each data set.

As described in Section 5.2, each document section-summary pair was treated independently during training. Table 12 gives the number of training and validation samples used in this method. For zoning methods, at test time, the summaries for each document zone were concatenated and evaluated against the complete reference summary.

	Train	Val
PubMed	157817	8955
ArXiv	62393	2098

Table 12. Counts of numbers of training and validation samples in each data set for the zoning methods.

In Table 13, the average number of tokens in each data set is given. For the original documents, the input was the full article, and the target was the document abstract. For the document divided into zones, the input was the source document section, and the target was the matched zone of the reference summary. As can be seen in Table 13, zoning methods significantly reduced the length of input and target tokens required in each training step, thus reducing memory requirements of the training hardware.

		Input tokens	Target tokens
Original	PubMed	3209	208
	ArXiv	6515	279
Zoned	PubMed	1070	103
	ArXiv	2493	177

Table 13. The average number of tokens in each data set.

5.3.2 Computational set up

All experiments were run on NVIDIA V100 instances, allocating one GPU instance per experiment. A decision was made to restrict GPU usage to this scale as the aim of this work was to understand how to derive value from abstractive text summarisation in a realistic setting, considering both environmental and economic resource consumption. Although more recent NVIDIA A100 instances have higher memory specifications and similar power consumption to V100s (Schoonhoven et al., 2022), they are still significantly more expensive than V100s, costing a minimum of $32.77/hr vs $3.06/hr when using AWS EC2 instances, as shown in the screenshots taken from their website in Figure 9 and 10[25]. Therefore, the use of A100s is not economically feasible for many people at present.

Instance name	On-Demand hourly rate	vCPU	Memory	Storage	Network performance
p4d.24xlarge	$32.7726	96	1152 GiB	8 x 1000 SSD	400 Gigabit
p4de.24xlarge	$40.96575	96	1152 GiB	8 x 1000 SSD	400 Gigabit

Figure 9. Table giving the cost of on-demand pricing of AWS EC2 instances with NVIDIA A100 GPUs for the US East (N. Virginia) region.

Instance name	On-Demand hourly rate	vCPU	Memory	Storage	Network performance
p3.2xlarge	$3.06	8	61 GiB	EBS Only	Up to 10 Gigabit
p3.8xlarge	$12.24	32	244 GiB	EBS Only	10 Gigabit
p3.16xlarge	$24.48	64	488 GiB	EBS Only	25 Gigabit
p3dn.24xlarge	$31.212	96	768 GiB	2 x 900 NVMe SSD	100 Gigabit

Figure 10. Table giving the cost of on-demand pricing of AWS EC2 instances with NVIDIA V100 GPUs for the US East (N. Virginia) region.

[25] https://aws.amazon.com/ec2/pricing/on-demand/ (accessed on 28/01/2023)

5.3.3 Implementation of abstractive methods

As baselines, several SOTA methods from prior literature were implemented. In addition, the abstractive method defined in Section 5.2 was implemented. Firstly, shown to have strong performance on the abstractive summarisation task, BERTSumExtAbs and BERTSumAbs methods (Liu and Lapata, 2019) were implemented. These methods both work by fine-tuning a BERT-based (Devlin et al., 2019) PLM for the summarisation objective. BERTSumAbs fine-tunes the BERT based model with an abstractive training objective, whilst BERTSumExtAbs first fine-tunes a BERT-based PLM with an extractive objective, and then fine-tunes the model further, from a saved checkpoint, with an abstractive objective. Since these models are BERT-based, the input document is truncated to 512 tokens.

DANCER (Gidiotis and Tsoumakas, 2020) was also implemented. Like the method proposed in Section 5.2, it splits the source document into sections for summarisation, but unlike the method proposed in this work, does not make use of the structure of the document or reference summary. DANCER splits a document into zones using keyword matching, then finds corresponding sections of the target abstract using ROUGE matching (Lin, 2004). It uses beam search decoding with 4 beams to generate the section summaries of a maxiumum of 120 tokens and combines the generated summaries of each section to form the complete article summary. . The original implementation, which fine-tunes a PEGASUS-based PLM, was followed. This method also truncates the text at 512 tokens, however, since the DANCER method splits the long document into sections, the truncation does not have a significant impact, and text from across the length of the input document is still considered in the generation. In addition, LED, a PLM designed to efficiently extend to long documents, is directly fine-tuned. However, on the NVIDIA v100 GPU, it was found that the document must be truncated to 1024 tokens for training. A hybrid extractive-abstractive method was also implemented. An extractive-abstractive model, which will be referred to as LEDExtAbs in this work, was trained using ORACLE extractive summaries as an input, optimised for the ROUGE recall metric, but limiting the number of sentences selected so that the total number of input tokens was less than 1024. At test time, the implemented unsupervised, extractive method GenCompareSum (Bishop et al., 2022) was used instead of the ORACLE method. GenCompareSum does not truncate the source document and has previously shown strong performance on the PubMed and ArXiv data sets. Lastly, the method outlined in Section 5.2 was implemented. For this, an LED PLM was fine-tuned with the zoned document section-summary pairs. For all models where an LED PLM was fine-tuned, a batch size of 3 was used and the model was trained for 3 epochs with a maximum input length of 1024 tokens and maximum generation length of 250 tokens.

5.4 Human evaluation study

The experiments were evaluated using summaries generated from the long document, English-language PubMed and ArXiv data sets described in Section 5.4.1. For the human evaluation, the summaries were generated using three different abstractive summarisation methods: DANCER, LEDExtAbs, and the method proposed in this work, described in Section 5.2. These methods are all able to consider text from across the entire length of a long document when generating a summary.

As the PubMed and ArXiv data sets included in the study are highly domain specific, six expert annotators were recruited (three per data set) to review the automatically generated summaries. The expert annotators reviewing the PubMed data set all had English as their first language and, at the time of evaluation, were, or were in the final years of study to be, qualified clinicians. The expert annotators for the ArXiv data set had all achieved a minimum of an undergraduate degree in a physical science. Although all annotators were fluent in English, two out of three annotators of the ArXiv data set did not have English as their first language. The annotators who participated in this study were all friends or colleagues of the author and therefore volunteered to participate in the study without payment. It was made clear to the annotators that the purpose of this evaluation was scientific research on abstractive summarisation and there was an intention to use the results for scientific publication.

In-line with prior literature (Fabbri et al., 2021), the abstractive summarisation methods were evaluated with respect to the coherence, fluency, consistency (factuality). The definitions provided to annotators were as follows:

1. **Coherence**: Whether the text is well structured and is non contradictory to itself. "The summary should be well-structured and well-organised. The summary should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic." (Fabbri et al., 2021)

2. **Fluency**: How well the text flows and the quality of the individual sentences. The text "should have no formatting problems, or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read" (Fabbri et al., 2021). Annotators are also instructed to penalise summaries that contain repetition. Note, all summaries are in lower-case so please ignore capitalization.

3. **Factuality**: “The factual alignment between the summary and the summarised source. A factually consistent summary contains only statements that are entailed by the source

document. Annotators are also asked to penalise summaries that contained hallucinated facts (facts that cannot be found in the source document)" (Fabbri et al., 2021). The main types of consistency errors which annotators should be made aware of (Huang et al., 2021) are:

- **Intrinsic error**: A fact that is contradicted to the source document, which is also referred to as "intrinsic hallucination", e.g., a numerical value in the source document being repeated in the wrong fact in the summary.
- **Extrinsic error**: A fact that is neutral to the source document (i.e., the content that is neither supported nor contradicted by the source document), i.e., a statement which seems to have been completely made up.

For each data set, each human annotator evaluated the same three summaries generated from the same fifteen randomly sampled documents, thus resulting in two-hundred and seventy scored summaries over the two data sets. A ranking-based metric for coherence and fluency was chosen due to ranking scales having been shown to be more effective than rating scales in prior literature (Kiritchenko and Mohammad, 2017), and particularly in cases where the metric is not easily quantifiable, e.g., coherence (Steen and Markert, 2021). For the factuality metric, a binary classification metric (entailed vs not entailed) was used and annotators to mark a sentence as 'not entailed' if there were any factual inconsistencies. The annotators were unaware of which method created each summary.

Given the nature of the domain-specific, long-document data sets, it is extremely laborious to thoroughly evaluate the summaries manually. This is especially true for the factuality metric; if a reviewer had simply been provided with the source document and predicted summary, they would have had to manually go through the several thousands of words in each source document to attempt to locate the source and assess entailment for each statement in the generated summary. To address this challenge, the following strategy was used for each measure:

1. **Coherence:** Ranked 1-3, where 1 indicates the most coherent summary and 3 indicates the least coherent. The generated summaries were assessed against each other with no reference. An example of the coherence ranking by one annotator on the PubMed data set is given in Figure 11.
2. **Fluency:** Ranked 1-3, where 1 indicates the most fluent summary and 3 indicates the least fluent. The generated summaries were assessed against each other with no

reference. An example of the fluency ranking by the same annotator for the same article is given in Figure 12.

3. **Factuality:** Three sentences were sampled from each generated summary and, for each, a sentence embedding (Reimers and Gurevych, 2019) was generated. Sentence embeddings were also generated for each sentence in the source document and the most similar two sentences from the source document to each sampled sentence were selected by comparing cosine similarity of their sentence embeddings. For the two sentences selected from the source document, the prior and following sentences from the source document were concatenated to them to create two longer text snippets. The reader was then given the three sentences sampled from the generated summary and two text snippets from the source document for each of these three sentences. The reader was then asked to decide whether, given the text snippets, if each sentence was entailed (and thus given a score of 1) or not entailed (and given a score of 0). As there were three sentences evaluated for each generated summary, the scores were averaged to give one score per summary. Screenshots of the factuality scoring for the three summaries in this PubMed sample are given in Figure 13, 14 and 15.

Rank the summaries 1,2,3 in terms of their coherence

Summary	Rank (1 best, 3 worst)
background : in recent years, there has been a much increase in the incidence of urinary tract infection (upec) in iran. we evaluated the virulence of the fimh gene in upec isolates from hospitalized patients and out - patients with upec. methods : we investigated the fimh gene in the upec isolates from hospitalized patients and out - patients with uti referred to the educational hospitals of shahrekord, iran. the bacterial isolates were subjected to screening for the presence of the fimh gene by pcr. in addition to analyzing all the upec isolates, the fimh gene was detected in other strains of e. coli. results : of the 130 upec isolates studied, the fimh gene was found in 92.8% of upec isolates. the fimh gene was detected in 98% of the e. coli isolates from uti. the fimh gene was detected in 98% of upec isolates from a patient with uti. for more subsequent investigations, the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti	3
background : urinary tract infections (utis) are one of the inflammatory diseases produced by high multiplication of many pathogens in the urinary apparatus , resulting in alterations in the perfect function of the urinary tract and kidneys . uropathogenic escherichia coli (upec) strains have special virulence factors , including pili or fimbriae , which mediate attachment to uroepithelial and vaginal cells , resistance to human serum bactericidal activity , haemolysin production , and increased amounts of k capsular antigen . furthermore , virulence factors of upec strains the fimh gene was found in 130 isolates (92.8%) of upec . of the 130 isolates positive for the fimh gene , 62 (47.7%) and 68 (52.3%) belonged to hospitalized patients and outpatients , respectively . the aim of this study was to determine the prevalence of the fimh gene in urothelial epithelial cells (upecs) isolated from hospitalized patients and outpatients with urinary tract infections (utis) referred to educational hospitals of shahrekord , iran.materials and methods : in this cross - sectional study ,130 upec isolates were isolated from hospitalized patients and outpatients with utis referred to educational hospitals of shahrekord , iran .	1
introduction: background urinary tract infections (utis) are one of the major causes of morbidity and mortality in many parts of the world. e. coli is one of the most frequent pathogen responsible for up to 80% of utis. objectives the aim of this study was to determine the virulence factors of type 1 fimbriae (fimh) isolated from uropathogenic e. coli isolates. methods: : one hundred and six e. coli strains were collected from hospitalized and non - hospitalized patients with uti during 2012. the isolates were characterized and identified using gram staining and biochemical tests . genomic dna templates for pcr amplification were gained from overnight growth of bacterial isolates on luria - bertani agar and subjected to screening for the presence of the fimh gene by pcr . results: : a total of 130 upec isolates were detected in the study . the fimh gene was detected in 92.8 % of upec isolates . the fimh gene was detected in 47.7 % of hospitalized patients and 52.3 % of outpatients . the high binding capacity of fimh was found to be in more than 95 % of the isolates .	2

Figure 11. Screenshot of coherence rankings by one annotator on a sample from the PubMed data set.

Rank the summaries 1,2,3 in terms of their fluency

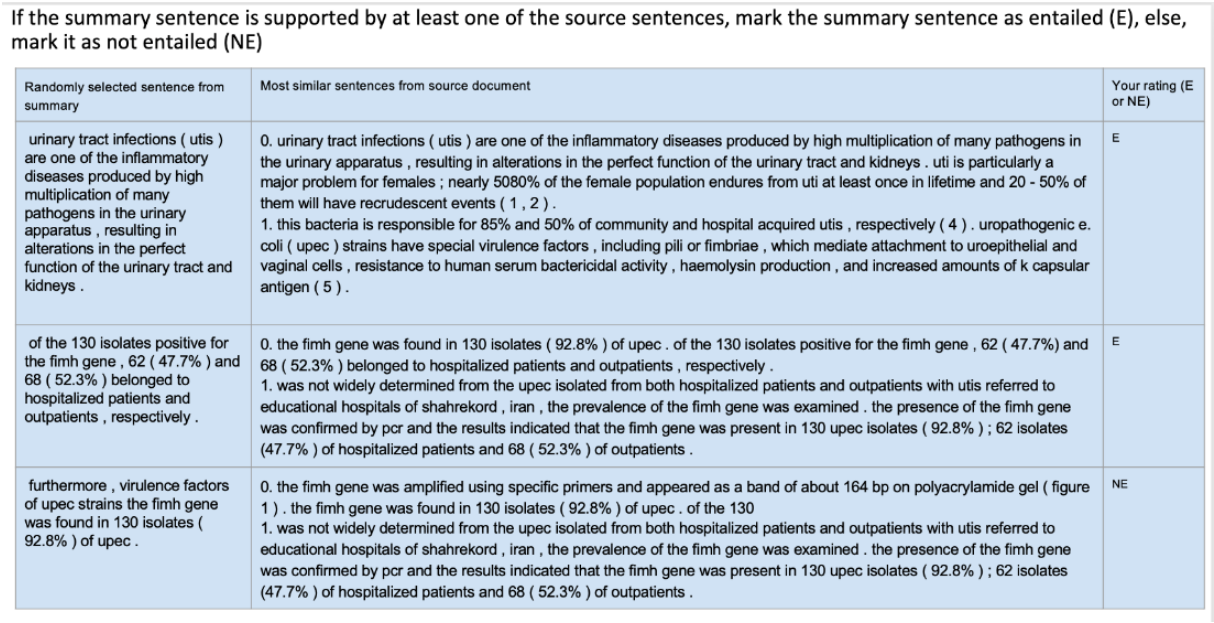
Summary	Rank (1 best, 3 worst)
background : in recent years, there has been a much increase in the incidence of urinary tract infection (upec) in iran. we evaluated the virulence of the fimh gene in upec isolates from hospitalized patients and out - patients with upec. methods : we investigated the fimh gene in the upec isolates from hospitalized patients and out - patients with uti referred to the educational hospitals of shahrekord, iran. the bacterial isolates were subjected to screening for the presence of the fimh gene by pcr. in addition to analyzing all the upec isolates, the fimh gene was detected in other strains of e. coli. results : of the 130 upec isolates studied, the fimh gene was found in 92.8% of upec isolates. the fimh gene was detected in 98% of the e. coli isolates from uti. the fimh gene was detected in 98% of upec isolates from a patient with uti. for more subsequent investigations, the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti	3
background : urinary tract infections (utis) are one of the inflammatory diseases produced by high multiplication of many pathogens in the urinary apparatus , resulting in alterations in the perfect function of the urinary tract and kidneys . uropathogenic escherichia coli (upec) strains have special virulence factors , including pili or fimbriae , which mediate attachment to uroepithelial and vaginal cells , resistance to human serum bactericidal activity , haemolysin production , and increased amounts of k capsular antigen . furthermore , virulence factors of upec strains the fimh gene was found in 130 isolates (92.8%) of upec . of the 130 isolates positive for the fimh gene , 62 (47.7%) and 68 (52.3%) belonged to hospitalized patients and outpatients , respectively . the aim of this study was to determine the prevalence of the fimh gene in urothelial epithelial cells (upecs) isolated from hospitalized patients and outpatients with urinary tract infections (utis) referred to educational hospitals of shahrekord , iran.materials and methods : in this cross - sectional study ,130 upec isolates were isolated from hospitalized patients and outpatients with utis referred to educational hospitals of shahrekord , iran .	2
introduction: background urinary tract infections (utis) are one of the major causes of morbidity and mortality in many parts of the world. e. coli is one of the most frequent pathogen responsible for up to 80% of utis. objectives the aim of this study was to determine the virulence factors of type 1 fimbriae (fimh) isolated from uropathogenic e. coli isolates. methods: : one hundred and six e. coli strains were collected from hospitalized and non - hospitalized patients with uti during 2012. the isolates were characterized and identified using gram staining and biochemical tests . genomic dna templates for pcr amplification were gained from overnight growth of bacterial isolates on luria - bertani agar and subjected to screening for the presence of the fimh gene by pcr . results: : a total of 130 upec isolates were detected in the study . the fimh gene was detected in 92.8 % of upec isolates . the fimh gene was detected in 47.7 % of hospitalized patients and 52.3 % of outpatients . the high binding capacity of fimh was found to be in more than 95 % of the isolates .	1

Figure 12. Screenshot of fluency rankings by one annotator on a sample from the PubMed data set.

If the summary sentence is supported by at least one of the source sentences, mark the summary sentence as entailed (E), else, mark it as not entailed (NE)

Randomly selected sentence from summary	Most similar sentences from source document	Your rating (E or NE)
for more subsequent investigations, the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti	0. in addition , single - nucleotide polymorphism (snp) analysis of fimh is a screening tool for epidemiological typing of upec (11 , 12) . therefore , the research on bacterial virulence factors can result in expansion and development of new methods for diagnosis and prevention of utis . for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . 1. for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . the present study was conducted for detection of the fimh virulence gene from upec isolated from both hospitalized patients and outpatients with uti , referred to educational hospitals of shahrekord , iran .	E
the fimh gene was detected in 98% of the e. coli isolates from uti.	0. for example , kaczmarek et al . (13) evaluated and detected the genes encoding virulence factors among e. coli strains with k1 antigen as well as the non - k1 e. coli strains . they found that the fimh gene existed in the whole tested e. coli k1 strains as well as in 97.0% of non - k1 strains . 1. (16) studied 18 upec isolates collected from females and found that the fimh gene was the most prevalent virulence factor and 100% of the isolates had that gene . in another study , 17) demonstrated that the fimh gene was the most frequent virulence gene and was detected in 98% of e. coli strains isolated from patient with utis .	E
we evaluated the virulence of the fimh gene in upec isolates from hospitalized patients and out - patients with upec.	0. in addition , single - nucleotide polymorphism (snp) analysis of fimh is a screening tool for epidemiological typing of upec (11 , 12) . therefore , the research on bacterial virulence factors can result in expansion and development of new methods for diagnosis and prevention of utis . for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . 1. for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . the present study was conducted for detection of the fimh virulence gene from upec isolated from both hospitalized patients and outpatients with uti , referred to educational hospitals of shahrekord , iran .	NE

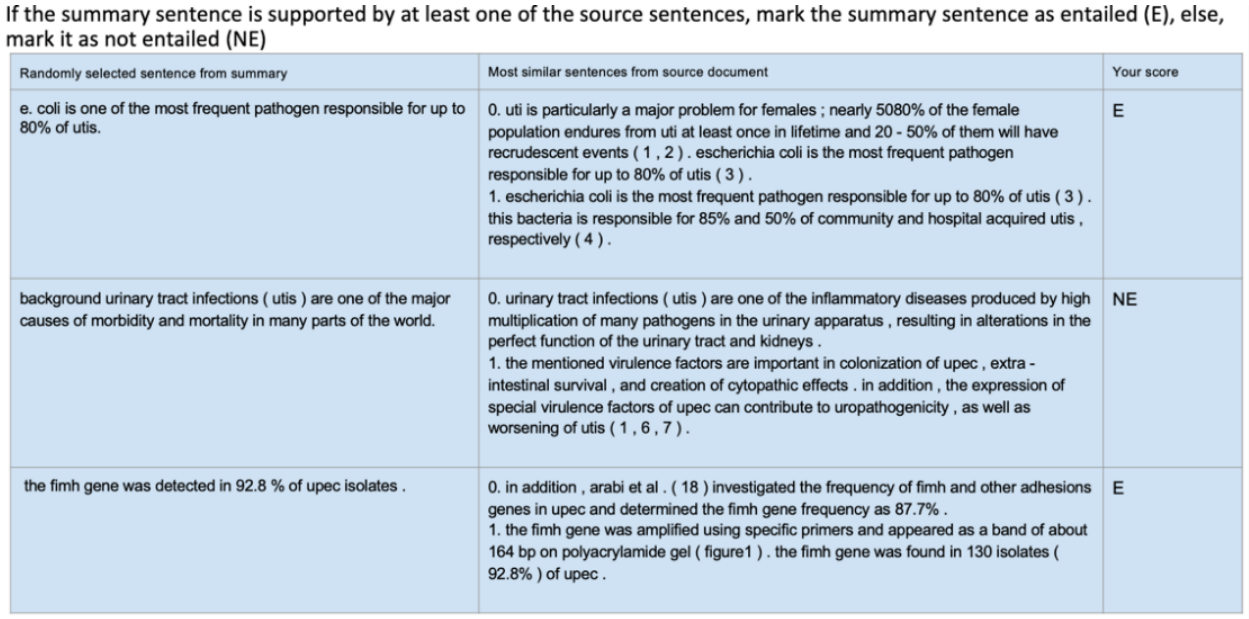
Figure 13. Screenshot of factual consistency annotations by one annotator for the extractive-abstractive method, LEDExtAbs, on a sample from the PubMed data set.



If the summary sentence is supported by at least one of the source sentences, mark the summary sentence as entailed (E), else, mark it as not entailed (NE)

Randomly selected sentence from summary	Most similar sentences from source document	Your rating (E or NE)
urinary tract infections (utis) are one of the inflammatory diseases produced by high multiplication of many pathogens in the urinary apparatus , resulting in alterations in the perfect function of the urinary tract and kidneys .	0. urinary tract infections (utis) are one of the inflammatory diseases produced by high multiplication of many pathogens in the urinary apparatus , resulting in alterations in the perfect function of the urinary tract and kidneys . uti is particularly a major problem for females ; nearly 5080% of the female population endures from uti at least once in lifetime and 20 - 50% of them will have recrudescent events (1 , 2) . 1. this bacteria is responsible for 85% and 50% of community and hospital acquired utis , respectively (4) . uropathogenic e. coli (upec) strains have special virulence factors , including pili or fimbriae , which mediate attachment to uroepithelial and vaginal cells , resistance to human serum bactericidal activity , haemolysin production , and increased amounts of k capsular antigen (5) .	E
of the 130 isolates positive for the fimh gene , 62 (47.7%) and 68 (52.3%) belonged to hospitalized patients and outpatients , respectively .	0. the fimh gene was found in 130 isolates (92.8%) of upec . of the 130 isolates positive for the fimh gene , 62 (47.7%) and 68 (52.3%) belonged to hospitalized patients and outpatients , respectively . 1. was not widely determined from the upec isolated from both hospitalized patients and outpatients with utis referred to educational hospitals of shahrekord , iran , the prevalence of the fimh gene was examined . the presence of the fimh gene was confirmed by pcr and the results indicated that the fimh gene was present in 130 upec isolates (92.8%) ; 62 isolates (47.7%) of hospitalized patients and 68 (52.3%) of outpatients .	E
furthermore , virulence factors of upec strains the fimh gene was found in 130 isolates (92.8%) of upec .	0. the fimh gene was amplified using specific primers and appeared as a band of about 164 bp on polyacrylamide gel (figure 1) . the fimh gene was found in 130 isolates (92.8%) of upec . of the 130 1. was not widely determined from the upec isolated from both hospitalized patients and outpatients with utis referred to educational hospitals of shahrekord , iran , the prevalence of the fimh gene was examined . the presence of the fimh gene was confirmed by pcr and the results indicated that the fimh gene was present in 130 upec isolates (92.8%) ; 62 isolates (47.7%) of hospitalized patients and 68 (52.3%) of outpatients .	NE

Figure 14. Screenshot of factual consistency annotations by one annotator for the DANCER method (Gidiotis and Tsoumakas, 2020) on a sample from the PubMed data set.



If the summary sentence is supported by at least one of the source sentences, mark the summary sentence as entailed (E), else, mark it as not entailed (NE)

Randomly selected sentence from summary	Most similar sentences from source document	Your score
e. coli is one of the most frequent pathogen responsible for up to 80% of utis.	0. uti is particularly a major problem for females ; nearly 5080% of the female population endures from uti at least once in lifetime and 20 - 50% of them will have recrudescent events (1 , 2) . escherichia coli is the most frequent pathogen responsible for up to 80% of utis (3) . 1. escherichia coli is the most frequent pathogen responsible for up to 80% of utis (3) . this bacteria is responsible for 85% and 50% of community and hospital acquired utis , respectively (4) .	E
background urinary tract infections (utis) are one of the major causes of morbidity and mortality in many parts of the world.	0. urinary tract infections (utis) are one of the inflammatory diseases produced by high multiplication of many pathogens in the urinary apparatus , resulting in alterations in the perfect function of the urinary tract and kidneys . 1. the mentioned virulence factors are important in colonization of upec , extra - intestinal survival , and creation of cytopathic effects . in addition , the expression of special virulence factors of upec can contribute to uropathogenicity , as well as worsening of utis (1 , 6 , 7) .	NE
the fimh gene was detected in 92.8 % of upec isolates .	0. in addition , arabi et al . (18) investigated the frequency of fimh and other adhesions genes in upec and determined the fimh gene frequency as 87.7% . 1. the fimh gene was amplified using specific primers and appeared as a band of about 164 bp on polyacrylamide gel (figure1) . the fimh gene was found in 130 isolates (92.8%) of upec .	E

Figure 15. Screenshot of factual consistency annotations by one annotator for the zoning method, outlined in Section 5.2, which generates a highly structured summary on a sample from the PubMed data set.

5.5 Evaluation of abstractive summarisation methods

5.5.1 Automatic metrics

Here, the results from extrinsic and intrinsic evaluations conducted to compare different abstractive summarisation methods in long document contexts are shown. ROUGE-1, -2, and -L metrics (Lin, 2004), and the BERTScore metric (Zhang et al., 2019b) are reported in Table 14. ROUGE-L scores are significantly lower than those reported in previous studies (Gidiotis and Tsoumakas, 2020) because new line tokens were not included between each summary sentence. This approach was taken, despite resulting in lower reported ROUGE-L scores, as including new line tokens between sentences seems like an artificial way to increase the automated score for the purpose of ranking highly on summarisation leader boards. BARTScore's recall, precision and F1 metrics, between the predicted and reference summary, are given in Table 15, along with the LDFACTS metric, details of which can be found in Section 6, which compares the predicted summary to a source document to measure factual consistency.

	PubMed				**ArXiv**			
	R1	**R2**	**RL**	**BERTScore**	**R1**	**R2**	**RL**	**BERTScore**
BERTSumAbs	32.67	13.6	21.21	61.15	32.67	10.47	20.18	61.15
BERTSumExtAbs	32.92	13.73	21.34	61.6	33.24	10.73	20.3	61.4
LED (1024)	41.38	15.36	24.14	65.04	39.39	11.6	21.23	63.17
LEDExtAbs	41.09	14.1	22.43	64.64	<u>39.88</u>	<u>12.22</u>	<u>21.47</u>	<u>63.21</u>
DANCER	**46.13**	**20.76**	**26.89**	**66.97**	**41.79**	**16.00**	**23.08**	**64.02**
THIS WORK	<u>44.93</u>	<u>17.97</u>	<u>25.85</u>	<u>66.94</u>	35.97	10.29	18.32	61.76

Table 14. Results for the automatically generated abstractive summaries, evaluated with the ROUGE-1, -2, -L and BERTScore metrics. Bold font indicates the best results and underlined font indicates the second-best results.

	PubMed				**ArXiv**			
	REC	**PREC**	**F1**	**LDFACTS**	**REC**	**PREC**	**F1**	**LDFACTS**
BERTSumAbs	-11.58	<u>-3.2</u>	-7.40	-4.10	-11.86	<u>-3.15</u>	-7.51	<u>-4.14</u>
BERTSumExtAbs	-11.62	**-3.15**	-7.38	-4.09	-11.63	**-3.10**	-7.40	-4.31
LED (1024)	-6.54	-3.74	<u>-5.14</u>	-4.18	-6.82	-4.95	-5.89	-4.56
LEDExtAbs	-6.57	-4.03	-5.30	-4.00	-6.19	-5.43	<u>-5.81</u>	-4.37
DANCER	**-4.71**	-4.64	**-4.67**	**-1.63**	<u>-5.81</u>	-5.34	**-5.57**	**-2.15**
THIS WORK	<u>-4.98</u>	-5.18	-5.08	<u>-3.74</u>	**-4.92**	-8.46	-6.69	-4.49

Table 15. Results for the automatically generated abstractive summaries, evaluated with BARTScore recall (REC), precision (PREC), F1 and LDFACTs metrics. Bold font indicates the best results and underlined font indicates the second-best results.

5.5.2 Human evaluation

Steen and Markert (2019) and Krishna et al., (2023) both criticised the majority of studies researching text summarisation for not including any human evaluation despite ROUGE being knowingly flawed. Where human evaluation is conducted, they state that most studies do not give details of the annotation design. In response, as part of this work an extrinsic evaluation of abstractive summarisation methods for long document summarisation was conducted and in Section 5.4, details of the human evaluation study were given.

As mentioned in Section 5.4, only the methods which could use text from the entire length of the document to generate their summaries were included in the human evaluation study – i.e., the methods which either use zoning or extractive-abstractive methods. In Table 16 the results of each method are reported, with respect to the human annotated coherence, fluency and factual consistency measures included in the study. As described in Section 5.4, the summaries were ranked between 1-3 for the coherence and fluency measures, with a score of 1 being the best and 3 the worst. Therefore, for these measures, the method will have a score between 1 and 3, with a lower score representing a better summary. For the factual consistency measure, the annotators gave sentences a score of 0 if it was not entailed by the section of the source document provided, and a score of 1 if it was. Therefore, for the factual consistency measure, scores will be between 0-1, with a higher score representing a more factually consistent method.

	PubMed			**ArXiv**		
	COH	**FLU**	**FAC**	**COH**	**FLU**	**FAC**
LEDExtAbs	2.42	2.44	0.28	**1.62**	**1.67**	<u>0.22</u>
DANCER	**1.78**	<u>1.91</u>	**0.75**	2.31	2.22	**0.78**
THIS WORK	<u>1.8</u>	**1.64**	<u>0.42</u>	<u>2.06</u>	<u>2.11</u>	0.13

Table 16. Results of the human evaluation study. Bold font indicates the best results and underlined font indicates the second-best results.

5.5.3 Discussion and analysis of results

Figure 16 and 17 give summaries generated for each method, for one article sampled from the PubMed and ArXiv data sets respectively. In these figures, the different styles of the six methods becomes apparent. BERTSumExtAbs and BERTSumAbs both produce short summaries, whilst the zoning method proposed in Section 5.2 (denoted 'THIS WORK') produces a highly

structured, longer summary. DANCER, LED and LEDExtAbs tend to produce summaries consisting of long blocks of text.

In Table 14, DANCER scored best in terms of ROUGE and BERTScore metrics across both data sets. The method proposed in Section 5.2 was the second highest performing on the PubMed data set, whilst the LEDExtAbs method was the second best performing for the ArXiv data set. It is noticeable that, across the three tables of results, models tended to perform better on the PubMed data set than on the ArXiv data set. This could be due to the noise in the ArXiv data set. Koh et al., (2022) found that over 60% of the ground truth summaries that they sampled from the ArXiv data set contained noise and over 15% of their samples were unreadable because of the noise. Additionally, on inspection of this data set, there are LaTeX artifacts present, as shown in Figure 17, making the text difficult to comprehend.

It is also of note that DANCER outperformed other methods significantly when evaluated with the LDFACTS metric. This metric compares the generated summary with the source document to assess factuality. Additionally, human evaluators scored DANCER significantly better for the factual consistency measure than the other methods included in the human evaluation. On inspection of the task for evaluating factual consistency it was found that, although DANCER is an abstractive method, it has extractive properties, and often phrases or sentences are directly copied from the source text. Consequently, it follows that DANCER is highly consistent with the source document. An example of this can be seen in Figure 14, where the first two sentences sampled for human annotation are direct copies of the source text, given on the right-hand side of the table.

For the BARTScore based evaluation metrics, the two methods which make use of zoning performed best on the recall metric. This is likely due to their ability to generate summaries from the most important sections across the source document. In terms of precision, BERTSumExtAbs and BERTSumAbs performed best. As BERTSumExtAbs and BERTSumAbs generate summaries which are significantly shorter than the other methods, it follows that they performed well with regards to the precision metric, but not as well on the other BARTScore metrics which consider recall in their calculations. Across all automated metrics, it can be observed that directly fine-tuning the LED model did not perform particularly well.

For the coherence and fluency measures, there was no clear winner across the methods evaluated by human annotators. However, the inter-annotator agreement score was low (see Table 17), suggesting that they were difficult measures to assess. Overall, DANCER and the method proposed in this work performed best on the PubMed data set and the LEDExtAbs performed best on the ArXiv data set. This could be due to the PubMed data set being more

consistently structured and therefore more suitable for zoning approaches that classify the sections into zones based on keywords in their headings. It is likely that if more advanced zoning methods were used, e.g., a probabilistic model for classifying the text itself, better performance would be achieved on data sets such as ArXiv, which do not have consistently named section headings. Using a more advanced zoning approach would also enable the processing of more documents as, in the current methodology, whole documents are discarded where there are not enough sections which can be classified.

Some common errors were observed when the generated summaries for each method were inspected. For all methods apart from this work's zoning method, the final sentence of the generated summaries was often only half complete. This can be seen in Figure 16 and 17. This is likely due to the token limit for generation causing the reference summary to be truncated at training time. This could be overcome by increasing the token generation limit. However, since there is a trade-off between the maximum input token limit and the maximum generation token limit, to increase the token generation limit and still fit the training job on a single NVIDIA V100 GPU, the method would be required to truncate the long document further than the 1024 tokens used these experiments. For the DANCER method, it was observed that it occasionally copies the same phrase or sentence repeatedly, resulting in redundancy and reducing its fluency and coherence. This can be seen in Figure 17 where the phrase "we focus on the role of two – body currents…" is repeated several times. Occasionally with the method proposed in Section 5.2, one or more summarised zones are very short and do not make a lot of sense. This is likely due to the method for pairing the document sections with their corresponding section of the reference summary, which sometimes will only include one sentence in the section summary; thus, the model learns to occasionally generate a single, non-coherent sentence.

Overall, these results show that there is still work to do to make abstractive summarisation of long documents fit-for-purpose, with good enough performance for it to be deployed in a production setting with users trusting its outputs. The results of this study indicate that the DANCER method was the strongest method evaluated, and that future work could look to extend this method by using more advanced zoning techniques, as well as potentially combining it with other methods, such as that proposed in Section 5.2, which looks to make the generated summaries more structured and therefore easier to comprehend.

BERTAbs	LEDExtAbs
lymphangiomas of the small bowel mesentery are rare , but have been reported in the mesentery , retroperitoneum , gastrointestinal tract and intra - abdominal solid viscera . the clinical presentation of lymphangiomas varies from incidental discovery on imaging to presenting with an acute abdomen . a 42 - year - old female presented with a history of chronic iron deficiency anemia and menorrhagia for several years . ultrasound demonstrated a complex cystic right adnexal mass measuring 9.7 5.2 6.3 cm that was presumed to originate from the right ovary . transabdom	context : cystic lymphangioma is a benign lymphoproliferative tumor of the mesenteric tract. the clinical presentation of lymphangiomas varies from incidental discovery on imaging to presenting with an acute abdomen. lymphangiomas are very common in the abdomen and most occur in the pelvic cavity.case report : a 42-year - old female with a history of chronic iron deficiency anemia and menorrhagia for several years and later diagnosed with cystic lymphangioma of the small bowel mesentery underwent pelvic ultrasound for investigation of menorrhagia and was found to have a complex cystic right adnexal mass. the mass effect was notably absent, and the cisterna chyli and thoracic duct were not enlarged. the lesion was considered to represent a mesenteric lymphangioma. conclusions : this case report documents a lymphangioma of the small bowel mesentery, incidentally discovered on pelvic ultrasound and initially thought to represent a complex cystic adnexal mass. the clinical presentation of lymphangiomas varies from incidental discovery on imaging to presenting with an acute abdomen.
BERTExtAbs	DANCER
lymphangiomas of the abdomen are rare , but have been reported in the mesentery , retroperitoneum , gastrointestinal tract and intra - abdominal solid viscera . we report a case of a 42 - year - old female who presented with a history of chronic iron deficiency anemia and menorrhagia for several years and later diagnosed with cystic lymphangioma of the small bowel mesentery .	the clinical presentation of lymphangiomas varies from incidental discovery on imaging to presenting with an acute abdomen . mesenteric lymphangiomas , in particular , can result in complications such as intestinal obstruction or volvulus leading to infarction . this case report documents a lymphangioma of the small bowel mesentery , incidentally discovered on pelvic ultrasound and initially thought to represent a complex cystic adnexal mass . intra - abdominal lymphangiomas are rare entities , which although are typically asymptomatic , can result in life - threatening complications . in our case , the lesion was discovered incidentally on imaging for an unrelated clinical indication and demonstrated interval growth at 1 year warranting a surgical referral . as of the present time , the surgical consultation has not yet been conducted and our diagnosis is based on imaging .
LED	THIS WORK
cystic lymphangioma of the mesenteric area is a rare benign mesenteric lymphangioma. our case report documents a lymphangioma of the small bowel mesentery incidentally discovered on pelvic ultrasonography. the lesion initially appeared to be a cystic adnexal mass but later was found to be a lymphangioma. the diagnostic difficulties and complications of this uncommon disease are discussed.	introduction: in this report, we aim to raise awareness and alert the clinician to an unusual case of mesenteric lymphangioma in a male patient who presented with an acute abdomen. the case presented as a large histological mass in the small bowel. at initial abdominal ultrasound, the mass was benign and normal. subsequent investigation revealed the presence of an adnexal mass in the mesentery. a histopathological report revealed a complex adnexal lymphangioma in the mesentery. the histopathology confirmed the diagnosis of mesenteric lymphangioma. in the case, histopathological diagnosis of malignant lymphangioma was confirmed with an operative laparoscopic approach. this case of the mesentery lymphangioma that developed early on abdominal ultrasound emphasizes that lymphangioma must be considered when suspicious of malignant lymphangioma.this case report describes a case of lymphangioma of the small bowel mesentery, incidentally discovered on pelvic ultrasound and initially thought to represent a complex cystic adnexal mass. results: lymphangiomas are common benign conditions in the general population , which generally result from intra - abdominal invasion of the lymph nodes . when they develop , their prognosis is poor , and the diagnosis can be difficult at time of diagnosis . lymphangiomas are benign tumors in the lymph nodes and usually appear in nodular fluid . they are usually detected incidentally on imaging and may pose diagnostic problems . conclusion: intra - abdominal lymphangiomas are rare entities , which although are typically asymptomatic , can result in life - threatening complications . in this case , the lesion was discovered incidentally on imaging for an unrelated clinical indication and demonstrated interval growth at 1 year warranting a surgical referral . as of the present time , the surgical consultation has not yet been conducted and our diagnosis is based on imaging . our findings were those of a thin - walled multiloculated cystic lesion , lacking both mass effect and enhancement , and which insinuated around vessels within the small bowel mesentery .

Figure 16. Automatically generated summaries for an article from the PubMed data set.

BERTAbs	LEDExtAbs
the @xmath0 reaction at @xmath1 (gev / c) @xmath2come has been studied at @xmath3 and @xmath4 mev / c at @xmath5 . the @xmath6 and @xmath7 cross sections have been measured at @xmath8 and @xmath9 mev / nucleon momentum transfer (@xmath10) @xmath3 . a total cross section of @xmath110 mev / @xmath3 has been measured for @xmath11 and @xmath11	a comparison of the predictions of the two - body currents in the proton and neutron double polarization spectra at high energies for proton - nucleus and proton - nucleus collisions is made with the aid of a relativistic transport theory which has been derived from the analysis of inclusive hadronic events. the net effect of the two - body currents is a reduction of the cross sections. the magnitude of the net effect of the two - body currents is found to be @xmath0 for proton and @xmath1 for neutron knockout. comparison with the one - body currents in the solid, shows that the calculated two - body current corrections should be neglected when comparing the results with the solid ones. the impact of the two - body currents on the nuclear cross sections is calculated using the framework of the optical potentials derived from the analysis of the hadronic induced reactions. the predicted relativistic effects attributed to the small components in the bound state wave functions are smallest for reactions with maximal two - body currents, whereas @xmath2 is larger for the latter. it is found that the effect of the two - body currents on the structure functions can be better estimated from the individual reactions.
BERTExtAbs	DANCER
we have performed a systematic study of the @xmath0 (gev / c) @xmath1 and @xmath2 (gev) @xmath3 cross sections for a (@xmath4) @xmath5 (@xmath6) @xmath7 (@xmath8) @xmath9 (@xmath10) @xmath11) @xmath11 (@xmath11) @xmath12 (@xmath12) @xmath13) @xmath10 (@xmath11)	we present the results of a systematic study of the a(@xmath0) reaction at saclay , nikhef , mainz and bates , performed at four - momentum transfers of the order @xmath1 0.2 (gev / c)@xmath3 . in particular , we focus on the a(@xmath4) reaction at saclay , nikhef , mainz and bates , which was performed at four - momentum transfers of the order @xmath1 0.2 (gev / c)@xmath3 . we we investigate the role of two - body currents in the kinematics of a single - knockout process . we focus on the role of two - body currents in the kinematics of a single - knockout process . we investigate the role of two - body currents in the kinematics of a single - knockout process . we focus on the role of two - body currents in the kinematics of a single - knockout process . we investigate the role of two - body currents in the kinematics of a single - knockout process . we focus on the role of two - body currents in the kinematics
LED	THIS WORK
we use a new theoretical approach to study double polarization observables (@xmath0) for a(@xmath1) reactions from finite nuclei. we solve the corresponding two - body current operators numerically in an @xmath2-parameterized framework by using the fermi - gas formalism within a fock - space approach with fermi - liquid coordinates. a comparison with earlier studies by means of a mean - field treatment shows that the experimental results of @xmath0, @xmath1 and @xmath2 measurements are significantly affected by the inclusion of additional two - body currents with different signs of final state interactions at small @xmath3. in particular, in @xmath4 @xmath5 (gev / c)@xmath6, a peak in the magnitude of the @xmath0 correlations in @xmath7 @xmath8 and @xmath9 @xmath9 is observed below the fermi momentum whereas above the fermi momentum the magnitude of the correlation is found to be rather independent. this suggests that the @xmath5 data from two - nucleon current observables (@xmath0, @x	introduction: the role of a background quark is discussed. the importance of the different (@xmath2) regimes for models aiming at understanding nuclei is given by the question whether the nuclear form factors can be successfully incorporated into the (@xmath3) formalism. the two - wave dispersion relation is derived for the case of mesonic gauge invariance and is shown to be insensitive to the experimental uncertainties. results: we present results of a @xmath0o(@xmath1) experiment comparing the spectroscopic characteristics of the @xmath2 wavefunctions obtained by combining the data from the ia with data obtained from the dwia. using a three - body interaction we find that it should lead to small corrections in the spectroscopic behavior of the @xmath2 wavefunctions compared to that produced by the interaction between unpolarized nuclei. conclusion: : : _ we find reasonable estimates of the two - body currents and find them to be of the order of 10 % in the @ xmath1 and @ xmath2 observables . _ the main contribution of the two - body currents is the increase in the cross section when one moves out of quasi - elastic conditions in quasi - elastic kinematics . the contribution of the two - body currents to @ xmath0 and (@ xmath1) cross sections can be calculated with some confidence .

Figure 17. Automatically generated summaries for an article from the ArXiv data set.

6 Automatic evaluation of long document abstractive summarisation

6.1 Challenges in the evaluation of abstractive summarisation of long documents

It is widely acknowledged within the community that, despite ROUGE scoring (Lin, 2004), being the traditional metric for automatic evaluation of text summarisation, it is flawed and does not correlate well with human judgement (Yuan, et al., 2021; Huang et al., 2020; Kryściński et al., 2019), due to not effectively capturing semantic, grammatical, and factual errors.

Factual inconsistency, i.e., when a generated summary is not entailed by its source document, is a well-documented limitation of modern abstractive summarisation methods (Maynez et al., 2020; Wallace et al., 2021). There have been efforts to develop improved, reference-free, metrics for measuring factual consistency (Scialom et al., 2021; Kryściński et al., 2020; Yuan, et al., 2021). A reference-free metric is a metric which does not require gold summaries for its evaluation of generated summaries and instead compares predicted summaries to their source documents. However, as PLMs are used in SOTA reference-free methods, the source document must be truncated due to hardware memory limitations. Thus, studies proposing these metrics run their evaluations on short document summarisation data sets (Hermann et al., 2015; Grusky et al., 2018; Narayan et al., 2018; Ribeiro et al., 2022) and the automated metrics they propose do not extend well when applied to evaluation of long document summarisation (Koh et al., 2022).

In addition to the limitations around memory complexity for SOTA automated metrics, the effort required to manually annotate long documents is likely a large contributing factor to the lack of long document summarisation data sets with human annotations which would enable research into automated metrics for evaluation of long document summarisation. This is particularly apparent for reference-free measures, such as factual consistency, where the annotator needs to read the long source document to assess whether a summary is factual.

To this end, in this work, the LDFACTS metric is proposed. This metric is intended for assessing the factual consistency of abstractive summarisation of long documents. LDFACTS is reference-free and was developed by adapting an existing metric, BARTScore (Yuan, et al., 2021), to consider an entire source document. The efficacy of the metric is assessed by evaluating the correlation of it and other automated metrics with human judgement. It is shown that that LDFACTS achieves a strong correlation with human annotations of factuality.

6.2 The LDFACTs metric

BARTScore (Yuan, et al., 2021) uses BART (Lewis et al., 2020) to calculate the log probability of generating a sequence of text, given a second sequence. It is a flexible framework which can

either be used to compare a predicted and reference summary to calculate precision, recall and F1 scores, or to compare a predicted summary with its source document to measure factual consistency. However, the source documents in long document data sets are on average 8.3x longer than their short document counterparts (Koh et al., 2022). Consequently, the PLMs used in SOTA metrics for assessing factuality, including in BARTScore, must truncate, on average, over half of the tokens of a source document in long document data sets (Koh et al., 2022). This makes them unsuitable for evaluating factual consistency of long documents. In this section, a metric LongDocFACTScore (LDFACTS), an adaptation of BARTScore's reference-free setting, is proposed. This metric is intended for the evaluation of factual consistency of long document abstractive summarisation.

The source document D is split into sentences using the nltk library[26], such that $D = \langle s_i, i \in I \rangle$ and the same is done for the generated summary $S = \langle s_j, j \in J \rangle$. For each of these sentences, sentence embeddings (Reimers and Gurevych, 2019) are generated using the sentence-transformers library[27] initialised with the 'bert-base-nmli-mean-tokens' model[28]. For each sentence in the predicted summary s_j, the cosine similarity between its sentence embedding and the sentence embedding of each sentence in the source document s_i is calculated. D is then reindexed by the cosine similarity scores, so that the new index k is sorted by $\arg\max_{i \in I} \left(cosine\_similarity(s_j, s_i) \right)$. The $K = 3$ most similar source document sentences are selected. These three highest scoring sentences are each concatenated with their preceding and following sentences, thus giving $s_k^* = s_{k-1} + s_k + s_{k+1}$, to create the sequence of slightly longer text snippets. For each text snippet s_k^*, BARTScore is calculated between it and the generated sentence s_j. The maximum BARTScore value (as they are negative, this means the score closest to zero) is taken as the assigned score for the sentence s_j. The method, applied to an individual sentence s_j of the generated summary, is illustrated in Figure 18. The method is repeated for each sentence, s_j, in S, and the average score for the summary is calculated by:

$$LDFACTS = \frac{1}{J} \sum_{j=1}^{J} \max_{k=\{1,2,3\}} (\text{BARTScore}(s_j | s_k^*)) \quad (7).$$

Consequently, there are two fundamental differences between LDFACTS and BARTScore. The first difference is that LDFACTS considers sections of text from the full length of the source document in its calculation whereas BARTScore truncates the source document to the first 1024

[26] https://www.nltk.org
[27] https://github.com/UKPLab/sentence-transformers
[28] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

tokens. If a predicted summary includes content from the latter part of a long document, it will be ignored with BARTScore, which is a problem when assessing factual consistency of long document summarisation. The second significant difference between the two metrics is that in LDFACTS the BARTScore calculation is done on short sections of text at one time, comparing one sentence in the predicted summary to a short section of the source document. In contrast, BARTScore's original implementation compares a full predicted summary with a full source document (or as much of it as it can fit within its token limit).

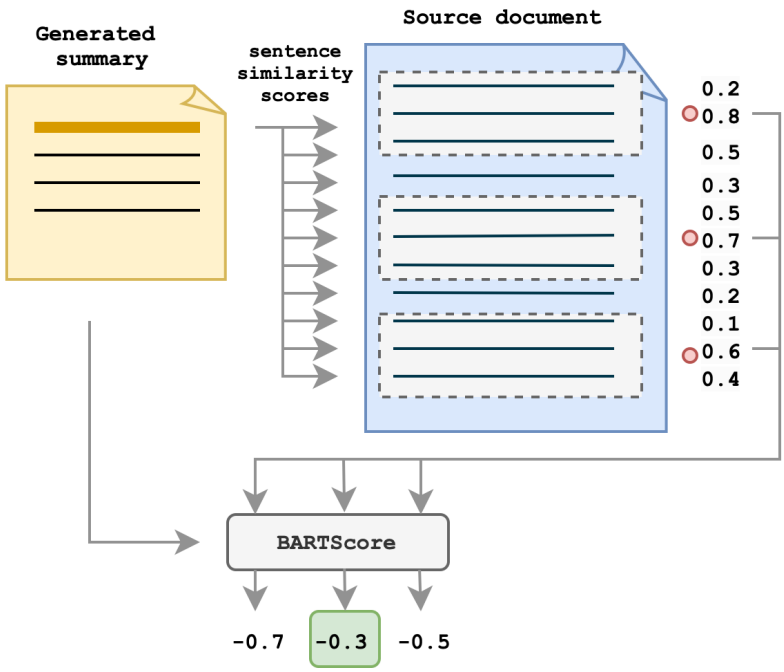


Figure 18. Calculation of the score for an individual sentence of a generated summary.

6.3 Extrinsic evaluation metrics for long document summarisation

6.3.1 Computational set up

As baselines to which to compare LDFACTS, ROUGE[29] (Lin, 2004), BERTScore (Zhang et al., 2019b) and BARTScore's (Yuan, et al., 2021) F1, precision and recall metrics were implemented, which all compare predicted and reference summaries. ROUGE scores measure the overlap in sequences of words between two texts, whilst BERTScore uses measures of cosine similarity between BERT-based (Devlin et al., 2019) token embeddings to assess the similarity. Additionally, other SOTA reference-free metrics, which have previously shown improved

[29] https://huggingface.co/spaces/evaluate-metric/rouge

correlation with the human judgement of factual consistency for short documents, were implemented. FactCC (Kryściński et al., 2020) uses a fine-tuned BERT-based classifier to predict, for each sentence of a summary, whether it is correct or incorrect, given its source document. QuestEval (Scialom et al., 2021) uses T5-based models (Raffel et al., 2020) for a question generation and answering approach. BARTScore's factuality metric was implemented as well as LDFACTS. BARTScore and LDFACTS were implemented with the 'bart-large' model[30]. The default settings were used for all metrics. All experiments were run on a single NVIDIA v100 GPU and all metrics, apart from ROUGE, make use of the GPU compute.

6.3.2 Inter-annotator agreement

Table 17 shows the inter-annotator agreement (IAA) of the human annotated data, calculated using the Krippendorff's alpha metric[31] (Krippendorff, 2011). The IAA for factual consistency was shown to be relatively high, averaging at 0.65 across the two data sets. In contrast, IAA for the coherence and fluency rankings was variable and significantly lower. This could be due to these measures being more subjective than the factual consistency measure. Overall, the agreement in the ArXiv data set is lower than for PubMed. This could be due to the noise in the ArXiv data set (Koh et al., 2022), e.g., the LaTeX artifacts present in the dataset, and the highly domain-specific nature of the data set.

	COH	**FLU**	**FAC**	**Avg.**
PubMed	0.33	0.44	0.76	0.51
ArXiv	0.44	0.28	0.54	0.42
Avg.	**0.39**	**0.36**	**0.65**	

Table 17. IAA of the human-annotated data for the coherence (COH), fluency (FLU) and factuality (FAC) metrics.

Metric	**PubMed**	**ArXiv**
FactCC	-0.06	-0.16
QuestEval	0.25	0.23
BARTScore (Factuality)	<u>0.39</u>	<u>0.49</u>
LDFACTS	**0.61**	**0.61**

Table 18. Kendall's tau correlations between the human factual consistency annotations and the four metrics which aim to measure factual consistency.

6.3.3 Correlation between metrics on long document summarisation

[30] https://huggingface.co/facebook/bart-large
[31] https://github.com/grrrr/krippendorff-alpha

Figure 19 gives a matrix of pairwise Kendall's tau (Kendall, 1938) correlations[32], calculated for all automated and human evaluation metrics included in this study. In this analysis, Kendall's tau correlations were calculated instead of Spearman correlations due to being more robust for data sets with smaller sample sizes. In Figure 19, BARTSc denotes BARTScore.

To calculate the correlations between human measures of performance and automatic metrics, for each measure, the scores were averaged over the three different annotators for each unique summary, thus giving a single score for each unique summary. The scores for each metric were then compared per each unique summary. Consequently, for each pair of metrics, the correlation was calculated between 90 summaries (3 unique summaries generated by different methods, created for 15 source documents, for 2 data sets).

Table 18 breaks down, by data set, the correlations between the human factual consistency annotations and the SOTA reference-free, automated metrics which aim to measure factual consistency. In both Table 18 and Figure 19, LDFACTS can be seen to have correlated better with the human judgement of factual consistency than any other metric. As Table 15 is split by data set, the lower IAA reported on the ArXiv data set can be taken into consideration when inspecting the results. In this table, the BARTScore results were higher on the ArXiv data set, where IAA was lower, whereas LDAFACTS performed consistently across both the PubMed and ArXiv data sets.

Comparatively, it was found that both FactCC and QuestEval showed a low correlation with human judgement. BARTScore's faithfulness metric had a reasonable correlation with the human factual consistency annotations, however, since it is required to truncate the source document, one could expect that it would become decreasingly correlated with human judgement as it is used to score texts of increasing length. No strong correlation can be seen between ROUGE or BERTScore metrics and any human annotated measures. Furthermore, only weak correlations between human measures of coherence and fluency and automated metrics were found in this long document study, thus indicating a need for further research into automated metrics which capture these measures.

Table 19 compares the average time taken, in seconds, to run each reference-free automated metric that aims to measure factual consistency. The time taken to score fifteen summaries is reported. Table 19 shows that LDFACTS is second fastest metric, despite evaluating the generated summary against the entire source document, rather than a truncated version.

[32] https://scipy.org

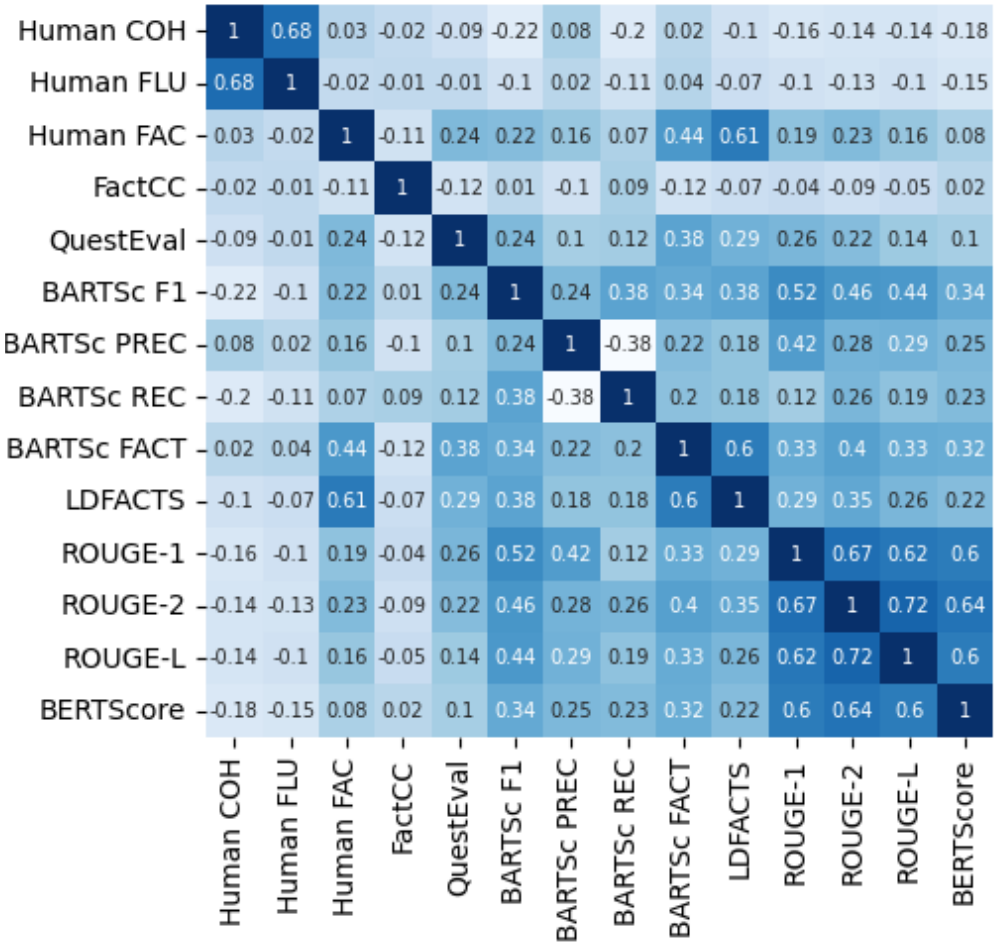


Figure 19. Pairwise Kendall's tau correlations between human and automated metrics.

Metric	Time taken (s)
FactCC	24
QuestEval	95
BARTScore	1
LDFACTS	8

Table 19. Average time taken (s) to run each factuality metric over 15 long document summaries. Bold font indicates the best results and underlined font indicates the second-best results.

6.3.4 Parameter study

In this section, the effects that different parameter settings have on the LDFACTS metric are studied. Table 20 and 21 provide Kendall's tau correlations of different implementations of BARTScore and LDFACTS with the human measure of factuality.

Table 20 shows the effect of varying K, the maximum number of candidate similar source sentences used for the BARTScore calculation. In LDFACTS, the maximum scoring BARTScore over the K candidate sentences is used at the score for the reference summary sentence. As in the original implementation of LDFACTS, each candidate similar source sentence was concatenated with a surrounding sentence from either side in this experiment. Row 1 of Table 20 gives the BARTScore results, and row 2 gives results of a baseline, which was implemented by excluding the sentence similarity calculation all together. Where the similarity calculation step was removed, BARTScore was simply calculated between each sentence of the predicted summary and the original source article, truncated to 1024 tokens; the scores were then averaged over each sentence of the predicted summary. In Table 20, all settings of LDFACTS are shown to outperform the baselines of BARTScore and LDFACTs (-sentence similarity). This is expected as neither BARTScore nor LDFACTs (-sentence similarity) consider the full length of a source document. Overall, the effects of varying K were seen to be small. Again, this is expected as the maximum BARTScore value of the K sentences is used in the LDFACTS metric, and it is likely that the highest scoring sentences with BARTScore correlate well with the most similar sentence embeddings. $K = 3$ averages as the best parameter across the two data sets.

Although not explicitly calculated, it is of note that by selecting the K=3 candidate sentences, BARTScore was calculated for around 1-2% of sentences in the PubMed and ArXiv data sets. By increasing the number of candidate similar sentences, the metric would become increasingly less efficient and, by extension, less suitable for use on long documents.

In Table 21, the number of candidate sentences was kept constant at $K = 3$ and the effect of concatenating the source sentence with the previous and following sentence(s) was examined. The correlation on the PubMed data set improved when no sentences were concatenated while the correlation on the ArXiv data set improved when one or two sentences were concatenated either side of the selected sentence. However, on average over the two data sets, there was little variation in the Kendall's tau correlation.

Method	PubMed	ArXiv	Average
BARTScore	0.39	0.49	0.440
LDFACTS (-sentence similarity)	0.44	0.37	0.405
LDFACTS K=1	0.61	0.60	<u>0.605</u>
LDFACTS K=3	0.61	0.61	**0.610**
LDFACTS K=5	0.60	0.60	0.600
LDFACTS K=7	0.59	0.61	0.600
LDFACTS K=9	0.59	0.61	0.595
LDFACTS K=11	0.57	0.61	0.590

Table 20. The effect of varying the number of similar sentences considered for the LDFACTS calculation on Kendall's tau correlation with human judgements of factuality. Bold font indicates the best results and underlined font indicates the second-best results.

Method	PubMed	ArXiv	Average
LDFACTS ($s_k^* = \ s$)	0.63	0.58	<u>0.605</u>
LDFACTS ($s_k^* = s_{k-1} + s_k + s_{k+1}$)	0.61	0.61	**0.610**
LDFACTS ($s_k^* = s_{k-2} + s_k + s_{k+2}$)	0.58	0.61	0.595

Table 21. The effect of varying the number of source document sentences concatenated for the LDFACTS calculation on Kendall's tau correlation with human judgement of factuality. Bold font indicates the best results and underlined font indicates the second-best results.

6.3.5 Correlation between metrics on short document summarisation

Although the intended use of LDFACTS is to evaluate the factual consistency of abstractive summarisation for long documents, the analysis conducted by Yuan et al. (2021) was repeated to evaluate LDFACTS against other automated metrics on a variety of human annotated, short document, abstractive summarisation data sets, to validate its performance in this setting. Their human annotated data and adapted code[33] was used to report the Spearman correlation results on three data sets, RealSumm (Bhandari et al., 2020); SummEval (Fabbri et al., 2021); and NER18 (Grusky et al., 2018) in Table 22. In this analysis, Spearman correlations, rather than Kendall's tau correlations were calculated to enable a direct comparison with the original analysis conducted by Yuan et al., (2022). BARTScore was re-implemented with a BART-large PLM and the scores for each summary in the data sets were regenerated. In addition to LDFACTS and the re-

[33] https://github.com/neulab/BARTScore

implemented BARTScore results, the scores for other metrics are taken directly from Yuan et al., (2021) and included. Following implementation by Yuan et al., (2019), the coverage metric (COV) was calculated with BARTScore's recall metric $(S^*|R,\theta)$, where S^* represents the system generated summary and S the reference summary. Coherence (COH), factuality (FAC), fluency (FLU), informativeness (INFO) and relevance (REL) are all calculated using what they describe as a faithfulness (or factuality) metric $p(S^*|D,\theta)$, where D represents the source document. Table 23 shows a similar analysis, in which the factuality metric was also used. Here, the accuracy scores calculated on human annotated data are given for the Rank19 data set (Falke et al., 2019). Additionally, the Pearson correlation between the automated metrics and the human factuality annotations for the two QAGS data sets (Wang et al., 2020b) are shown. For both tables, where the BARTScore results are taken directly from the original paper, this is denoted BARTScore*. Where the results from the re-implemented version are reported, this is denoted BARTScore (this work). These tables show that BARTScore and LDFACTS perform comparably in their ability to align to human judgment when used to evaluate abstractive summaries generated from short documents.

	RealSumm	SummEval				NER18			
	COV	**COH**	**FAC**	**FLU**	**INFO**	**COH**	**FLU**	**INFO**	**REL**
ROUGE-1	0.498	0.167	0.16	0.115	0.326	0.095	0.104	0.13	0.147
ROUGE-2	0.423	0.159	0.157	0.129	0.318	0.161	0.12	0.188	0.195
ROUGE-L	0.488	0.128	0.115	0.105	0.311	0.064	0.072	0.089	0.106
MoverScore	0.372	0.184	0.187	0.159	0.29	0.026	0.048	0.079	0.091
BERTScore	0.440	0.284	0.11	0.193	0.312	0.147	0.17	0.131	0.163
BARTScore*	0.441	0.322	0.311	0.248	0.264	0.679	0.67	0.646	0.604
BARTScore (this work)	**0.451**	0.308	0.313	0.250	0.265	**0.689**	**0.672**	**0.666**	**0.606**
LDFACTS	0.402	**0.356**	**0.349**	**0.296**	**0.252**	0.589	0.59	0.536	0.524

Table 22. Spearman correlation results between automated metrics and human annotated data on the RealSumm, SummEval and NER18 data sets. Bold font indicates the best results and underlined font indicates the second-best results.

	Rank19 Accuracy	QAGS_CNN Pearson	QAGS_XSUM Pearson
ROUGE-1	0.587	0.338	-0.008
ROUGE-2	0.63	0.459	<u>0.097</u>
ROUGE-L	0.568	0.357	0.024
MoverScore	0.713	0.414	0.054
BERTScore	0.713	0.576	0.024
FactCC	0.700	-	-
QAGS	<u>0.721</u>	0.545	**0.175**
Human	**0.839**	-	-
BARTScore*	0.684	0.661	0.009
BARTScore (this work)	0.681	**0.657**	0.006
LDFACTS	0.681	<u>0.648</u>	0.036

Table 23. Accuracy scores and Pearson correlations for the Rank19 and QAGS data sets. Bold font indicates the best results and underlined font indicates the second-best results.

6.3.6 Discussion of LDFACTS

In this section, it has been shown that, in line with previous research (Koh et al., 2022), existing automated metrics for assessing factual consistency which have previously shown good performance on short document data sets do not extend well to long document settings. However, the LongDocFACTScore (LDFACTS) metric, a reference-free metric for evaluating factual consistency of long document abstractive summarisation, has a stronger correlation with the human judgement of factual consistency for the long document data sets included in the study than any other metric evaluated.

For this evaluation, expert annotators were recruited as the long document data sets evaluated were domain specific. It is difficult to recruit large numbers of expert annotators and therefore work building on this should consider conducting a larger human evaluation study with more annotators evaluating more documents. One issue for both LDFACTS and BARTScore is that the raw scores are negative numbers, with scores closer to zero indicating a better summary. These scores are useful when comparing different summarisation methods against each other, but when calculated for an individual summary, could be difficult to interpret. Therefore, future research should investigate what constitutes a 'good score' for these metrics, to enable the use of these metrics to evaluate models for putting into production.

There were differences between the methods by which LDFACTS was evaluated and by which the human evaluation of factuality was conducted. Firstly, the human annotators were presented

with only three sentences sampled from the predicted summary, while LDFACTS considers all generated sentences when evaluating a predicted summary. Additionally, the human reviewers were asked to consider both text snippets presented to them to make a judgement on entailment, whereas LDFACTS only considers the maximum scoring text snippet, therefore it only considers one in its calculation. However, it is acknowledged that there are some similarities between the way that the human evaluation was conducted and the method that LDFACTS takes, which could raise concerns about bias of results. The reason for conducting the human analysis in this way was due to the source documents for these data sets being extremely long. Asking an annotator to go through all source documents searching for each fact would be highly labour intensive and it is likely they would have missed facts due to the quantity of information they would need to comprehend. As the annotators were asked to mark something as 'entailed' when they were confident a statement in the predicted summary was true given the text snippets presented to them, it's unlikely entailed documents were marked incorrectly. However, admittedly, there is a risk that both the LDFACTS metric and the human evaluation missed a sentence in the similarity calculation which would have supported a document being entailed. In future work, a more thorough analysis of the source documents should be done in the cases where annotators marked summaries as 'not entailed'.

Furthermore, recently, a long document human evaluation framework and data set for the summarisation task has been proposed (Krishna et al., 2023). Future work should look to evaluate LDFACTS against this framework.

7 Conclusion

This work begins to address some of the limitations of automatic long document text summarisation. Pretrained language models (PLMs) have enabled vast improvements across an array of natural language processing tasks, including text summarisation. However, they have also introduced limitations, many of which are around the computational resources required to train and deploy such models. This has resulted in most of the research on automatic text summarisation to date focusing on the comprehension of short documents. Since the value of automatic text summarisation comes from distilling long documents into shorter texts, whilst still communicating the same key information, by truncating documents before summarising them, or only summarising short documents, much of the value is lost.

The aims of this research were: (1) to evaluate existing methods for extractive and abstractive long document text summarisation and to propose novel summarisation methods, (2) to study and propose methods for domain-specific summarisation, and (3) to research the efficacy of automatic evaluation metrics for assessing factual consistency of long document abstractive summarisation. In this work, methods were explored which make use of SOTA PLMs, but which also use strategies to overcome some of the limitations that they traditionally impose. To address Research Aim 1, in Section 3, extractive methods were evaluated and an unsupervised hybrid abstractive-extractive model, GenCompareSum. This model harnesses the semantic understanding of PLMs but uses a strategy which does not impose a token limit on the document length. Furthermore, being an unsupervised approach, it does not require human annotated training data for the summarisation task. This method showed strong performance compared to unsupervised and SOTA supervised extractive baselines. Additionally, in Section 5, intrinsic and extrinsic evaluation was conducted to assess the efficacy of abstractive methods for the summarisation of long documents. Since many long document data sets are highly domain-specific, but most PLMs are trained on data of the general domain, in this work a lightweight framework for injecting domain knowledge, KeBioSum, was proposed, addressing Research Aim 2. Again, this approach demonstrated improved performance over strong extractive baselines, and was shown to be particularly useful when a PLM fine-tuned on in-domain data is unavailable. Throughout this work, to evaluate the different summarisation methods, the ROUGE scoring metric was used – the standard and most popular metric for evaluating text summarisation. However, this metric is understood to be flawed, particularly when being used to evaluate abstractive summarisation methods. Therefore, to answer Research Aim 3, in Section 6, the LDFACTS metric was proposed for evaluating the factual consistency of long documents. It was shown that LDFACTS correlated better with human annotations of factual consistency than any existing metric evaluated in a long document setting.

In Section 5 of this work, it is shown that the abstractive methods evaluated are not factually consistent and it is apparent that there is much more work to be done to develop reliable models suitable for deployment in high-risk production settings, such as healthcare. Furthermore, in fields such as healthcare, where factual consistency is vital, extractive approaches are advantageous in that they allow for a reader to quickly refer back to the source document and check the context of the text which appears in the summary. Future work should look to explore more advanced summarisation methods which can use extractive techniques to improve factual consistency and trust in high-risk settings. As Large Language Models (LLMs) have shown promise in their ability to be able to cite sources and summarise extractive content, an interesting future direction of

research could be to use sentence embeddings to find relevant sections in long documents (or even multiple documents) and use a LLM to summarise them, citing their sources.

Another interesting direction for future work would be to understand how the methods proposed in this work could extend to other domains and settings. For example, GenCompareSum should easily extend to multi-document settings, however, in order to make the final summary coherent to a user, some thought should be put in to how to communicate the origins of each extractive segment. Furthermore, it would be interesting to see how the adapter-fusion approach in KeBioSum extends to other technical domains, such as legal or technical scientific documentation. LDFACTS should extend to other domains and multi-document summarisation data sets, however, an interesting direction for future work would be to explore how it can effectively evaluate highly abstractive but factually consistent summaries of multiple sources, as currently the score it produces will favour more extractive and single-document approaches.

Although this work begins to address some limitations of PLMs for long document summarisation and follow-on research for each of the methods explored is suggested in this thesis, there are some bigger, more general challenges in the field of text summarisation which need addressing to enable significant advances. Firstly, one can hypothesise that issues with fluency, coherence and factual consistency are at least partly a result of noisy training data (for example a fact appearing in the summary of a training example which doesn't appear in its source document). Due to the quantity of data required to train a neural supervised summarisation model, data sets for text summarisation tend to be automatically created, and a proxy is used for the target summary, e.g., the abstract is used in data sets of scientific papers, and titles are used as reference summaries in some news article summarisation data sets. In all domains of machine learning, clean, curated, in-domain data is required to train the best models. Therefore, to make significant advancements in text summarisation, methods for curation of better-quality data sets are vital. An interesting future direction of work could be to use methods such as LDFACTS to curate training data and remove noisy examples where target summaries are factually inconsistent with a source document. Furthermore, as highlighted in Section 6, improved metrics are required for capturing human measures of quality, particularly fluency and coherence. Until we have strong metrics that can capture these measures, we will struggle to know whether to performance of our models is adequate. Lastly, although this research attempted to keep GPU utilization at a practical level during the experiments conducted, this level of compute is not environmentally or economically sustainable if we expect all companies, researchers, and institutions to use the technologies we are developing. Therefore, to significantly progress the field of text summarisation, and deep-

learning more generally, new, less resource-intensive strategies must be developed for the training and usage of these models.

References

Afzal, Muhammad, Fakhare Alam, Khalid Mahmood Malik, and Ghaus M. Malik. "Clinical context–aware biomedical text summarization using deep neural network: model development and validation." *Journal of medical Internet research* 22, no. 10 (2020): e19810.

Akiyama, Kazuki, Akihiro Tamura, and Takashi Ninomiya. "Hie-BART: Document Summarization with Hierarchical BART." *NAACL-HLT 2021* (2021): 159.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Bajaj, Payal, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder et al. "Ms marco: A human generated machine reading comprehension dataset." *arXiv preprint arXiv:1611.09268* (2016).

Hoffmann, Falk, Katharina Allers, Tanja Rombey, Jasmin Helbach, Amrei Hoffmann, Tim Mathes, and Dawid Pieper. "Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019." Journal of Clinical Epidemiology138 (2021): 1-11.

Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615-3620. 2019.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." *arXiv preprint arXiv:2004.05150* (2020).

Bhandari, Manik, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. "Re-evaluating Evaluation in Text Summarization." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9347-9359. 2020.

Bishop, Jennifer, Qianqian Xie, and Sophia Ananiadou. "GenCompareSum: a hybrid unsupervised summarization method using salience." In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 220-240. 2022.

Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30, no. 1-7 (1998): 107-117.

Brockmeier, Austin J., Meizhi Ju, Piotr Przybyła, and Sophia Ananiadou. "Improving reference prioritisation with PICO recognition." *BMC medical informatics and decision making* 19 (2019): 1-14.

Bui, Duy Duc An, Guilherme Del Fiol, John F. Hurdle, and Siddhartha Jonnalagadda. "Extractive text summarization system to aid data extraction from full text in systematic review development." *Journal of biomedical informatics* 64 (2016): 265-272.

Cachola, Isabel, Kyle Lo, Arman Cohan, and Daniel S. Weld. "TLDR: Extreme Summarization of Scientific Documents." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4766-4777. 2020.

Cao, Meng, Yue Dong, and Jackie Chi Kit Cheung. "Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3340-3354. 2022.

Cao, Meng, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. "Factual Error Correction for Abstractive Summarization Models." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6251-6258. 2020.

Chopra, Sumit, Michael Auli, and Alexander M. Rush. "Abstractive sentence summarization with attentive recurrent neural networks." In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 93-98. 2016.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).

Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W. and Goharian, N., 2018, June. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 615-621).

Contractor, Danish, Yufan Guo, and Anna Korhonen. "Using argumentative zones for extractive summarization of scientific articles." In *Proceedings of COLING 2012*, pp. 663-678. 2012.

Demner-Fushman, Dina, and Jimmy Lin. "Answering clinical questions with knowledge-based and statistical techniques." *Computational Linguistics* 33, no. 1 (2007): 63-103.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, pp. 4171-4186. 2019.

Dong, Yue, Andrei Mircea, and Jackie Chi Kit Cheung. "Discourse-Aware Unsupervised Summarization for Long Scientific Documents." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1089-1102. 2021.

Durmus, Esin, He He, and Mona Diab. "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5055-5070. 2020.

Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of artificial intelligence research* 22 (2004): 457-479.

Fabbri, Alexander R., Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. "Summeval: Re-evaluating summarization evaluation." *Transactions of the Association for Computational Linguistics*9 (2021): 391-409.

Fabbri, Alexander Richard, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. "QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2587-2601. 2022.

Falke, Tobias, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2214-2220. 2019.

Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894-6910. 2021.

Gehrmann, Sebastian, Yuntian Deng, and Alexander M. Rush. "Bottom-up abstractive summarization." *arXiv preprint arXiv:1808.10792* (2018)

Gidiotis, Alexios, and Grigorios Tsoumakas. "A divide-and-conquer approach to the summarization of long documents." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 3029-3040.

Gnehm, Ann-Sophie. "Text zoning for job advertisements with bidirectional LSTMs." (2018): 1-9.

Goodwin, Travis R., Max E. Savery, and Dina Demner-Fushman. "Towards zero-shot conditional summarization with adaptive multi-task fine-tuning." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2020, p. 3215. NIH Public Access, 2020.

Grail, Quentin, Julien Perez, and Eric Gaussier. "Globalizing BERT-based transformer architectures for long document summarization." In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, pp. 1792-1810. 2021.

Grusky, Max, Mor Naaman, and Yoav Artzi. "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 708-719. 2018.

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing." *ACM Transactions on Computing for Healthcare (HEALTH)* 3, no. 1 (2021): 1-23.

Guo, Mandy, Joshua Ainslie, David C. Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. "LongT5: Efficient Text-To-Text Transformer for Long Sequences." In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724-736. 2022.

He, Bin, Xin Jiang, Jinghui Xiao, and Qun Liu. "Kgplm: Knowledge-guided language model pre-training via generative and discriminative learning." *arXiv preprint arXiv:2012.03551*(2020).

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. "Teaching machines to read and comprehend." *Advances in neural information processing systems* 28 (2015).

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-efficient transfer learning for NLP." In *International Conference on Machine Learning*, pp. 2790-2799. PMLR, 2019.

Huang, Dandan, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. "What Have We Achieved on Text Summarization?." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 446-469. 2020.

Huang, Luyang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. "Efficient Attentions for Long Document Summarization." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419-1436. 2021.

Huang, Yichong, Xiachong Feng, Xiaocheng Feng, and Bing Qin. "The factual inconsistency problem in abstractive text summarization: A survey." *arXiv preprint arXiv:2104.14839*(2021).

Jin, Qiao, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. "PubMedQA: A Dataset for Biomedical Research Question Answering." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567-2577. 2019.

Kamoi, Ryo, Tanya Goyal, and Greg Durrett. "Shortcomings of Question Answering Based Factuality Frameworks for Error Localization." *arXiv preprint arXiv:2210.06748* (2022).

Kendall, Maurice G. "A new measure of rank correlation." *Biometrika* 30, no. 1/2 (1938): 81-93.

Kiritchenko, Svetlana, and Saif Mohammad. "Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 465-470. 2017.

Koh, Huan Yee, Jiaxin Ju, Ming Liu, and Shirui Pan. "An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics." *ACM computing surveys* 55, no. 8 (2022): 1-35

Kornilova, Anastassia, and Vladimir Eidelman. "BillSum: A Corpus for Automatic Summarization of US Legislation." In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 48-56. 2019.

Koupaee, Mahnaz, and William Yang Wang. "Wikihow: A large scale text summarization dataset." *arXiv preprint arXiv:1810.09305* (2018).

Krippendorff, Klaus. "Computing Krippendorff's alpha-reliability." (2011).

Krishna, Kalpesh, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. "LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization." arXiv preprint arXiv:2301.13298 (2023).

Kryściński, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher. "Evaluating the Factual Consistency of Abstractive Text Summarization." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332-9346. 2020.

Kryściński, Wojciech, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. "Neural Text Summarization: A Critical Evaluation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 540-551. 2019.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871-7880. 2020.

Liang, Xinnian, Shuangzhi Wu, Mu Li, and Zhoujun Li. "Improving unsupervised extractive summarization with facet-aware modeling." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1685-1697. 2021.

Lin, Bill Yuchen, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. "CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823-1840. 2020.

Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In *Text summarization branches out*, pp. 74-81. 2004.

Liu, Yang, and Mirella Lapata. "Text Summarization with Pretrained Encoders." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730-3740. 2019.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019)

Liu, Yixin, Ansong Ni, Linyong Nan, Budhaditya Deb, Chenguang Zhu, Ahmed H. Awadallah, and Dragomir Radev. "Leveraging locality in abstractive text summarization." *arXiv preprint arXiv:2205.12476* (2022).

Liu, Yixin, and Pengfei Liu. "SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 1065-1072. 2021.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. "S2ORC: The Semantic Scholar Open Research Corpus." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969-4983. 2020

Luhn, Hans Peter. "The automatic creation of literature abstracts." *IBM Journal of research and development* 2, no. 2 (1958): 159-165.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP natural language processing toolkit." In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55-60. 2014.

Mao, Yuning, Xiang Ren, Heng Ji, and Jiawei Han. "Constrained abstractive summarization: Preserving factual consistency with constrained generation." *arXiv preprint arXiv:2010.12723* (2020).

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. "On Faithfulness and Factuality in Abstractive Summarization." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906-1919. 2020.

Meng, Zaiqiao, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. "Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4672-4681. 2021.

Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404-411. 2004.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Mizuta, Yoko, Anna Korhonen, Tony Mullen, and Nigel Collier. "Zone analysis in biology articles as a basis for information extraction." *International journal of medical informatics* 75, no. 6 (2006): 468-487.

Möller, Timo, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. "COVID-QA: A question answering dataset for COVID-19." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. 2020.

Nan, Feng, Cicero dos Santos, Henghui Zhu, Patrick Ng, Kathleen Mckeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. "Improving Factual Consistency of Abstractive Summarization via Question Answering." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6881-6894. 2021.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797-1807. 2018.

Nenkova, Ani, and Lucy Vanderwende. "The impact of frequency on summarization." *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005* 101 (2005).

Nentidis, Anastasios, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. "Overview of BioASQ 2021: the ninth BioASQ challenge on large-scale biomedical semantic indexing and question answering." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pp. 239-263. Springer International Publishing, 2021.

Nogueira, Rodrigo, Jimmy Lin, and A. I. Epistemic. "From doc2query to docTTTTTquery." *Online preprint* 6 (2019).

Nye, Benjamin, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J. Marshall, Ani Nenkova, and Byron C. Wallace. "A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature." In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018, p. 197. NIH Public Access, 2018

Pagnoni, Artidoro, Vidhisha Balachandran, and Yulia Tsvetkov. "Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812-4829. 2021.

Pang, Bo, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. "Long document summarization with top-down and bottom-up inference." *arXiv preprint arXiv:2203.07586* (2022).

Pfeiffer, Jonas, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. "AdapterFusion: Non-Destructive Task Composition for Transfer Learning." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487-503. 2021.

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. "AdapterHub: A Framework for Adapting Transformers." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46-54. 2020.

Pilault, Jonathan, Raymond Li, Sandeep Subramanian, and Christopher Pal. "On extractive and abstractive neural document summarization with transformer language models." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9308-9319. 2020.

Qin, Yiwei, Weizhe Yuan, Graham Neubig, and Pengfei Liu. "T5Score: Discriminative Fine-tuning of Generative Evaluation Metrics." *arXiv preprint arXiv:2212.05726* (2022).

Radev, Dragomir R., Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek et al. "MEAD-a platform for multidocument multilingual text summarization." (2004).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21, no. 1 (2020): 5485-5551.

Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982-3992. 2019.

Repke, Tim, and Ralf Krestel. "Bringing back structure to free text email conversations with recurrent neural networks." In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pp. 114-126. Springer International Publishing, 2018.

Ribeiro, Leonardo FR, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. "Investigating Pretrained Language Models for Graph-to-Text Generation." In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pp. 211-227. 2021.

Ribeiro, Leonardo FR, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations. *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics. 2022.

Rohde, Tobias, Xiaoxia Wu, and Yinhan Liu. "Hierarchical learning for generation with long source sequences." *arXiv preprint arXiv:2104.07545* (2021).

Ruan, Qian, Malte Ostendorff, and Georg Rehm. "HiStruct+: Improving Extractive Text Summarization with Hierarchical Structure Information." In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1292-1308. 2022.

Rush, Alexander M., Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379-389. 2015.

Schoonhoven, Richard, Bram Veenboer, Ben van Werkhoven, and Kees Joost Batenburg. "Going green: optimizing GPUs for energy efficiency through model-steered auto-tuning." *arXiv preprint arXiv:2211.07260* (2022).

Scialom, Thomas, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. "Questeval: Summarization asks for fact-based evaluation." *arXiv preprint arXiv:2103.12693* (2021).

See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017).

Sharma, Eva, Chen Li, and Lu Wang. "BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2204-2213. 2019.

Stede, Manfred, Heike Bieler, Stefanie Dipper, and Arthit Suriyawongkul. "Summar: Combining linguistics and statistics for text summarization." *Frontiers in Artificial Intelligence and Applications* 141 (2006): 827.

Steen, Julius, and Katja Markert. "How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1861-1875. 2021.

Steinberger, Josef, and Karel Jezek. "Using latent semantic analysis in text summarization and summary evaluation." *Proc. ISIM* 4, no. 93-100 (2004): 8.

Surita, Gabriela, Rodrigo Nogueira, and Roberto Lotufo. "Can questions summarize a corpus? Using question generation for characterizing COVID-19 research." *arXiv preprint arXiv:2009.09290* (2020).

Teufel, Simone, and Marc Moens. "Summarizing scientific articles: experiments with relevance and rhetorical status." *Computational linguistics* 28, no. 4 (2002): 409-445.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Wallace, Byron C., Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. "Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization." *AMIA Summits on Translational Science Proceedings* 2021 (2021): 605.

Wan, David, and Mohit Bansal. "FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1010-1028. 2022.

Wang, Alex, Kyunghyun Cho, and Mike Lewis. "Asking and Answering Questions to Evaluate the Factual Consistency of Summaries." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008-5020. 2020a.

Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide et al. "CORD-19: The COVID-19 Open Research Dataset." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. 2020b.

Wang, Ruize, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. "K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1405-1418. 2021.

Xiao, Wen, and Giuseppe Carenini. "Extractive Summarization of Long Documents by Combining Global and Local Context." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3011-3021. 2019.

Xiao, Wen, and Giuseppe Carenini. "Systematically Exploring Redundancy Reduction in Summarizing Long Documents." In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 516-528. 2020.

Xie, Qianqian, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. "Pre-trained language models with domain knowledge for biomedical extractive summarization." *Knowledge-Based Systems* 252 (2022): 109460.

Xu, Shusheng, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. "Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1784-1795. 2020.

Yuan, Weizhe, Graham Neubig, and Pengfei Liu. "Bartscore: Evaluating generated text as text generation." *Advances in Neural Information Processing Systems* 34 (2021): 27263-27277.

Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham et al. "Big bird: Transformers for longer sequences." *Advances in neural information processing systems* 33 (2020): 17283-17297.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In *International Conference on Machine Learning*, pp. 11328-11339. PMLR, 2020a.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." *arXiv preprint arXiv:1904.09675* (2019b).

Zhang, Xiang, Ping Geng, Tengteng Zhang, Qian Lu, Peng Gao, and Jing Mei. "Aceso: PICO-guided evidence summarization on medical literature." *IEEE journal of biomedical and health informatics* 24, no. 9 (2020b): 2663-2670.

Zhang, Xingxing, Furu Wei, and Ming Zhou. "HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5059-5069. 2019a.

Zheng, Hao, and Mirella Lapata. "Sentence Centrality Revisited for Unsupervised Summarization." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6236-6247. 2019.

Zhong, Ming, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. "Extractive Summarization as Text Matching." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6197-6208. 2020.

Zhu, Chenguang, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. "Enhancing Factual Consistency of Abstractive Summarization." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 718-733. 2021.

Zhu, Ming, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. "Question answering with long multiple-span answers." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3840-3849. 2020.